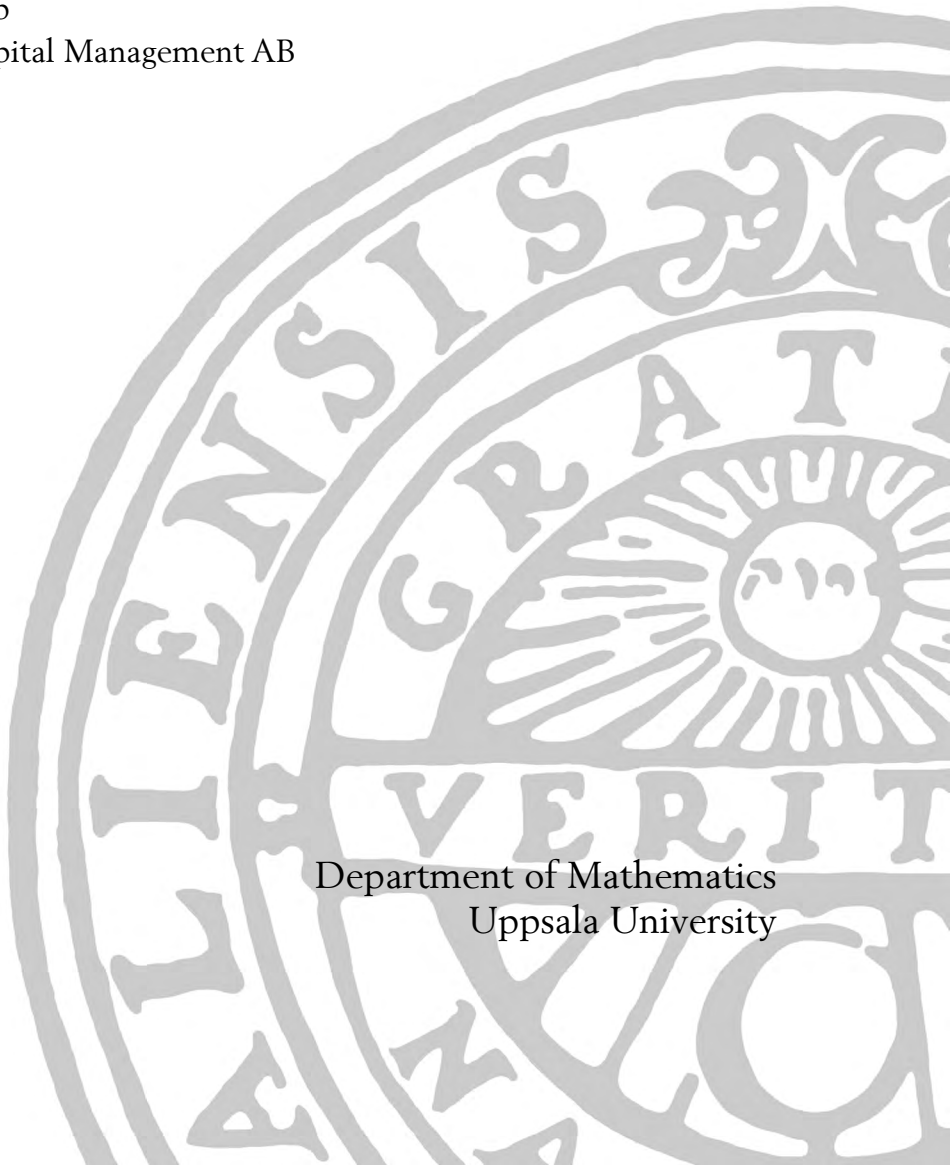
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2017:27

Volatility modelling with exogenous binary variables

William Gustafsson

Examensarbete i matematik, 30 hp
Handledare: Jukka Harju, Volt Capital Management AB
Ämnesgranskare: Rolf Larsson
Examinator: Kaj Nyström
Juni 2017



Department of Mathematics
Uppsala University

Speculation does not determine prices; it has to accept the prices that are determined in the market. Its efforts are directed to correctly estimating future price-situations, and to acting accordingly. The influence of speculation cannot alter the average level of prices over a given period; what it can do is to diminish the gap between the highest and the lowest prices.

– Ludwig von Mises, *The Theory of Money and Credit*

Abstract

Volatility modelling has been attracting both academic and practical interest for a long time, and a huge amount of different methods and models have been proposed. One reason for the abundance of different methods is the fact that volatility is an unobservable variable. In general it is taken to mean a measure of the variability of the price of an asset. This paper begins with an overview over volatility in general, after which a number of different models are tested on 66 different futures price series. Here the goal is to incorporate known prior and future economic data release dates, known to cause excess volatility, in the models. This is shown to improve model fitness for most models, whereas the improvement in forecasting performance is less clear.

Acknowledgements

I would like to thank my supervisor Jukka Harju at Volt Capital for giving me the opportunity to do this project, which has given me an insight into the financial business and a chance to see the practical applications of my education.

I would also like to thank my supervisor Rolf Larsson at Uppsala University for his support during the entire project.

Contents

1 Introduction **1**
- 1.1 Volatility . . . 1
- 1.2 Data . . . 3

2 Measuring historical volatility **4**
- 2.1 Realized volatility . . . 4
- 2.2 Yang-Zhang volatility estimate . . . 4

3 Forecasting **6**
- 3.1 GARCH models . . . 6
- 3.2 GARCH-MIDAS . . . 7
  - 3.2.1 Parameters . . . 10
- 3.3 HAR-RV . . . 10

4 Adding exogenous variables **11**
- 4.1 Multiplicative . . . 11
- 4.2 Additive . . . 12

5 Parameter estimation and model evaluation **13**
- 5.1 QMLE . . . 13
- 5.2 Other . . . 14

6 Results **16**
- 6.1 GARCH-MIDAS . . . 16
- 6.2 HAR-RV . . . 16
- 6.3 Comparison . . . 17

7 Conclusions **18**
- 7.1 Further research . . . 19

References **20**

A Option pricing **21**

B Tables **23**
- B.1 Data . . . 23
- B.2 GARCH-MIDAS . . . 25
- B.3 HAR-RV . . . 26
- B.4 Comparison . . . 27

C Plots **28**
- C.1 GARCH-MIDAS . . . 28
- C.2 Comparison . . . 29

1 Introduction

In asset management and automated trading, robust and accurate volatility measures and predictions are of utmost importance, affecting trading decisions and risk assessment. Experience shows that certain unexpected—or expected—events tend to increase market volatility when occurring, and if your volatility model does not factor in this excess volatility as pertaining to a certain event, for example your risk assessment could be unnecessarily conservative following this sudden volatility spike, leading to missed opportunities. This gets particularly embarrassing when the volatility increasing event was actually readily expected, or even known for a certain to come to occur on a certain date.

This paper will investigate different volatility models incorporating known prior and future economic data release dates, known to cause excess volatility, when trying to forecast volatility for individual assets.

1.1 Volatility

First of all, a word on notation is in order. Volatility is a term used in economics in general, and in finance in particular. It is not a strictly defined mathematical term, hence the wide array of different ways to measure and model volatility mathematically. In finance, the term refers to the dispersion of returns for an asset, i.e. how much the asset price tends to fluctuate. Higher volatility tends to be associated with higher risk, and thus higher potential rewards, and the opposite for lower volatility. For the rest of this paper, the term volatility will refer to this general financial meaning.

The problem with volatility is that it is not actually observable on the market. The only thing we can observe is the assets price, perhaps at an arbitrarily accurate level, but since volatility lacks an exact definition we do not have a unique way to infer it from the prices. One measure of historical volatility is the sample standard deviation of asset returns. This is an intuitive measure, as sample standard deviation is a generally used measure for the dispersion of the values in a sample. The problem is that different choices of sample size, period, and frequency give different measures, that can also have different economic or financial interpretations.

Let c_t denote the daily closing price of an asset. The daily log return is then

$$r_t = \log\left(\frac{c_t}{c_{t-1}}\right)$$

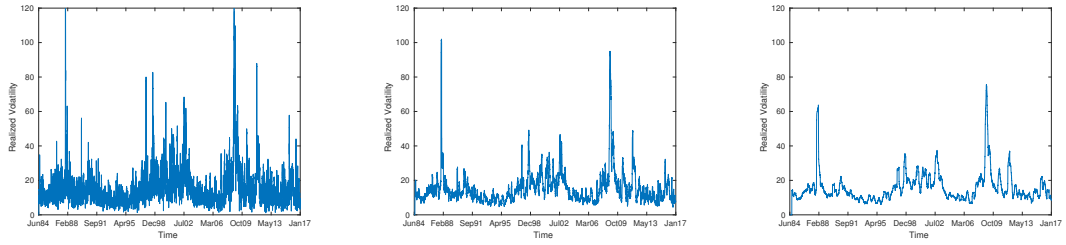
and the simple close-to-close historical volatility measure is

$$\sigma_{cc} = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N} r_t^2}$$

assuming zero drift. Volatility is often quoted in percentage and on an annualized basis, yielding the expression

$$\sigma_{cc} = 100 \cdot \sqrt{\frac{252}{N-1}\sum_{t=1}^{N} r_t^2}$$

Using this measure as an example, with $N = 5$ (one trading week), $N = 21$ (one trading month), and $N = 63$ (one trading quarter), gives the following historical realized volatilities for the same underlying asset:



(a) Weekly (b) Monthly (c) Quarterly

Figure 1.1: Realized volatility for different sample lengths

The three measures seem to capture the same underlying property of the asset returns, but at different "resolutions". This will be expanded upon in Section 2.1.

