UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

MIDAS and GARCH: A comparison of predictive ability using real world data

DEPARTMENT OF ECONOMICS

BACHELOR'S THESIS IN STATISTICS

Author:
Robin SÄRNÅ

Supervisor:
Dr. Mattias SUNDÉN

June 9, 2017

Abstract

I compare GARCH and MIDAS one-day-ahead forecasts of volatility using high frequency data from the CRSP U.S. Mega Cap Index. The MIDAS models are estimated using high frequency data sampled at 5, 15 and 30 minute intervals and estimated using both exponential Almon and beta lag distributions with two shape parameters. The GARCH(1,1) model with a skewed t-distribution is the benchmark model to which the MIDAS models are compared. The study finds that MIDAS models have superior predictive ability in volatility spikes due to its ability to incorporate high frequency data and that the GARCH model is more prone to underestimate volatility but is able to produce smaller forecast errors during calm periods. The MIDAS models using data sampled at a frequency of 5 minutes perform poorly suggesting that high frequency noise plays an important role when sampling at this frequency. Sampling frequency appears to be more important than lag length when deciding on which MIDAS model to use.

Keywords: MIDAS, GARCH, high frequency data

Contents

1 Introduction

The aim of this thesis is to test and compare GARCH and MIDAS one-day-ahead forecasts of volatility using high frequency data from the CRSP U.S. Mega Cap Index. The MIDAS models are estimated using high frequency data sampled at 5, 15 and 30 minute intervals and estimated using both exponential Almon and beta lag distributions with two shape parameters. Before estimating the models I employ the augmented Dickey-Fuller test to check for the presence of a unit root in the data, Ljung-Box test to test for autocorrelation and two CUSUM tests to test for structural breaks in mean and variance. The estimated models are finally scored using several different measures to identify potential strengths and weaknesses in the forecasts.

Volatility modeling is of interest to both academia and the practitioner given the key role it plays in option pricing, hedging, VaR-calculations, other risk measures and asset pricing. Testing the predictive ability of MIDAS models using high frequency data is especially interesting given the relatively large literature on high frequency data that has emerged over the last one or two decades. As GARCH is a common and well researched methodology for forecasting volatility this is used as the benchmark model to which the MIDAS models performance is compared.

Emre Alper et al (2012) evaluates weekly out of sample volatility forecasts using MIDAS and GARCH(1,1) and find that MIDAS produces significantly smaller mean squared prediction error during market stress due to its ability to incorporate additional information provided by the high frequency data. They also find that shocks to volatility in weekly GARCH models tend to persist for a longer time increasing the error after a volatility spike whereas the shocks in the MIDAS models die out more quickly. Çelik and Ergin (2013) use high frequency data from the Turkish stock market and find that MIDAS is superior at forecasting volatility during financial crisis when compared to GARCH models.

This study adds to previous studies in that it is using a different dataset in combination with high frequency data sampled at several frequencies for comparison. I also use a more varied approach to forecast evaluation, previously used by Hansen and Lunde (2005), to better identify the strengths and weaknesses of each model. The data used in this paper is collected over the relatively calmer period between 2012 through 2016 whereas the above studies include the turmoil generated by the 2007 to 2008 financial crisis. This study on the other hand is affected by the flash crash of August 24th 2015 that generated a temporary spike in volatility preceded and followed by periods of relatively low volatility. This temporary increase in volatility is however much different from the protracted kind of volatility that was present during the 2007 to 2008 financial crisis and presents a unique angle to this study.

2 Method

2.1 MIDAS

Mixed frequency data is common in economics and many macroeconomic variables are reported at different intervals. Employment for example is often reported monthly while GDP is reported quarterly. In contrast stock prices are reported on an almost continuous basis when the trading window is open. Even if variables are sampled at different frequencies there can be a strong link between them and the more frequently sampled variable can provide important information when forecasting the less frequently sampled variable (Armesto, Engemann and Owyang, 2010). Various techniques for dealing with this type of forecasting exists. One way is to scale back the data, through for example averaging, so that the number of observations in the dependent and independent variable match at the cost of informational value in the high frequency data. Another approach can be to use some weighting scheme that is either intuitive or has to be parametrized. Unfortunately, all of these approaches suffer from either large estimation errors or require a large number of parameters to be estimated. If the number of estimated parameters is too large the estimated model becomes useless. This is what Ghysels, Santa-Clara and Valkanov (2004) refers to as parameter proliferation and can be solved using the MIDAS methodology.

