Predicting Monthly Volatility

He Li *lihezizou@gmail.com, Nov. 16th, 2021*

1. Data description and pre-processing

We are handling the dataset containing stock price data for 6 different stocks, sampled at 1 minute frequency, on the trading days from Jan. 2nd, 2017 to Dec. 29th, 2017. The dataset contains 252 trading days with trading time from 9:30 am to 4:00 pm (391 minutes in total). After examining the data, we discover the following irregular records and modify them accordingly:

(1) Day 327 in the dataset, Nov. 24th, 2017, is a half trading day and contains only 211 trading minutes. We delete this day for all 6 stocks.

(2) Stock a and d both contain 93 irregular price values, (0.0 for stock a and 1.0 for stock d), we replace these prices by the prices one-minute prior to them.

(3) Stock f contains 6 trading days without any price changes (no volatility), we delete these 6 days for stock f.

(4) Stock a, c, d, f contain $164, 31, 111, 1371$ missing values, respectively. For stock f, most missing values appear on the 6 zero-volatility trading days mentioned above. We fill these NaN values by their most recent available prices (indicating a zero volatility over this period). If the missing period is at the beginning of a trading day (starts from 9:30 am), we fill these NaN values by their next available prices.

(5) As indicated by the price time series plot in Figure 1 in the appendix, stock c experienced a 45% price drop at the opening minute on day 149. We delete this day for stock c.

Our goal is to forecasting the volatility over the next month following the end of the samples for all 6 stocks. Consider the price sequence as a stochastic process, for any time interval $[t-1, t]$, the empirical estimation of the quadratic variation of a stochastic process is given by the following *realized variance*, $RV_{t+1} = \sum_{j=1}^{1/\Delta} r^2_{t+j\Delta,\Delta}$, where $r_{t+j\Delta,\Delta}$ is the return (either percentage or logarithm return) over the period $[t+(j-1)\Delta, t+j\Delta]$. Here we consider the partition of time interval $[t, t+1]$ into $1/\Delta$ equal sub-intervals. For this project, t and $t+1$ can be seen as the closing minute for day t and day $t+1$, respectively, and we consider the even mesh over this period. Therefore, $\sqrt{RV_t}$ measures the daily volatility on day t, including the intraday volatility as well as the close-to-open volatility.

Barndorff-Nielsen and Shephard (2002) derives the consistency of RV to quadratic variation and asymptotic normality result under the stochastic volatility model. The theory suggests that we should sample prices as often as possible to better approximate the underlying quadratic variation based on realized variance. However, in practice, the benefit of very high frequency sampling is swamped by the market micro-structure effects, such as liquidity effects, bid/ask bounce and mis-recordings (Merton, 1980; Roll, 1984). For this project, we consider the 5-minute sampling frequency, which is a common choice in many references (see, e.g., Ghysels et al. (2006); Hansen and Lunde (2011)) and in practice.

We include the daily volatility time series plot for all 6 stocks in Figure 2 in the appendix. We discover that stock b has a low volatility period at the first half of the year, and a high volatility period at the second half of the year. To give a better prediction for the next monthly volatility, we only use the latter part data for stock b. It also should be noted that all stocks have some period with very large volatility. These data points should be handled properly during the training process.

Table 1 in the appendix presents summary statistics of stocks' daily volatility.

2. Model description

In this section, we give a brief introduction to some candidate models we use later.

Inspired by the HARCH model, Corsi (2009) proposes an additive cascade model of different volatility components designed to mimic the actions of different types of market participants. This model is known as the *heterogeneous autoregressive model of realized variance (HAR-RV)*, and has the following form,

$$\mathrm{OP}(RV_{t,h}) = \beta_0 + \beta_d \mathrm{OP}(RV_t) + \beta_w \mathrm{OP}(RV_{t-5,5}) + \beta_m \mathrm{OP}(RV_{t-21,21}) + \varepsilon_t. \tag{2.1}$$

where $RV_{t,h}$ is the averaged daily realized variance over the period $[t, t+h]$ and OP is some operator, which could be $\mathrm{OP}(x) = \sqrt{x}$, $\mathrm{OP}(x) = \log(x)$ or identity mapping $\mathrm{OP}(x) = x$. Therefore, the model predicts future volatility using a daily, a weekly and a monthly component, respectively. The HAR-RV model can be seen as a prediction which uses the exponential smoothing (piecewise constant) of lagged values of RV_t.