Another slightly different concept when discussing volatility is the so called *Implied Volatility*. This comes from the area of option pricing, where the underlying asset is assumed to follow some stochastic process; in the simplest case it's the standard geometric Brownian motion, as in the original Black-Scholes model. The asset price S_t is then assumed to be given by

$$S_t = S_0 e^{(r - \frac{\sigma^2}{2})t + \sigma W_t}$$

where W_t is a standard Brownian motion, and the theoretical price of an European call option is given by

$$C_\theta = \mathcal{N}(d_1)S_0 - \mathcal{N}(d_2)Ke^{-rT} \tag{1.1}$$

where $\mathcal{N}(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and

$$d_1 = \frac{\log\left(\frac{S_0}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

A European call option is a contract giving the holder the right—but not the obligation, as opposed to a futures contract as we shall see—to buy the underlying asset S at some future maturity time for the strike price K. Given that the asset can be bought and sold on the market at this time for some price that will be different from the strike price, the value of the option at maturity ranges from 0 to the positive difference between the strike and the spot price. The pricing formula in (1.1) is derived in Appendix A.

Here S_0 denotes the current asset price, K the option's strike price, r the risk free interest rate, T the time to maturity, and σ the underlying volatility. All parameters except σ could be said to be observable on the market, and due to the way transactions are settled on most financial markets—with a bid-ask spread and then settling on the

price—one can infer the markets assessment of the volatility by observing an option market price C_M and then solving

$$C_\theta - C_M = 0$$

for the parameter σ, which is then called the implied volatility.

Official estimates of the implied volatility can be obtained from the market, but the problem is that it relies heavily on the underlying model assumption which almost certainly does not capture the "true" mechanics of the price process. Thus this thesis will not focus on implied volatility, other than noting that some of the models that will be used could use it as an external variable for prediction.

1.2 Data

All models will be built and evaluated using historical price data for 66 futures contracts, shown in Table B.1. The column $Price_C$ indicates from what date closing prices are available, and $Price_{OHLC}$ from what date *Open*, *High*, *Low*, and *Close* prices are available. The closing price is the price at the end of the trading day, the price usually referred to when quoting historical asset prices, opening prices are at the beginning of the trading day, and high and low are the highest and lowest prices of the trading day respectively. All time series extend to January 2017.

A futures contract is a contract between a buyer and a seller of an asset, agreeing to make a trade for a certain agreed upon price at some future delivery date. The origin of this type of contract was in agricultural commodities, where a buyer wanted to secure a certain price for a good which would have to be purchased later in the year, or conversely where a producer wanted to secure a certain income in the future independent of the total production (e.g. harvest of crops). This is a crucial financial instrument, allowing for actors on the market to plan their cash flows over a longer period of time and acting as an insurance against unexpected events. This type of contract has later expanded to include all types of underlying assets as seen by the examples in Table B.1, e.g. index and currency futures, but fills the same function still.

Once the contract has been negotiated, it can be traded on an open exchange to third parties. This is where the interest of this project lies, as the futures contract then takes the form of most other traded financial assets, with daily fluctuating prices. The price fluctuations stem from the fact that the contract can become more or less valuable vis-à-vis the originally negotiated price as the time of delivery draws nearer—will the spot price at the delivery date be lower or higher than the price of the futures contract?[1]

A more practical reason for using futures for this particular project is the fact that it is the second largest global financial market, surpassed only by the Foreign Exchange (currency) market. This means that the market has a very high liquidity, allowing for fast and automated trade.

[1]This type of trading performed by the much berated "speculator" is, contrary to popular belief, a crucial and integral part of a working economy. Increasing prices is a signal to producers that the future demand will be high, directing production from less to more demanded goods, thus satisfying more consumer needs.

2 Measuring historical volatility

2.1 Realized volatility

The simple realized volatility measure mentioned in the introduction can be extended to include arbitrarily many observations per day. Let s_t denote the price of an asset at time t. Given m observations of the price each day, and N days of observations, the corresponding intraday returns are given by

$$r_t = \log\Big(\frac{s_t}{s_{t-\frac{1}{m}}}\Big), \quad t = \frac{1}{m}, \frac{2}{m}, \ldots, \frac{(m-1)\cdot(N-1)}{m}, N-1$$

The *N-day* Realized Volatility for the time period $[t - N + \frac{1}{m}, t]$ is then

$$RV_t(N) = \sum_{j=1}^{m\cdot N} r^2_{t-N+\frac{j}{m}}$$

and an annualized daily volatility estimate is given by

$$\sigma_t^{RV} = 100 \cdot \sqrt{\frac{252}{N-1} RV_t(N)}$$

As can be seen this is a more general form of the simple close-to-close measure mentioned in the introduction, where $m = 1$.

One problem with this approach is that it requires very high frequency data, which is not always readily available[2]. Another problem could be the introduction of measurement noise when using very high frequency data. In the next section I introduce a model trying to counter these problems.

2.2 Yang-Zhang volatility estimate

One model proposed in Yang and Zhang (2000) as an improved volatility measure over the standard close-to-close, and which does not require more than four prices each day, is the Yang-Zhang model. Let o_t be the opening price at day t, c_t denote the closing price, h_t the highest price during day t, and l_t the lowest price. The N-day Yang-Zhang volatility measure is then given by

$$YZ_t(N) = \sigma^2_{ON} + k \cdot \sigma^2_{OC} + (1-k)\sigma^2_{RS}$$

where

$$\sigma^2_{ON} = \sum_{j=1}^{N-1} \left(\log\Big(\frac{o_{t-N+1+j}}{c_{t-N+j}}\Big) - \overline{\log\Big(\frac{o_{t-N+1+j}}{c_{t-N+j}}\Big)} \right)^2$$

denotes the *Overnight* volatility, i.e. the volatility between the closing price the previous day and the opening price today,

[2]For this project, the highest frequency available is hourly prices, which may be too sparse. Also this data is available for a much shorter time span than the daily data.

$$\sigma_{OC}^2 = \sum_{j=1}^{N} \left(\log\left(\frac{c_{t-N+j}}{o_{t-N+j}}\right) - \overline{\log\left(\frac{c_{t-N+j}}{o_{t-N+j}}\right)} \right)^2$$

denotes the *Open-to-Close* volatility, i.e. the volatility between the opening and closing price the same day, and

$$\sigma_{RS}^2 = \sum_{j=1}^{N} \left(\log\Big(\frac{h_{t-N+j}}{c_{t-N+j}}\Big)\log\Big(\frac{h_{t-N+j}}{o_{t-N+j}}\Big) + \log\Big(\frac{l_{t-N+j}}{c_{t-N+j}}\Big)\log\Big(\frac{l_{t-N+j}}{o_{t-N+j}}\Big) \right)$$

is the *Rogers-Satchell* volatility measure, proposed in Rogers and Satchell (1991). Here the first term descirbes the variation in the day's prices as the difference between the high and the open and close prices, and similarly in the second term but for the lowest price. This has a few good properties, e.g. that if the price is monotonically increasing ($h_t = c_t$ and $l_t = o_t$) or decreasing ($h_t = o_t$ and $l_t = c_t$), the volatility that day is zero. One drawback is that it does not take into account opening jumps, i.e. changes between yesterday's closing price and today's opening price. Hence the addition of the overnight volatility estimate σ_{ON}^2.

The Yang-Zhang model is the sum of the overnight volatility and a weighted average of the open-to-close and RS volatilities, where the weight k is a choice variable but empirical tests have led to the generally accepted form

$$k = \frac{0.34}{1.34 + \frac{N+1}{N-1}}$$

An annualized daily volatility measure using YZ is then given by

$$\sigma_t^{YZ} = 100 \cdot \sqrt{\frac{252}{N-1} YZ_t(N)}$$

3 Forecasting

3.1 GARCH models

The original ARCH (AutoRegressive Conditional Heteroskedasticity) model was proposed in Engle (1982). Here we let $r_t = \mu_t + a_t$, $\mu_t = \mathrm{E}_{t-1}[r_t]$, and the residuals a_t are modelled with an ARCH(p) model as

$$\begin{aligned} a_t &= \sqrt{h_t}\varepsilon_t \\ h_t &= \alpha_0 + \sum_{i=1}^{p} \alpha_i a_{t-j} \end{aligned}$$

where ε_t are i.i.d. random variables with mean 0 and variance 1, often assumed to be normally distributed $\mathcal{N}(0,1)$ or Student's t–distributed, and $\alpha_0 > 0$, $\alpha_i \geq 0$.