The MIDAS model was introduced by Ghysels, Santa-Clara and Valkanov (2004) and the main area of application is generally within mixed frequency data. It is a model that has been developed for the specific purpose of creating regression models where the input data is sampled at a higher frequency than the examined variable. This makes MIDAS modeling very useful in volatility forecasting when using high frequency data to forecast daily variance. MIDAS does not use an autoregressive methodology which means that regressors different than the regressand can be used and that variations of return other than volatility estimates can be used as regressors. This makes for a more flexible model with regards to the independent variable when compared to ARCH/GARCH models that require some lagged volatility measure when forecasting volatility. Since the procedure is exactly the same independently of the type of regressand the MIDAS methodology can easily be used to compare different independent variables for predictive ability. (Ghysels et al, 2006)

The MIDAS model is in many ways similar to a distributed lag model but instead of having fixed weights or weights that are individually estimated the weights are parameterized which reduces the researchers need to know the theoretical weights while still being able to avoid the problem of parameter proliferation. In other words, it allows for more information to be accounted for in the model while retaining a relatively parsimonious model (Ghysels, Santa-Clara and Valkanov, 2004).

Ghysels, Santa-Clara and Valkanov (2004) show that MIDAS can be used with multiple regressors using a maximum likelihood approach. However, as the intention of this thesis only is to work with security returns and high frequency data I consider this to be beyond

the scope of this thesis.

There are several possible ways to parameterize the weights and Ghysels et al (2016) gives examples such as U-MIDAS, normalized beta probability density function, normalized exponential Almon lag polynomial, and polynomial specification with step functions. They further mention the Gompertz probability density function, Log-Cauchy probability density function and Nakagami probability density function. Armesto, Engemann and Owyang (2010) write that "the desired function should achieve flexibility while maintaining parsimony". The methodology to use depends on the lag structure the researcher is trying to model. Ghysels (2016) suggests a model selection process where all potential models are estimated over the estimation window and the best model is then selected using a selection criteria such as AIC (Akaike, 1973) or BIC (Schwarz, 1978).

Foroni et al (2015) assume a linear high frequency model and derive the Unrestricted-MIDAS (U-MIDAS) that easily can be estimated by OLS using frequency alignment. This is computationally the easiest MIDAS model to estimate. The downside to this approach is that it is not as parsimonious as the beta or exponential Almon weighted approach and using OLS does only work in a linear setting.

Andreou et al (2010) show that non-linear least squares (NLS) is able to parameterize the models where the lag structure is not linear. This approach works well with both exponential Almon and beta lags described below.

Ghysels et al. (2006) use a regression based method and using their original notation for a daily volatility regression MIDAS can be written as:

$$V_{t+H,t}^{(Hm)} = \mu + \Phi_H \Sigma_{k=0}^{k^{max}} b_H(k.\theta) \tilde{X}_{t-k,t-k-1}^{(m)} + \epsilon_{Ht} \tag{1}$$

Where $b_h(k, \theta)$ is some polynomial that parameterizes the weights in the model. The weights are normalized and add up to unity which allows for the estimation of the scale parameter Φ. The parameterization of the weights is a key element in MIDAS modeling as it allows for a parsimonious model that does not need the parameters to be estimated individually for every lag.

Where k_{max} is the number of steps of high frequency data to be used when predicting the steps in the forecasted series. This is also equivalent to the number of parameters that would have been estimated using a regular distributed lag model. In a MIDAS setting Ghysels and Valkanov (2012) write that the only downside to using a large k_{max} is the loss of data in the beginning of the sample when the model is estimated. The strength being that the lag structure might be able to detect correlation over longer periods of time. They suggest a conservative approach using a relatively large k_{max} and let weights die out in the parameter estimation.