The HAR-RV model has some variants. One may consider a continuous-time jump diffusion process for price. Barndorff-Nielsen and Shephard (2004) define the standardized *realized bipower variation* as $BV_{t+1} = \mu_1^{-2} \sum_{j=2}^{1/\Delta} |r_{t+j\Delta,\Delta}| \left|r_{t+(j-1)\Delta,\Delta}\right|$, where $\mu_1 = \sqrt{2/\pi}$. We can therefore detect the jump component by $J_{t+1} = \max\{RV_{t+1} - BV_{t+1}, 0\}$.

Andersen et al. (2007) propose the following *HAR-RV-J* model,

$$\mathrm{OP}(RV_{t,h}) = \beta_0 + \beta_d \mathrm{OP}(RV_t) + \beta_w \mathrm{OP}(RV_{t-5,5}) + \beta_m \mathrm{OP}(RV_{t-21,21}) + \beta_J \mathrm{OP}(J_t) + \varepsilon_t. \tag{2.2}$$

Huang and Tauchen (2005) propose the following ratio-statistic to identify significant jumps,

$$Z_{t+1} = \Delta^{-1/2} \frac{(RV_{t+1} - BV_{t+1}) RV_{t+1}^{-1}}{\left(\mu_1^{-4} + 2\mu_1^{-2} - 5\right) \max\left\{1, TQ_{t+1} BV_{t+1}^{-2}\right\}}.$$

To take market micro-structure noise into consideration, we adopt the modification of realized bipower variation and tripower quarticity from Andersen et al. (2007) as follows,

$$BV_{t+1} = \mu_1^{-2}(1-2\Delta)^{-1} \sum_{j=3}^{1/\Delta} |r_{t+j\Delta,\Delta}| \left|r_{t+(j-2)\Delta,\Delta}\right|,$$

$$TQ_{t+1} = \Delta^{-1} \mu_{4/3}^{-3} (1-4\Delta)^{-1} \sum_{j=5}^{1/\Delta} |r_{t+j\Delta,\Delta}|^{4/3} \left|r_{t+(j-2)\Delta,\Delta}\right|^{4/3} \left|r_{t+(j-4)\Delta,\Delta}\right|^{4/3},$$

where $\mu_{4/3} = 2^{2/3}\Gamma(7/6)\Gamma(1/2)^{-1}$. Based on this ratio-statistic, we can split realized variance into two parts, significant jump and residual,

$$J_{t+1,\alpha} = \mathbf{1}[Z_{t+1} > \Phi_\alpha][RV_{t+1} - BV_{t+1}], \;\; C_{t+1,\alpha} = \mathbf{1}[Z_{t+1} \leq \Phi_\alpha]RV_{t+1} + \mathbf{1}[Z_{t+1} > \Phi_\alpha]BV_{t+1}].$$

Here Φ_α is the α-quantile for standard normal distribution, we let $\alpha = 0.975$ in our case. We now introduce our next candidate model, the *HAR-RV-CJ* model, proposed by Andersen et al. (2007) as follows,

$$\begin{aligned}\mathrm{OP}(RV_{t,h}) = \beta_0 + \beta_{CD}\mathrm{OP}(C_t) + \beta_{CW}\mathrm{OP}(C_{t-5,5}) + \beta_{CM}\mathrm{OP}(C_{t-21,21}) \\ + \beta_{JD}\mathrm{OP}(J_t) + \beta_{JW}\mathrm{OP}(J_{t-5,5}) + \beta_{JM}\mathrm{OP}(J_{t-21,21}) + \varepsilon_t. \end{aligned} \tag{2.3}$$

Finally, we consider the *Mixed Data Sampling (MIDAS)* regression model proposed by Ghysels et al. (2006, 2007). Specifically,

$$\mathrm{OP}(RV_{t,h}) = \beta_0 + \beta_1 \sum_{k=1}^{k_{\max}} b(k,\theta)\mathrm{OP}(X_{t-k+1}) + \varepsilon_t, \tag{2.4}$$

where we consider the parameterization of $b(k,\theta)$ as follows, $b(k,\theta) = \frac{e^{\theta_1 k + \ldots + \theta_Q k^Q}}{\sum_{j=1}^{k_{\max}} e^{\theta_1 j + \ldots + \theta_Q j^Q}}$. The main idea of MIDAS regression is to use regressors which may have different frequency from the response variable. We consider three type of regressors below in the experiment, the realized variance RV_t, the absolute return AR_t, $AR_t = \sum_{j=1}^{1/\Delta} |r_{t+j\Delta,\Delta}|$ and the modified realized bipower variation BV_t. It has been stated in Barndorff-Nielsen and Shephard (2004); Ghysels et al. (2006) that the use of absolute return (and realized bipower variation) could capture the volatility better.