This was later extended in Bollerslev (1986) to the Generalized ARCH (or GARCH) model, where previous values of h_t are added to the volatility process. This helps to model the often observed phenomenon of *volatility clustering*, i.e. that higher volatility tends to be followed by a period of increased volatility. The GARCH(p,q) model is given by

$$\begin{aligned} a_t &= \sqrt{h_t}\varepsilon_t \\ h_t &= \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} \end{aligned}$$

with ε_t as in the ARCH model, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1$. The final constraint can be seen by letting

$$\eta_t = a_t^2 - h_t \quad \Longrightarrow \quad h_t = a_t^2 - \eta_t$$

and plugging this into the GARCH equation, yielding

$$a_t^2 = \alpha_0 + \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) a_{t-i}^2 + \eta_t + \sum_{j=1}^{p} \beta_j \eta_{t-j}$$

Thus the unconditional variance is given by

$$\begin{aligned} \mathrm{Var}\big[a_t\big] &= \mathrm{E}\big[a_t^2\big] - \mathrm{E}\big[a_t\big]^2 = \mathrm{E}\big[a_t^2\big] - \mathrm{E}\big[\sqrt{h_t}\varepsilon_t\big]^2 = \mathrm{E}\big[a_t^2\big] - \Big(\mathrm{E}\big[\sqrt{h_t}\big] \underbrace{\mathrm{E}\big[\varepsilon_t\big]}_{=0}\Big)^2 \\ &= \mathrm{E}\big[a_t^2\big] = \alpha_0 + \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) \mathrm{E}\big[a_{t-i}^2\big] + \mathrm{E}\big[\eta_t\big] + \sum_{j=1}^{p} \beta_j \mathrm{E}\big[\eta_{t-j}\big] \end{aligned}$$

Since a_t is stationary, and thus $\mathrm{E}\big[a_t^2\big] = \mathrm{E}\big[a_{t-1}^2\big]$, and $E\big[\eta_t\big] = 0$, we have that

$$\mathrm{Var}\big[a_t\big] = \mathrm{E}\big[a_t^2\big] = \frac{\alpha_0}{1 - \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i)}$$

which has to be positive. Also, since $\mathrm{E}\big[a_t^2\big] = \mathrm{E}\big[h_t\varepsilon_t^2\big] = \mathrm{E}\big[h_t\big]$, the unconditional mean of the volatility process is also

$$\mathrm{E}\big[h_t\big] = \frac{\alpha_0}{1 - \sum_{i=1}^{\max(p,q)}(\alpha_i + \beta_i)}$$

Even the simple GARCH(1,1) model captures a lot of market features, and has been one of the most popular volatility models for a long time. Due to its rather simple structure, innumerable variations of the original model have been proposed, adding new features to either the mean equation, volatility equation, or both. Notable examples include, but are in no way restricted to, the following:

- IGARCH (Integrated GARCH), where the above mentioned constraint on $\sum_{i=1}^{\max(p,q)}(\alpha_i + \beta_i)$ is actually changed so that $\sum_{i=1}^{\max(p,q)}(\alpha_i + \beta_i) = 1$. This leads to all past shocks η_t having an effect on a_t^2, which has certain good properties, but also to the unconditional variance being undefined.

- EGARCH (Exponential GARCH), where the logarithm of the variance is modelled and then the exponential is used. This leads to less restrictions on the parameters in order to maintain positivity, since the exponential is always positive. There are different formulations of the EGARCH model, one being

$$\log(h_t) = \alpha_0 + \sum_{i=1}^{q} \alpha_i \frac{|a_{t-i}| + \gamma a_{t-i}}{h_{t-i}} + \sum_{j=1}^{p} \beta_j \log(h_{t-j})$$

  This has the added advantage of modelling the so called leverage effect often observed on the market, which means that negative shocks tend to have a larger impact on volatility than positive ones.

- GARCH-M (GARCH-in-mean), which adds the volatility directly to the returns equation, $r_t = \mu + ch_t + a_t$, with a_t as before and c a constant.

3.2 GARCH-MIDAS

The so called GARCH-MIDAS model was proposed in Engle, Ghysels, and Sohn (2008). It follows the model proposed by Campbell in Campbell (1991) where the unexpected (log) return, i.e. the difference between the actual return r_{it} on day t of month i and the expected return on day $t-1$, is given by

$$\begin{aligned} r_{it} - \mathrm{E}_{i,t-1}[r_{it}] = &\left(\mathrm{E}_{it}\left[\sum_{j=0}^{\infty}\rho^j \Delta d_{it+j+1}\right] - \mathrm{E}_{i,t-1}\left[\sum_{j=0}^{\infty}\rho^j \Delta d_{it+j+1}\right]\right) \\ &- \left(\mathrm{E}_{it}\left[\sum_{j=1}^{\infty}\rho^j r_{it+j+1}\right] - \mathrm{E}_{i,t-1}\left[\sum_{j=1}^{\infty}\rho^j r_{it+j+1}\right]\right) \end{aligned}$$

where Δd_{it} is a one-period dividend difference and $\rho < 1$. An economic interpretation of this can be found in Campbell (1991) and Campbell and Shiller (1988), but the point is that negative unexpected returns indicate a larger than expected future expected

return or a lower future expected dividend growth, or both.[3] More important is the way of modelling the left hand side as

$$r_{it} - \mathrm{E}_{i,t-1}\left[r_{it}\right] = \sqrt{\tau_t g_{it}}\,\varepsilon_{it}, \quad \varepsilon_{it}|\mathcal{F}_{i,t-1} \sim \mathcal{N}(0,1)$$

where τ_t represents the long-run, and g_{it} the short-run, behavior of the asset, as proposed in Engle and Rangel (2008). Here $\mathcal{F}_t$ denotes the information set available at time t. The idea is that unexpected returns are affected by both long term information, as in future expected cash flows, economic performance on the macro level, etc., and short term factors such as daily trading volumes and market liquidity. Denoting the expected return by $\mathrm{E}_{i,t-1}\left[r_{it}\right] = \mu$, we arrive at the model formulation

$$r_{it} = \mu + \sqrt{\tau_i g_{it}}\,\varepsilon_{it}$$

where the short term part is given by the mean-reverting GARCH(1,1) process

$$g_{it} = (1 - \alpha - \beta) + \alpha\frac{(r_{i,t-1} - \mu)^2}{\tau_i} + \beta g_{i,t-1}$$

The long-run component τ_i is estimated using MIDAS (Mixed Data Sampling) regression, which is a weighted sum of data sampled at different frequencies, perhaps at a lower frequency than the daily price data. Different types of data could be used here; proposed examples include weekly/monthly/quarterly/yearly sampled realized volatility or macroeconomic variables. When using realized volatility, the long term part is given by

$$\tau_i = m + \theta\sum_{k=1}^{K}\varphi_k(\omega)RV_{i-k} \tag{3.1}$$

where K is the number of weeks/months/quarters used, and

$$RV_i = \sum_{t=1}^{N_i} r_{it}^2$$

Here N_i denotes the number of days of week/month/quarter/year i. The more general model for using L different macroeconomic variables is given by

$$\tau_i = m + \sum_{j=1}^{L}\sum_{k=1}^{K}\varphi_k(\omega_j)\theta_j X_{i-k}^{j}$$

In both cases the weights $\varphi_k(\omega)$ are used, which are typically decreasing the further from today's date they are. It is suggested to use *Beta* or *Exponential* weights, i.e.

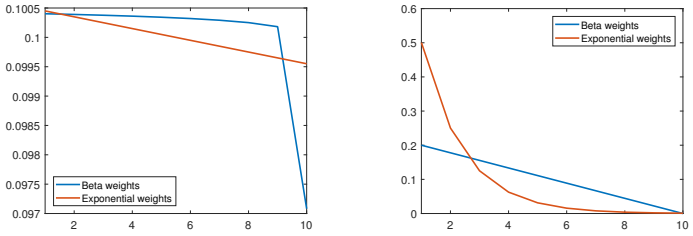
$$\varphi_k(\omega) = \frac{\left(1 - \frac{k}{K}\right)^{\omega-1}}{\sum_{j=1}^{K}\left(1 - \frac{j}{K}\right)^{\omega-1}}, \qquad \text{(Beta)}$$

or

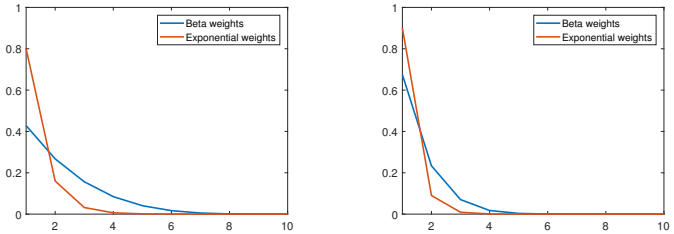
[3] The economic reasoning behind this seems to assume that all negative returns must at some point be followed by a subsequent price increase, i.e. that the asset can never become completely worthless; "Capital losses cannot continue forever." While this is obviously not universally true, it can probably be safely assumed in most cases.

$$\varphi_k(\omega) = \frac{\omega^k}{\sum_{j=1}^{K} \omega^j}, \qquad \text{(Exponential)}$$

In Figure 3.1 below the different weight structures are shown for a range of values of ω. Both schemes offer similar properties, with the Beta weights being a bit more flexible than the Exponential. For the numerical tests, Beta weights will be used.