The variable k is a specific step in the high frequency data between $t - 1$ and t where t measures the steps in the forecasted series.

Ghysels et al (2004) propose the same model as in the 2006 paper but also include a trend function on the lagged dependent variable. Ghysels and Valkanov (2012) specifically focus on MIDAS daily volatility forecasting using high frequency data.

From here on follows an introduction to two of the earliest and what appears to be the most common weight parameterization polynomials in the literature; the exponential Almon and beta functions. Note that both specifications are normalized to make sure weights sum to one.

Ghysels et al. (2004) and Ghysels et al. (2006) use a beta function for the weight parametrization. Using a combination of the original notation and the notation of Armesto, Engemann and Owyang (2010):

$$b_H(k; \theta) = \frac{f\left(\frac{k}{k^{\max}}, \theta_1; \theta_2\right)}{\sum_{j=1}^{k^{\max}} f\left(\frac{j}{k^{\max}}, \theta_1; \theta_2\right)} \tag{2}$$

Where $\theta = [\theta_1; \theta_2]$ and θ_1, θ_2 are shape parameters of the beta weighting function:

$$f(i, \theta) = \frac{i^{\theta_1 - 1}(1 - i)^{\theta_2 - 1}\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} \tag{3}$$

And Γ is the regular gamma function:

$$\Gamma(\theta_p) = \int_0^\infty e^{-i} i^{\theta_p - 1} di \tag{4}$$

The beta function can take on a wide range of shapes depending on the specified parameters. The beta function for a few different parameter values is plotted in figure 1 to illustrate the flexibility of this specification.

Beta Distribution
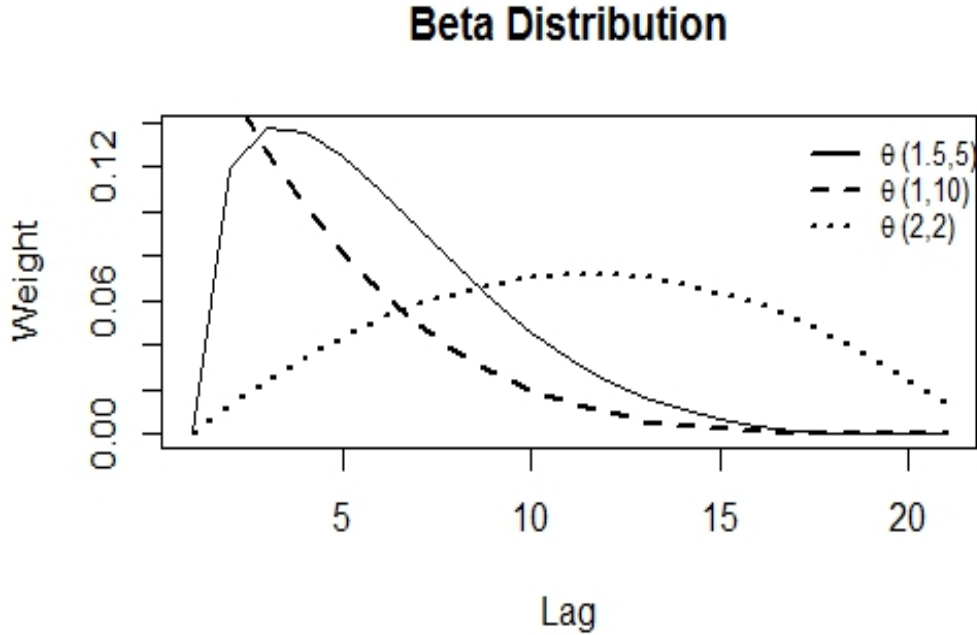
Figure 1: Various parameter values of the beta density function

Ghysels, Santa-Clara and Valkanov (2005) and Ghysels, Sinko and Valkanov (2007) use the Almon lag for parametrization:

$$b_h(k;\theta) = \frac{e^{\theta_1 k + ... + \theta_Q k^Q}}{\sum_{k=1}^{K} e^{\theta_1 k + ... + \theta_Q k^Q}} \tag{5}$$

Where k is a step in the explanatory high frequency series and Q is the number of parameters to be estimated. A common and simple approach, such as the one used by Armesto, Engemann and Owyang (2010) is to only use two shape parameters θ.