3. Numerical results

In this section, we perform the model fitting and selection on all 6 stocks, using models mentioned above. Specifically, we consider the following 6 models as candidate models, HAR-RV model (2.1), HAR-RV-J model (2.2), HAR-RV-CJ model (2.3) with modified realized bipower variation and tripower quarticity, and MIDAS regression (2.4) with regressor RV_t, AR_t, BV_t, respectively. Our setups are summarized as follows,

(1) For MIDAS regression, we set $k_{\max} = 9$ and use the past 10 day volatility information to forecast the future monthly volatility. We consider the case where $\theta = (\theta_1, \theta_2)$ in $b(k,\theta)$.

(2) Each stock has approximate 200 samples to work on. We split the dataset into two part, the training set contains the first 70% of the data and is used to fit different models. We evaluate the performances of different models on the rest part of the data, conduct model selections and estimate the standard deviations for the models' predictions.

(3) Stock f has a step-wise price pattern, and the realized bipower variations for f are always 0. In this case, we exclude the HAR-RV-CJ model (2.3) and the MIDAS regression (2.4) with regressor BV_t.

(4) To remove the effect of some large values in the dataset, we consider the logarithm transformation on both sides of the regression.

(5) The HAR-type models are fitted by least squares and are performed by scikit-learn Ridge regressor in Python. We use scipy optimization method to solve θ and Linear regressor in scikit-learn to solve β_0, β_1 for the MIDAS models.

We now report the best model for each stock, the performance of the model, the prediction for the next month volatility, and the fitted parameters. A comparison between the ground truths and our predictions is plotted in Figure 3 and Figure 4 in the appendix, for both the training dataset and the validation dataset.

Stock a: the best model is the HAR-RV-CJ model (2.3), with validation MSE 6.44. The point estimation for next month is 8.03 with the 95% confidence interval $[5.35, 12.05]$. The fitted parameters are,

$$(\beta_0, \beta_{JD}, \beta_{JW}, \beta_{JM}, \beta_{CD}, \beta_{CW}, \beta_{CM}) = (4.15, -0.04, 0.04, -0.16, 0.12, 0.1, 0.38).$$

Stock b: the best model is the HAR-RV model (2.1), with validation MSE 28.65. The point estimation for next month is 71.94 with the 95% confidence interval $[62.34, 83.02]$. The fitted parameters are,

$$(\beta_0, \beta_d, \beta_w, \beta_m) = (6.84, 0.13, 0.03, 0.15).$$

Stock c: the best model is the MIDAS regression (2.4) with regressor BV_t, with validation MSE 3.77. The point estimation for next month is 10.77 with the 95% confidence interval $[8.75, 13.26]$. The fitted parameters are,

$$(\theta_1, \theta_2, \beta_0, \beta_1) = (-0.68, 0.04, 4.32, 0.59).$$

Stock d: the best model is the HAR-RV-CJ model (2.3), with validation MSE 30.03. The point estimation for next month is 105.57 with the 95% confidence interval $[95.42, 116.79]$. The fitted parameters are,

$$(\beta_0, \beta_{JD}, \beta_{JW}, \beta_{JM}, \beta_{CD}, \beta_{CW}, \beta_{CM}) = (9.53, 0.00, -0.03, 0.06, 0.00, 0.02, -0.09).$$

Stock e: the best model is the MIDAS regression (2.4) with regressor BV_t, with validation MSE 0.14. The point estimation for next month is 13.64 with the 95% confidence interval $[12.91, 14.41]$. The fitted parameters are,

$$(\theta_1, \theta_2, \beta_0, \beta_1) = (0.02, -0.1, 5.37, -0.07).$$

Stock f: the best model is the HAR-RV model (2.1), with validation MSE 19.86. The point estimation for next month is 13.07 with the 95% confidence interval $[7.62, 22.45]$. The fitted parameters are,

$$(\beta_0, \beta_d, \beta_w, \beta_m) = (4.71, 0.04, 0.1, 0.06).$$

4. Discussions

From the results above, generally speaking, the two most promising model would be the MIDAS regression (2.4) with regressor BV_t and HAR-RV-CJ model (2.3) with modified realized bipower variation and tripower quarticity. These two models also give strong out-of-sample performances on different data experiment in the literature.

Additional table and plots can be found in the separate appendix file. We also provide the jupyter notebook file with code and extra results.