(a) $\omega = 1.001$ (b) $\omega = 2$



(c) $\omega = 5$ (d) $\omega = 10$

Figure 3.1: Beta and exponential weights for different parameters ω

For the numerical tests a few versions of the RV model for τ will be investigated. In Equation (3.1), if we use e.g. monthly RV, τ will only change once per month. If we let $K = 12$ for example, then for each day of month i, τ_i will be a weighted sum of the RVs of the previous 12 months. One alternative to this is to use a rolling window of RVs instead, meaning that we update the components of τ daily. For each day t then, τ_t will be a weighted sum of the 12 previous windows of 21 days, i.e.

$$\tau_t = m + \theta \sum_{k=1}^{K} \varphi_k(\omega) RV_{t-k}^{rw} \tag{3.2}$$

where

$$RV_t^{rw} = \sum_{t=1}^{N} r_{t-j}^2$$

Thus the one day ahead volatility forecast is given by

$$\sigma_t^{GM} = 100 \cdot \sqrt{252 \cdot \tau_i g_t}$$

for the rolling window model, and

$$\sigma_{it}^{GM} = 100 \cdot \sqrt{252 \cdot \tau_i g_{it}}$$

for the fixed window model. In both cases the current day is $t-1$.

Finally, I propose using the Yang-Zhang volatility measure as presented in Section 2.2 in place of RV, either with a fixed or rolling window.

3.2.1 Parameters

In the GARCH-MIDAS model the free parameters that need to be estimated are $\Theta = \{\mu, \alpha, \beta, m, \theta, \omega\}$, that is the expected return μ, the GARCH coefficients α and β, the MIDAS coefficients m and θ, and the MIDAS weight parameter ω. This can be done using numerical optimization methods, minimizing e.g. the log-likelihood function.

We also have the choice parameters N and K, that is the length and number of RVs used in the MIDAS scheme. Since these are discrete parameters that have certain economic interpretations, a number of natural combinations will be tested manually.

3.3 HAR-RV

One suggested method, here called HAR-RV (Heterogeneous AR model in the RV), of forecasting volatility is to use a number of different period RV estimates as regressors in a linear regression, or as rolling inputs to an AR model, e.g.

$$\widehat{RV}_{t+1} = \beta_0 + \beta_1 RV_t(1) + \beta_2 RV_t(5) + \beta_3 RV_t(21) + \varepsilon_{t+1}, \quad \varepsilon \sim \mathcal{N}(0,1)$$

using daily ($N=1$), weekly ($N=5$), and monthly ($N=21$) realized volatilities. A number of different extensions have been proposed, e.g. in Corsi, Mittnik, Pigorsch, and Pigorsch (2008). Here a GARCH part is added, modelling the volatility of the volatility. It is suggested to model either the square root or the logarithm of the RV, here denoted $RV^{log/sqrt}$, giving the model

$$\begin{aligned}
\widehat{RV}_{t+1}^{log/sqrt} &= \gamma_0 + \gamma_1 RV_t^{log/sqrt}(1) + \gamma_2 RV_t^{log/sqrt}(5) + \gamma_3 RV_t^{log/sqrt}(21) + \sqrt{h_t}\varepsilon_{t+1} \\
h_t &= \alpha_0 + \sum_{i=1}^{q} \alpha_i a_{t-j}^2 + \sum_{j=1}^{p} \beta_j h_{t-j}
\end{aligned}$$

where

$$RV_t = \sum_{j=1}^{m} r_{t-\frac{j}{m}}^2 = \mu + a_t, \quad a_t = \sqrt{h_t}\varepsilon$$

$$RV_t^{log}(N) = \frac{1}{N}\sum_{j=1}^{N} \log(RV_{t-j}), \quad RV_t^{sqrt}(N) = \frac{1}{N}\sum_{j=1}^{N} \sqrt{RV_{t-j}}$$

As mentioned previously, these realized volatility measures require very high frequency data, which unfortunately is not available here. Thus I suggest using the Yang-Zhang measure introduced in Section 2.2 in the above model, which will be evaluated further on.

4 Adding exogenous variables

The idea is to add a number of binary variables to the models of interest, corresponding to the release of financial and economic reports on certain dates. These variables will then be 1 on the days of release, and 0 all other. For the numerical tests 13 variables are available, presented in the table below.

Name	No. of events 1984–2017
Federal Funds Target Rate - Up	164
US Employees on Nonfarm Payrol	241
Conference Board Consumer Confidence	239
GDP US Chained 2009 Dollars QoQ	239
US CPI Urban Consumers MoM SA	241
US Trade Balance of Goods and Services	242
US Treasury Federal Budget Debt	242
ISM Manufacturing PMI SA	243
University of Michigan Consumer Survey	423
US Durable Goods New Orders Industries	244
US Manufacturers New Orders Total	242
Markit US Manufacturing PMI SA	110
Federal Open Market Committee	129

As the actual time of release on the specified days are unknown, there will be three data series for each variable; one shifted forward and the other backward, corresponding to the report being released prior to, during, or after the opening hours of the market. Another reason for this is that different assets of interest are traded in different time zones.

4.1 Multiplicative

Let $I_t^j,\ j = 1, \ldots, 39$ denote the binary time series above. Inspired by the work in Bomfin (2003), I suggest multiplying the volatility model in question by a scaling factor

$$S_t = 1 + \sum_{j=1}^{39} \delta_j I_t^j$$

where δ_j are coefficients estimated when calibrating the model. This means that the I_t^j:s only affect the model on the days t when they are non-zero, scaling the volatility by a factor determined by the coefficients δ_j, and on the other days the scaling factor $S_t = 1$. Here we can also impose restrictions on the parameters δ_j, e.g. letting $\delta_j > 0$ as a way to model exclusively the historically observed increases in volatility as these are the most interesting.

The models of interest would then be as follows: the GARCH-MIDAS model becomes

$$r_t = \mu + \sqrt{\tau_t \, g_t \, S_t} \, \varepsilon_t$$

and the HAR-RV model

$$\widehat{RV}_{t+1} = \left(\gamma_0 + \gamma_1 RV_t(1) + \gamma_2 RV_t(5) + \gamma_3 RV_t(21) + \sqrt{h_t}\varepsilon_{t+1}\right) \cdot S_t$$

This approach would be in line with the initial hypothesis, that certain events increase the volatility locally but should not affect the underlying model too much, leading to a overinterpretation of sudden and somewhat expected jumps.

4.2 Additive

Another way to include the exogenous binary variables is to add them directly to the GARCH part of the model. This could have the potential benefit of letting the effect of the binary variables lag one step, leading to a slightly longer but still temporary effect. We could for example let the short term part of the GARCH-MIDAS model be

$$g_t = (1 - \alpha - \beta) + \alpha \frac{(r_{t-1} - \mu)^2}{\tau_t} + \beta g_{t-1} + \sum_{j=1}^{39} \delta_j I_t^j$$

If and insofar as all binary variables are zero on a given day t, the unconditional expectation is given by

$$\mathrm{E}\left[g_t\right] = \frac{1 - \alpha - \beta}{1 - (\alpha + \beta)} = 1$$

Otherwise we have that

$$\mathrm{E}\left[g_t\right] = 1 + \frac{\sum_{j \in \{j | I_t^j = 1\}} \delta_j I_t^j}{1 - (\alpha + \beta)} > 1$$

given that $\delta_j > 0$. This gives a similar interpretation as the previous method, with a standard model for regular days that get scaled up on days where the special events take place.

Similarly we can add the exogenous variables to the GARCH part of the HAR-RV model, i.e.