Otherwise the notation is the same as in the beta function above. The exponential Almon function with two shape parameters is also a flexible specification with similarities to the beta function.

9

The exponential Almon function for a few different parameter values is plotted in figure 2.

Exponential Almon Distribution

Figure 2: Various parameter values of the exponential Almon density function

Armesto, Engemann and Owyang (2010) compare different forecasting methods for mixed frequency data and find that in some cases MIDAS provides an ad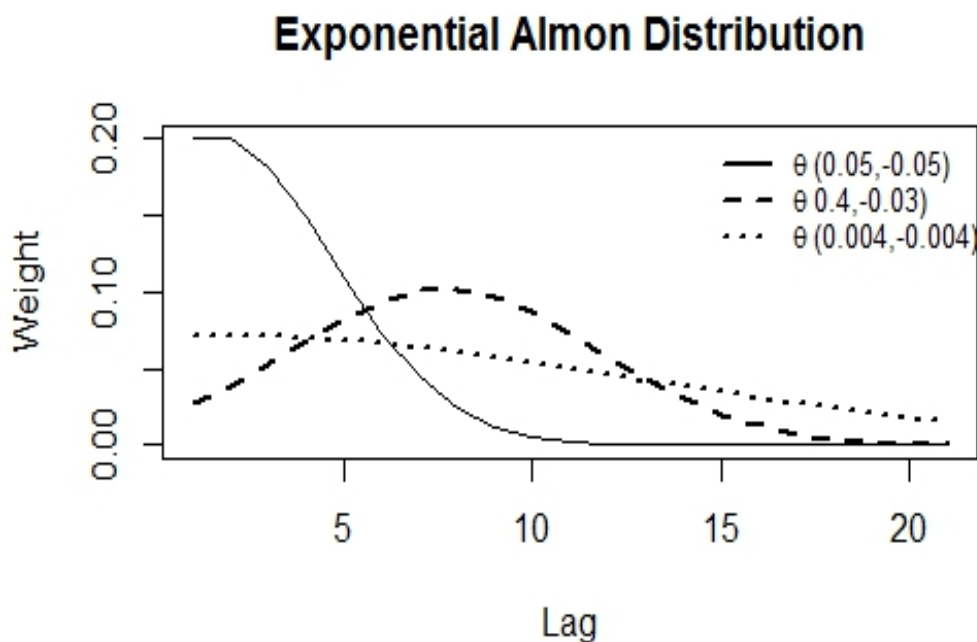vantage but that in some cases it does not improve the results compared to averaging the higher frequency data to match that of the less frequently sampled dataset.

Ghysels et al (2009) find that MIDAS models perform better than GARCH models in 10-day ahead forecasts of stock variance but that they are outperformed in 1-day ahead forecasts.

Ghysels et al (2006) use prediction horizons of 1 day and 1-4 weeks for their volatility forecasts using S&P 500 data. They chose this forecast horizon as it captures most of the relevant cases for practitioners and use 50 lags as this appears to be the optimal lag length for capturing the variance structure. They find that using data sampled at 5 minutes does not improve forecasts compared to data being sampled at a lower frequency. Emre Alper et al (2012) also used a 50 day estimation window for their MIDAS model in their comparisons against GARCH models.

I estimate the MIDAS models using the R plugin "rmidasr" developed by and well documented in Ghysels, Kvedaras and Zemlys (2016). I decide to use 1-day, 5-day, 20-day and

50-day estimation windows. I chose these estimation windows as they correspond to a single trading day, trading week and trading month. The 50-day estimation window is included as is it is a lag length that has previously been used for the purpose of forecasting variance by Ghysels et al (2006) and Emre Alper et al (2012) with favorable results. I estimate models for each estimation window using intraday variance sampled at 5, 15 and 30 minutes as independent variables and the 15 minutes realized variance as dependent variable. All models are estimated in two sets, one using a normalized beta lag structure and the other using a normalized exponential Almon lag structure. This generates 18 models.