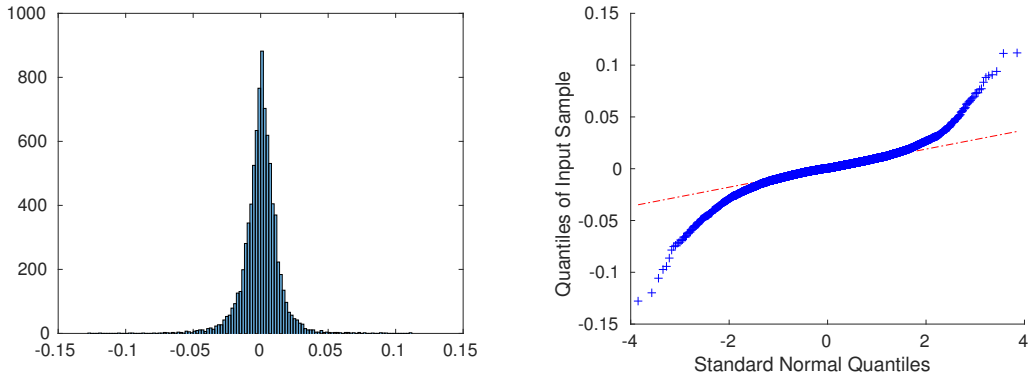
$$\begin{aligned} \widehat{RV}_{t+1} &= \gamma_0 + \gamma_1 RV_t(1) + \gamma_2 RV_t(5) + \gamma_3 RV_t(21) + \sqrt{h_t}\varepsilon_{t+1} \\ h_t &= m + \sum_{j=1}^{q} \alpha_j a_{t-j}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} + \sum_{j=1}^{39} \delta_j I_t^j \end{aligned}$$

5 Parameter estimation and model evaluation

When evaluating a model we need something to compare the results to, that is to see if the model fits the observed data. Here lies one of the problems with volatility modelling, since the actual volatility is not observable on the market. Thus a proxy is needed, where the simplest and most common is the daily squared returns r^2, or innovations $a^2 = (r - \mu)^2$. This has a few drawbacks, in that it is a noisy proxy only using the closing price each day and ignoring all intra-day price movements. As this is the only proxy available for all data series however, it will be the one that is primarily used.

5.1 QMLE

Maximum Likelihood Estimation (MLE) is a method used for parameter estimation, where the idea is to maximize the likelihood that the observed data could have been generated by the given model and parameters. For the GARCH type models used here, we can obtain the likelihood function by assuming that the innovations a_t are normally distributed. This is generally not the case for financial data, as can be seen by plotting the histogram and QQ-plot for e.g. the S&P 500 E-mini series.



(a) Histogram (b) QQ-plot

Figure 5.1: S&P 500 E-mini innovations

Here we see that the kurtosis, the fourth moment, is not consistent with a Gaussian distribution, but that the distribution has heavier tails. Disregarding this, and assuming normally distributed innovations anyway, is called Quasi Maximum Likelihood Estimation (QMLE). With a few not-so-restricting assumptions, it can be shown that the QMLE estimate is consistent and asymptotically normal. The likelihood function for the observations a_t and volatility estimates $\hat{\sigma}_t^2(\theta)$, $t = 1, \ldots, T$, is then given by

$$\mathcal{L}(\theta) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\hat{\sigma}_t^2(\theta)}} \exp\left(-\frac{a_t^2}{\hat{\sigma}_t^2(\theta)}\right)$$

where $\theta \in \Theta$ are the model parameters to be estimated. Equivalently, since the logarithm is monotonically increasing, we can take the log-likelihood function in order to get a sum instead of a product, which is given by

$$\ell(\theta) = \log(\mathcal{L}(\theta)) = \sum_{t=1}^{T} -\frac{1}{2}\Big(\log(2\pi) + \log(\hat{\sigma}_t^2(\theta)) + \frac{a_t^2}{\hat{\sigma}_t^2(\theta)}\Big)$$

The QMLE parameter estimate is then

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta) = \arg\max_{\theta\in\Theta} \ell(\theta)$$

In practice the MATLAB function `fmincon` is used to optimize, which uses an interior point algorithm to find the minimum of a constrained optimization problem, and thus the problem becomes to find

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \; -\ell(\theta)$$

such that θ satisfies the given constraints. This will be the main method used to estimate parameters for my models.

5.2 Other

A few other measures will be used to compare different models. In Patton (2011), the most robust loss functions when using an imperfect volatility proxy are shown to be

- Mean Squared Error (MSE), which is one of the standard measures for model fit, given by

  $$MSE = \frac{1}{T}\sum_{t=1}^{T}(a_t^2 - \hat{\sigma}_t^2)^2$$

  or the Root MSE (RMSE) which is

  $$RMSE = \sqrt{MSE}$$

- Quasi-Likelihood loss function, which is shown to be preferred over MSE due to less strict conditions for robustness and better distributional properties, given by

  $$QL = \sum_{t=1}^{T}\Big(\frac{a_t^2}{\hat{\sigma}_t^2} - \log\big(\frac{a_t^2}{\hat{\sigma}_t^2}\big) - 1\Big)$$

When adding more variables to a model, as in the case of the exogenous binary variables in this case, there is a risk of overfitting. To take this into account, the following measures that penalizes the number of model parameters can be used:

- Akaike Information Criterion (AIC), given by

  $$AIC = 2k - 2\,\ell(\theta)$$

  where k is the number of free parameters $\theta = \theta_1, \ldots, \theta_k$ and $\ell(\theta)$ is the log-likelihood.

- Bayesian Information Criterion (BIC), given by

$$BIC = \log(T)\cdot k - 2\,\ell(\theta)$$

  where T is the number of observations, and k and ℓ as before.

Here lower values indicate a better fit of the model. BIC penalizes the number of parameters more than AIC, but the latter has many theoretical and practical advantages over the former.

Another test to compare different models, that also takes into account the number of free parameters, is the likelihood ratio test. Let the model with less parameters be the *null* model, and the one with more parameters the *alternative* model. Fit the models and calculate the likelihoods $\mathcal{L}_0$ (null) and $\mathcal{L}_1$ (alternative). The test statistic

$$D = -2\log\Big(\frac{\mathcal{L}_0}{\mathcal{L}_1}\Big) = 2\log\Big(\frac{\mathcal{L}_1}{\mathcal{L}_0}\Big) = 2\big(\log(\mathcal{L}_1) - \log(\mathcal{L}_0)\big) = 2\big(\ell_1 - \ell_0\big)$$

is then approximately χ^2 distributed with $k = k_1 - k_0$ degrees of freedom, where k_1, k_0 are the number of free parameters for the alternative and null model respectively. The test statistic D is then compared to the corresponding quantile of the χ^2 distribution at e.g. 5% level, and the hypothesis that the simpler null model is at least as good as the alternative can be rejected if D is greater.

6 Results

6.1 GARCH-MIDAS

The first goal is to choose a suitable N and K to be used for all data series. This is since the other parameters are automatically estimated for each series, whereas N and K have to be chosen beforehand, and a general choice for these would simplify the practical use of the model. This is done for the standard model, without the exogenous variables, and will then be used for further comparisons. A number of natural choices for N is tested: 5 (weekly), 21 (monthly), 63 (quarterly), 126 (semiannually), and 252 (annually). This is combined with K's such that $N \cdot K$ is equal to approximately 126 (half year), 252 (one year), and 504 (two years). For each combination the model parameters are estimated by maximizing the log-likelihood, for each of the 66 data series individually. The sum of the log-likelihood, AIC, and BIC measures over all series are shown in Figures C.1, C.2, and C.3. These indicate that for the fixed window version, $N = 5$ and $K = 50$ is the best choice, corresponding to one year of weakly RVs used in the MIDAS part. For the rolling window version $N = 5$ and $K = 101$, two years of weekly RVs, gives the best result. This prompted further investigations, to see whether certain series gave a disproportionate addition to the sum. Checking each series individually, the N and K that most often gave the largest log-likelihood was for both versions $N = 5$ and $K = 101$, with 32 and 29 series respectively, shown in Table B.2. The same result was obtained from AIC and BIC. Thus this will will be the choice of N and K used from now on, when introducing the exogenous variables.

Proceeding to check if the addition of the exogenous variables improved model performance, the two different versions from Chapter 4 are fitted using the above settings for N and K. This is done for all 66 data series, and a likelihood ratio test is performed for each series individually, comparing to the same model without exogenous variables. AIC and BIC measures are also compared individually. Table B.3 shows for how many series the likelihood ratio test indicated a significant improvement when comparing to the χ^2 distribution for each model, as well as for how many the AIC/BIC was smaller for the models with exogenous variables.

Next an out-of-sample 5 day ahead forecast was made, corresponding to the user case of fitting the model and using it for one week before re-fitting it to the new data. Here the RMSE and Quasi Likelihood loss is compared, checking for how many series it was smaller when adding the exogenous variables, indicating a better forecasting performance. The results are shown in Table B.4.

Finally the QL loss and RMSE is compared for the days of particular interest, viz. the days when the exogenous variables are non-zero, to see of the performance was enhanced on these particular days. The results are shown in Table B.5.

6.2 HAR-RV

Both versions of the model are tested, modelling the square root (called "Sqrt" in the result tables) and the logarithm (called "Log") of the volatility. Both are tested with and without the exogenous variables added, using both the multiplicative and additive method, performing the same tests as above. The likelihood ratio test results and

AIC/BIC comparison are shown in Table B.6, the forecast performance comparison is shown in Table B.7, and the interesting dates comparison is shown in Table B.8. All numbers in the tables indicate for how many out of the 66 series the current test showed an improvement when adding exogenous variables.

6.3 Comparison

Finally all different models are compared, checking for each data series and fitness measure which model performed the best. The results are shown in Table B.9, where the number of data series for which each model performed the best on each measure is shown. Also an example of the execution time for each method is shown, when tested on the S&P 500 E-mini series which contains prices for 8,471 days.

Figures C.4 and C.5 show an example of the data used, and the corresponding volatility output for two of the best performing models; fixed window GARCH-MIDAS with the multiplicative exogenous variables, and the logarithmic version of HAR-RV also with multiplicative exogenous variables. Figure C.6 shows the five-day forecast for the above mentioned models, with and without the exogenous variables. This is compared to the realized squared innovations and one-day Yang-Zhang measure.

7 Conclusions

Looking at the addition of the exogenous variables, a few observations can be made. First of all the multiplicative procedure described in Section 4.1 seems to perform better in general than the additive one in Section 4.2. This is seen in particular for the GARCH-MIDAS model, where the in-sample fit is improved the most for the multiplicative version, whereas the difference is less evident for the HAR-RV model. Both models see a significant improvement of in-sample fitness for between 67% and 87% of the data series when looking at the likelihood ratio test and AIC, as well as the interesting dates comparison. The improvement is not seen in the BIC measure, which due to the addition of 39 extra variables gets lower. The best performing model when looking at in-sample fit is the fixed window GARCH-MIDAS model with multiplicative exogenous variables.

When looking at the out-of-sample forecast performance, the improvement when adding the exogenous variables is not that great—for most models less than half of the series see an improvement. The best performing model in this case, both in regards to the improvement compared to without exogenous variables and the overall comparison, is the logarithmic HAR-RV model.

Looking at an example of the forecasting output for these two models in Figure C.6, some interesting things can be seen. First of all the HAR-RV model with exogenous variables displays a quite impressive likeness to the actual measured volatility, and also the improvement compared to the standard model is obvious. For the GARCH-MIDAS model the fit is not as good, but the difference compared to the standard version is interesting. Here we see that the exogenous variables, which are non-zero for all of the days in question, cause the forecast to change more between each day compared to the standard model which is nearly flat, making for a more interesting model that could be said to be more similar to the observed values in shape, even if not in size. This is the case in general when looking at other series as well, and in this regard the models are improved. As said before, the "real" volatility cannot be observed, and thus comparisons to volatility proxies may not tell the whole picture. Given that the goal was to incorporate the exogenous variables in a clean way into the model, this goal can be said to have been achieved.

In Figure C.5 the volatility outputs for the entire series is shown. Here we see that both models are similar overall, with the biggest difference that the HAR-RV model could be said to have a larger volatility in itself, in that it has a larger variation from day to day. The 505 first days of the GARCH-MIDAS model are flat due to the fact that this is how many days are needed for the MIDAS part, so the actual model output starts after that. Looking close at the HAR-RV model some missing values can be seen in the first half of the series, which are in fact zero. The cause of this is not perfectly clear, but is related to missing values in the input data that are filled in to make the model work.

One interesting fact about the GARCH-MIDAS model, which is also noted in Engle et al. (2008), is that the performance when using a fixed window instead of a rolling is not decreased, and in some cases it is even increased. This is a good thing when looking at the execution time in Table B.9, which is three times greater for the rolling window version. Here we can also see that the HAR-RV is faster overall compared to GARCH-MIDAS, but for the best versions the difference is not that big. Note that

this is just an example for one series; the execution time can vary a lot depending on how fast the parameter optimization is for a certain series, but overall I would say that that the relationship between the times for the different models is representative for the general performance.

7.1 Further research

Using more high frequency data for the realized volatility measures and for the volatility proxy when fitting the models would be the next step to test. Also the likelihood function could be changed to reflect the non-Gaussian distribution of returns, as in e.g. Corsi et al. (2008) the normal inverse Gaussian distribution is used.

The forecasting performance is what would be most interesting to continue to improve upon; here further testing could be done by testing the performance for other days than the last ones available as is done now, and different measures of fit could be tested. Also the practical consideration that it would be worse, in a risk perspective, to predict a too low volatility than a too high could be implemented by some weighted measure.

References

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

Bomfin, A. N. (2003). Pre-announcement effects, news effects, and volatility: Monetary policy and the stock market. *Journal of Banking and Finance*, *27*(1), 133–151.

Campbell, J. Y. (1991). A variance decomposition for stock returns. *Economic Journal*, *101*(405), 157–179.

Campbell, J. Y., & Shiller, R. J. (1988, July). Stock prices, earnings, and expected dividends. *The Journal of Finance*, *43*(3), 661–671.

Corsi, F., Mittnik, S., Pigorsch, C., & Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews*, *27*, 46–78.

Engle, R. F. (1982, July). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, *50*(4), 987–1007.

Engle, R. F., Ghysels, E., & Sohn, B. (2008, August). On the economic sources of stock market volatility. *AFA 2008 New Orleans Meetings Paper*.

Engle, R. F., & Rangel, J. G. (2008). The spline-garch model for low frequency volatility and its global macroeconomic causes (2004). *Review of Financial Studies*, *21*(405).

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

Rogers, L. C. G., & Satchell, S. E. (1991). Estimating variance from high, low and closing prices. *Annals of Applied Probability*, *1*(4), 504-512.

Yang, D., & Zhang, Q. (2000, July). Drift independent volatility estimation based on high, low, open, and close prices. *The Journal of Business*, *73*(3).

A Option pricing

The pay-off, that is the amount you can earn by exercising the option, for a European call option is given by

$$\varphi_T = (S_T - K)\,\mathbb{1}_{\{S_T > K\}}$$

where $K > 0$ is the strike price, S_T the price of the underlying asset at time of maturity T. Using the risk-neutral valuation framework, the price at time t is given by

$$F(t,s) = \mathrm{E}_{\mathbb{Q}}^{t,s}\left[e^{-r(T-t)}\varphi_T\right]$$

under the pricing measure $\mathbb{Q}$, where

$$\begin{cases} dS_u = rS_u dt + \sigma S_u dW_u \\ S_t = s \end{cases}$$

and W_t is a Brownian motion under $\mathbb{Q}$. The solution to the stochastic differential equation of this geometric Brownian motion is

$$S_u = se^{(r-\frac{\sigma^2}{2})(u-t)+\sigma(W_u - W_t)}$$

Thus we have that

$$\begin{aligned} F(t,s) &= \mathrm{E}_{\mathbb{Q}}^{t,s}\left[e^{-r(T-t)}\varphi_T\right] = \mathrm{E}_{\mathbb{Q}}^{t,s}\left[e^{-r(T-t)}\,(S_T - K)\,\mathbb{1}_{\{S_T>K\}}\right] \\ &= \mathrm{E}_{\mathbb{Q}}^{t,s}\left[e^{-r(T-t)}\left(se^{\left(r-\frac{\sigma^2}{2}\right)(T-t)+\sigma(W_T-W_t)} - K\right)\mathbb{1}_{\{S_T>K\}}\right] \\ &= \left[W_T - W_t = X\cdot\sqrt{T-t},\ \text{where } X\sim\mathcal{N}(0,1)\right] \\ &= \int_{-\infty}^{\infty} e^{-r(T-t)}\left(se^{\left(r-\frac{\sigma^2}{2}\right)(T-t)+\sigma\sqrt{T-t}\,x} - K\right)\mathbb{1}_{\{S_T>K\}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx \end{aligned}$$

Because of the indicator function, the integral is only non-zero from where $S_T > K$, that is

$$se^{\left(r-\frac{\sigma^2}{2}\right)(T-t)+\sigma\sqrt{T-t}\,x} > K \quad\Longrightarrow\quad x > -\frac{\log\frac{s}{K} + \left(r-\frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} = -d_2$$

This gives the lower limit of integration, and the calculations above can continue as follows:

$$F(t,s) = \int_{-d_2}^{\infty} e^{-r(T-t)}\left(se^{\left(r-\frac{\sigma^2}{2}\right)(T-t)+\sigma\sqrt{T-t}\,x} - K\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx$$

$$= \underbrace{\int_{-d_2}^{\infty} s e^{-r(T-t)+\left(r-\frac{\sigma^2}{2}\right)(T-t)+\sigma\sqrt{T-t}\,x-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}}\,dx}_{=I_1} - \underbrace{K e^{-r(T-t)} \int_{-d_2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx}_{=I_2}$$