2.2 GARCH

In this paper GARCH(1,1) is used as the benchmark model to which the predictive ability of the MIDAS model is compared.

The Autoregressive Conditional Heteroskedastic (ARCH) model was originally developed by Engle (1982). Before the introduction of ARCH few methods were available for modelling conditional heteroscedasticity and the methods that were available usually involved rolling estimates or weighted averages of previous heteroscedasticity (Engle, 2001) or models where the heteroscedasticity was modeled using an exogenous variable (Engle,1982).

By assuming that the dependent variable could be described by a linear combination $x\beta$ and by assuming normality in the residuals Engle (1982) showed that an AR model could be constructed to describe the variance and that the parameters for such an equation could be found using a maximum likelihood methodology. The ARCH model was further developed by Bollerslev (1986) into what is called the Generalized Autoregressive Conditional Heteroskedastic (GARCH) model. The difference from the ARCH model is that instead of using solely an AR process to describe the conditional variance Bollerslev (1986) described the conditional variance as an ARMA process. This gives the model more flexibility and allows for a better fit to many processes observed in empirical work. The GARCH model is usually best estimated using a maximum likelihood approach and just as in the ARCH model the assumption of an identifiable mean process and normally distributed residuals were made. The GARCH(1,1) process can be described as:

$$\sigma_t^2 = w + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + v \tag{6}$$

Where σ_t^2 is the conditional variance, w is an intercept, ε is a residual of a mean process (typically an ARMA process), v is an error term, β is an estimated parameter and σ_{t-1}^2 is the conditional variance in the previous period.

Enders (2015, p. 129) states that the advantage of the GARCH model to an ARCH model is that it can be modeled more parsimoniously and more exact depending on the lag structure.

The original assumption of normality in the residuals was made to allow for the maximum likelihood procedure but other distributions can also be used. As described by Rydberg (2000) and others security returns have heavier tails than the normal distribution and are often negatively skewed. To accommodate for this Engle and Bollerslev (1986) and Bollerslev (1987) extend the ARCH and GARCH methodology to allow for conditionally t-distributed errors rather than strictly assuming normality as had been done before. It also turns out that the degrees of freedom and the skewness of the t-distribution can be estimated together with the other parameters in the maximum likelihood procedure which has made the t-distribution a very flexible approach in GARCH modeling.

GARCH models require the sum of the parameters α and β to be less than one and that the model parameters are positive to ensure that the conditional variance is strictly positive. Setting the parameter sum equal to one generates the more parsimonious IGARCH model introduced by Engle and Bollerslev (1986). Nelson (1990) provides a detailed explanation of the stationarity requirements and persistence of the GARCH(1,1) model.

A good way to test if the model has been correctly specified is to look at the remaining autocorrelation within the squared residuals. If the autocorrelation has disappeared it can be concluded that the ARCH or GARCH effect has been captured to a satisfying degree. Formally this is usually done using a Ljung-Box test with 15 lags (Engle, 2001).

Given the importance the conditional mean plays in estimating the model it is important that a researcher looks for structural breaks in the data according to Enders (2015).

Enders (2015) warn researchers of deciding on estimators purely using AIC (Akaike, 1973) and BIC (Schwarz, 1978) as guidelines since there is a substantial risk of overfitting the model to the data. AIC and BIC are two commonly used informational criteria used to evaluate statistical models and additional information can be found in respective paper. Using an informational criteria risks creating a model that does not match the data generating process and is only representative of the sample at hand. Worse, if the model is used to forecast out of sample the specification might be incorrect. The main focus should be on statistically significant parameters and parameters that are consistent with theory describing the data.

Hansen and Lunde (2005) compared 16 different GARCH models and concluded that GARCH(1,1) was the best model for forecasting out of sample conditional volatility in datasets with deutsche mark - dollar exchange rates but that other models performed better than GARCH(1,1) when forecasting the conditional volatility of IBM stocks out of sample.