Where

$$I_1 = \int_{-d_2}^{\infty} s e^{-r(T-t)+\left(r-\frac{\sigma^2}{2}\right)(T-t)-\frac{(x-\sigma\sqrt{T-t})^2}{2}+\frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi}}\,dx$$

$$= s \int_{-d_2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\sigma\sqrt{T-t})^2}{2}}\,dx = \left[z = x - \sigma\sqrt{T-t},\, dz = dx\right]$$

$$= s \int_{-d_2-\sigma\sqrt{T-t}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz = \left[-d_1 = -d_2 - \sigma\sqrt{T-t}\right]$$

$$= s\,\mathbb{P}\left(Z \geq -d_1\right) = s\mathcal{N}(d_1)$$

and

$$I_2 = K e^{-r(T-t)}\,\mathbb{P}\left(X \geq -d_2\right) = K e^{-r(T-t)}\,\mathcal{N}(d_2)$$

Thus we have that

$$F(t,s) = I_1 - I_2 = s\mathcal{N}(d_1) - K e^{-r(T-t)}\,\mathcal{N}(d_2)$$

where

$$d_2 = \frac{\log\frac{s}{K} + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}$$

$$d_1 = d_2 + \sigma\sqrt{T-t}$$

or

$$d_1 = \frac{\log\left(\frac{s}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T-t}$$

B Tables

B.1 Data

Name	Price_{C}	$\text{Price}_{\text{OHLC}}$
AMSTERDAM IDX FUT	21-Jun-1984	02-Jan-1989
CAC40 10 EURO FUT	09-Jul-1987	07-Dec-1988
DAX INDEX FUTURE	22-Jun-1984	23-Nov-1990
OBX INDEX FUTURE	02-Jan-1996	13-Aug-1997
EURO STOXX 50	31-Dec-1986	22-Jun-1998
HANG SENG IDX FUT	04-Oct-1984	01-Apr-1992
FTSE/MIB IDX FUT	31-Dec-1997	22-Mar-2004
NASDAQ 100 E-MINI	04-Feb-1985	21-Jun-1999
NIKKEI 225 (OSE)	11-Oct-1984	05-Sep-1988
OMXS30 IND FUTURE	18-Dec-1986	14-Feb-2005
Russell 2000 Mini	21-Jun-1984	17-Aug-2007
S&P 500 E-mini	21-Jun-1984	09-Sep-1997
SPI 200 FUTURES	29-May-1992	02-May-2000
MSCI TAIWAN INDEX	21-Jan-1997	21-Jan-1997
KOSPI2 INX FUT	03-Jan-1990	03-May-1996
FTSE CHINA A50	21-Jul-2003	04-Jan-2007
BOVESPA INDEX FUT	21-Dec-1989	12-Jul-1995
MEX BOLSA IDX FUT	19-Jan-1994	03-May-1999
IBEX 35 INDX FUTR	05-Jan-1987	03-Aug-1992
SWISS MKT IX FUTR	01-Jul-1988	01-Oct-1998
FTSE/JSE TOP 40	30-Jun-1995	03-Jul-1995
MSCI SING IX ETS	07-Sep-1998	07-Sep-1998
SGX Nifty 50	03-Jul-1990	25-Sep-2000
SET50 FUTURES	16-Aug-1995	28-Apr-2006
BIST 30 FUTURES	02-Jan-1997	05-Oct-2005
US 2YR NOTE (CBT)	25-Jun-1990	25-Jun-1990
US 5YR NOTE (CBT)	20-May-1988	25-May-1988
US 10YR NOTE (CBT)Sep16	06-Jan-1986	06-Jan-1986
US LONG BOND(CBT)	21-Jun-1984	21-Jun-1984
JPN 10Y BOND(OSE)	04-Feb-1986	04-Feb-1986
EURO-BUND FUTURE	01-Jul-1991	01-Jul-1991
EURO-BOBL FUTURE	04-Oct-1991	04-Oct-1991
AUDUSD Crncy Fut	21-Jun-1984	13-Jan-1987
CAD CURRENCY FUT	07-Nov-1984	04-Apr-1986
EURO FX CURR FUT	28-Aug-1984	19-May-1998
BP CURRENCY FUT	21-Jun-1984	27-May-1986
JPN YEN CURR FUT	21-Jun-1984	22-May-1986
NZD FUT	07-Nov-1984	07-May-1997

Name	**Price$_{\mathbf{C}}$**	**Price$_{\mathbf{OHLC}}$**
ICE US Cocoa futures	01-Oct-1985	01-Oct-1985
COTTON NO.2 FUTR	01-Apr-1986	01-Apr-1986
COFFEE C FUTURE	01-Oct-1985	01-Oct-1985
FCOJ-A FUTURE	14-Oct-1985	14-Oct-1985
SUGAR #11 (WORLD)	01-Oct-1985	01-Oct-1985
COFF ROBUSTA 10tn	30-Aug-1991	16-Jan-2008
CATTLE FEEDER FUT	17-Mar-1986	17-Mar-1986
LIVE CATTLE FUTR	13-Jan-1986	13-Jan-1986
LEAN HOGS FUTURE	01-Apr-1986	01-Apr-1986
CORN FUTURE	20-Sep-1985	20-Sep-1985
Oat Future	01-Aug-1996	01-Aug-1996
ROUGH RICE (CBOT)	11-Jun-1992	11-Jun-1992
SOYBEAN FUTURE	25-Sep-1985	25-Sep-1985
SOYBEAN OIL FUTR	21-Jun-1984	30-Sep-2002
Soybean Meal	21-Jun-1984	23-Oct-1985
WHEAT FUTURE(CBT)	16-Jan-1986	16-Jan-1986
GOLD 100 OZ FUTR	28-Aug-1984	08-Apr-1985
Silver	30-Apr-1985	01-Apr-1986
Copper CME	01-Apr-1986	06-Dec-1988
PALLADIUM FUTURE	01-Apr-1986	01-Apr-1986
PLATINUM FUTURE	01-Apr-1986	01-Apr-1986
WTI CRUDE FUTURE	10-Oct-1984	01-Apr-1986
BRENT CRUDE FUTR	02-Nov-1984	01-Sep-1989
NY Harb ULSD Fut	10-Oct-1984	04-Apr-1986
NATURAL GAS FUTR	03-Apr-1990	03-Apr-1990
Low Su Gasoil G	03-Jul-1989	03-Jul-1989
GASOLINE RBOB FUT	04-Nov-2003	08-Mar-2006
MILL WHEAT EURO	04-Jan-1999	04-Jan-1999

Table B.1: Data used

B.2 GARCH-MIDAS

N	K	Fixed	Rolling
5	25	10	8
5	50	7	6
5	101	**32**	**29**
21	6	1	1
21	12	4	0
21	24	3	4
63	2	0	1
63	4	2	1
63	8	4	3
126	1	0	3
126	2	0	0
126	4	3	5
252	1	0	1
252	2	0	4

Table B.2: Number of series where certain choice of N, K performs best

	Multiplicative		Additive	
	Fixed	Rolling	Fixed	Rolling
Significant likelihood ratio test	58	58	29	27
AIC smaller	52	54	24	16
BIC smaller	5	3	2	0

Table B.3: In-sample comparison when adding exogenous variables

	Multiplicative		Additive	
Test	Fixed	Rolling	Fixed	Rolling
QL smaller	20	20	26	29
RMSE smaller	18	17	20	18

Table B.4: Out-of-sample forecast comparison when adding exogenous variables

	Multiplicative		Additive	
Test	Fixed	Rolling	Fixed	Rolling
QL smaller	59	58	58	56
RMSE smaller	43	45	51	50

Table B.5: Interesting dates comparison when adding exogenous variables

B.3 HAR-RV

	Multiplicative		Additive	
	Sqrt	Log	Sqrt	Log
Significant likelihood ratio test	49	50	42	42
AIC smaller	42	41	43	42
BIC smaller	3	5	21	10

Table B.6: In-sample comparison when adding exogenous variables

	Multiplicative		Additive	
Test	Sqrt	Log	Sqrt	Log
QL smaller	16	35	27	23
RMSE smaller	13	38	27	24

Table B.7: Out-of-sample forecast comparison when adding exogenous variables

	Multiplicative		Additive	
Test	Sqrt	Log	Sqrt	Log
QL smaller	52	36	47	22
RMSE smaller	43	59	46	25

Table B.8: Interesting dates comparison when adding exogenous variables

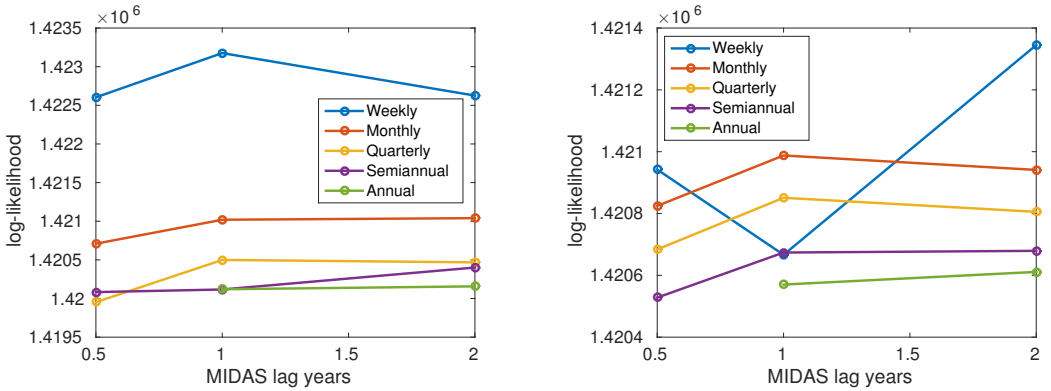
B.4 Comparison

		Log-likelihood	AIC	BIC	QL forecast	RMSE forecast	QL interesting	RMSE interesting	Time (s)
GARCH-MIDAS	Fixed	2	5	20	6	9	3	5	1.49
	Rolling	0	2	11	8	7	1	1	4.64
	Fix. ex. mult.	24	12	1	4	4	20	10	12.98
	Roll. ex. mult.	13	15	1	3	4	13	7	39.81
	Fix. ex. add.	4	1	0	2	2	6	7	11.23
	Roll. ex. add.	0	1	0	4	3	2	1	26.55
HAR-RV	Sqrt	0	3	19	4	5	0	3	1.16
	Log	0	0	0	13	8	8	0	1.37
	Sqrt ex. mult.	22	15	1	1	2	1	17	5.34
	Log ex. mult.	0	0	0	13	14	9	0	10.80
	Sqrt ex. add.	1	12	13	6	3	0	15	13.89
	Log ex. add.	0	0	0	2	5	3	0	6.99

Table B.9: Comparison between all different models, with the number of series where each model performed the best for each measure. Also an example of execution time for the longest series available, taken as an average over three runs.

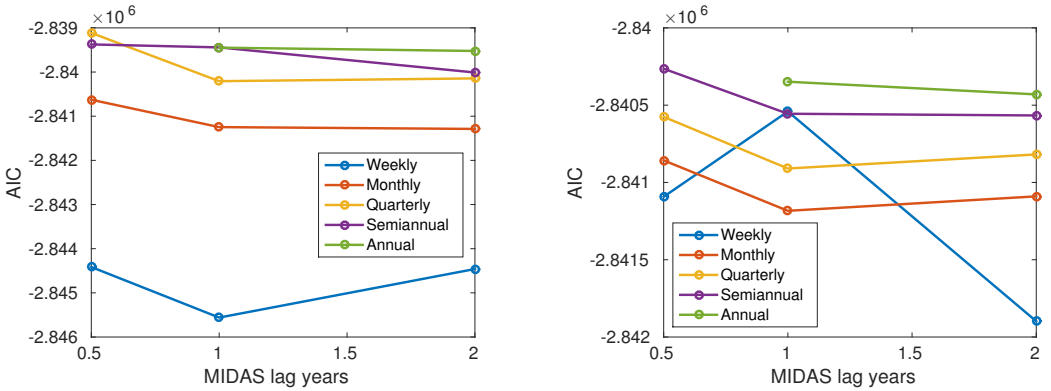
C Plots

C.1 GARCH-MIDAS



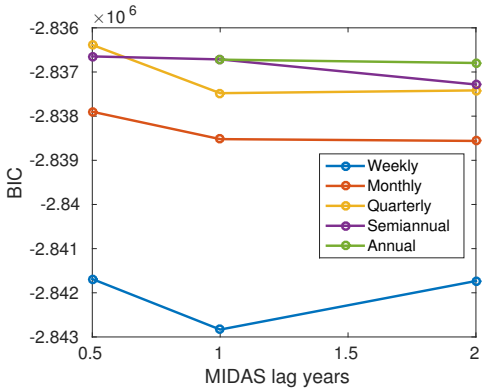
(a) Fixed window (b) Rolling window

Figure C.1: Sum of log-likelihoods for all series

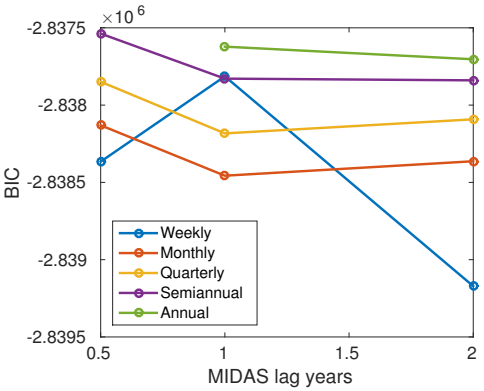


(a) Fixed window (b) Rolling window

Figure C.2: Sum of AIC for all series



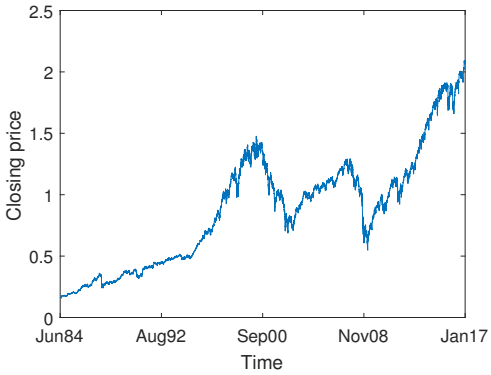
(a) Fixed window



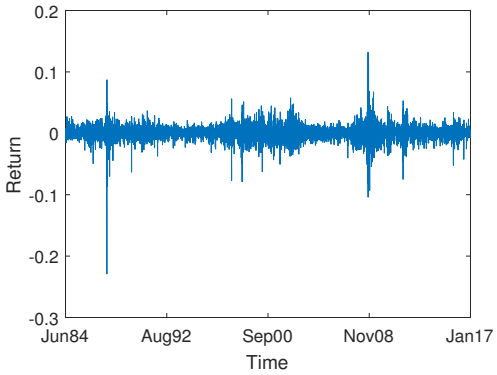
(b) Rolling window

Figure C.3: Sum of BIC for all series

C.2 Comparison

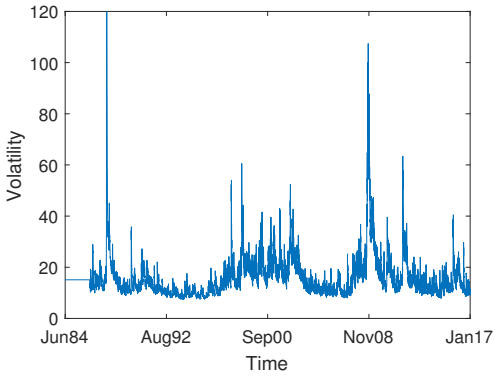


(a) Closing prices

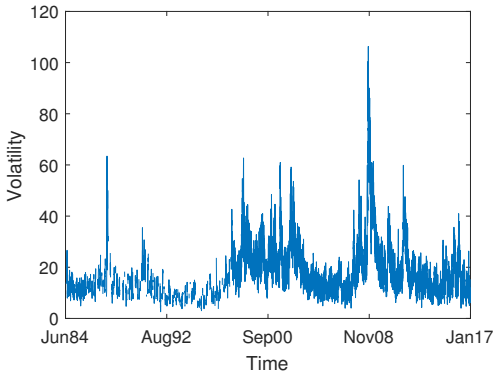


(b) Returns

Figure C.4: Example data: S&P 500 E-mini

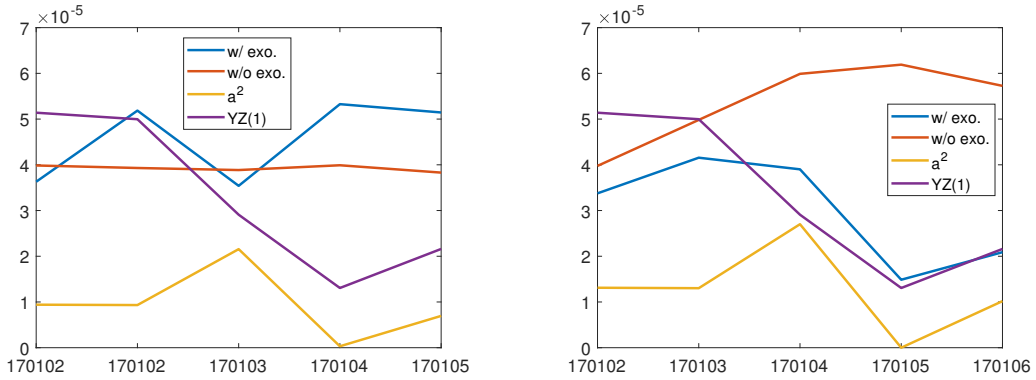


(a) GARCH-MIDAS



(b) HAR-RV

Figure C.5: Example volatility output for S&P 500 E-mini data



(a) GARCH-MIDAS

(b) HAR-RV

Figure C.6: Comparison of forecast volatility with and without exogenous variables, compared to the actual squared innovations and one-day Yang-Zhang RV.