Hwang and Pereira (2006) found that GARCH models need to be estimated on at least 500 data points or else the model is likely to be biased downwards.

A downside to the GARCH approach is that it is an autoregressive process meaning that it solely relies on historical data from the same time series. This means that it cannot incorporate information from other independent variables or fully integrate information sampled at a higher frequency.

The models are estimated and the forecasts are created using the ruGARCH package for R. I use this package as it has a good functionality when it comes to producing rolling forecasts with a moving estimation window. For this modeling I use the daily return data to forecast the conditional variance. I estimate three ARMA(0,0)-GARCH(1,1) models using Gaussian, t-distributed and skewed t-distributed conditional distributions using ML-methodology.

2.3   Data processing methodology

2.3.1   Augmented Dickey-Fuller test

It is often considered that the efficient market hypothesis hold to some degree and that prices follow some type of non-stationary process as explained by Enders (2015, p. 184-187). The unit-root process will not allow itself to be modelled in a reliable way as the process is driven by the sum of the errors that do not disperse over time. As time pass the variance will only increase and the time series will never converge to a forecastable value. Luckily, the unit-root process can easily be differenced into a stationary process as this cancels out the error terms and makes the new process possible to model. If more than one unit root is present, the time series should be differenced an equal number of times to the number of unit roots to become stationary. The stationarity is essential if we are to be able to model the time series and this is the rationale for using returns rather than the asset price itself. To test if we have any unit root remaining we can use the Dickey-Fuller (DF) test, which is explained below. According to Enders (2015, p. 189-192) the modern view is that macro-economic data generally contain stochastic trends, meaning that the trend is best removed by differencing rather than some other method (such as regression) to detrend the data. This is supported by Nelson and Plosser (1982) who suggest that macroeconomics variables are difference stationary.

The DF test was introduced by Dickey and Fuller (1979) as a way to test for the presence of a unit root. The problem with testing for a unit root lay in the characteristics of the autocorrelation function of a non-stationary process that makes the test biased when using the ordinary t or F-statistic (Enders, 2015).

The DF test is done by calculating one of three specified test statistics: τ_1, τ_2 or τ_3. The test statistic to be chosen depends on if we have a pure random walk process, a random walk with drift or if we both have a drift and linear trend (Ender, 2015).

The test for a pure random walk process is performed as described by Engle (2015) based on Dickey and Fuller (1979).

Regress:

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t \tag{7}$$

Where $\gamma = a_1 - 1$ and a_1 is the potential unit root under examination making the null hypothesis $\gamma = 0$. As such, if γ is 0 we have a unit root present. The estimated value γ from the regression is normalized by dividing it by the standard deviation similar to an ordinary t-test. As the t-statistic here is biased Dickey and Fuller (1979) derived the unbiased DF statistic that is used for diagnostics of this test. The DF statistics can be found in the Dickey and Fuller (1979) paper. The null hypothesis of the DF test is that the process is non-stationary, i.e. a unit root is present.

Above is the procedure for a pure random walk model. Enders (2015) explains that if the examined data generating process has an intercept (called drift term) the model should be defined as:

$$\Delta y_t = a_0 + \gamma y_{t-1} + \varepsilon_t \tag{8}$$

If there is a linear time trend in the data this should be accounted for and the model should be specified as:

$$\Delta y_t = a_0 + \gamma y_{t-1} + a_2 t + \varepsilon_t \tag{9}$$

Depending on the specification the test statistic will vary and it is therefore important to pick the correct specification and corresponding test statistic. Díckey and Fuller (1981) show how to use an F-test to select the appropriate specification using their own derived test statistic.

The F-test is simply constructed as:

$$\Phi_i = \frac{[SSR(restricted) - SSR(unrestricted)/r]}{SSR(unrestricted)/(T-k)} \tag{10}$$

Where $SSR(restricted)$ is the sum of squared residuals in the restricted model, $SSR(unrestricted)$ is the sum of squared residuals in the unrestricted model, r is the number of restrictions (the difference in independent variables), T is the number of observations in the data and k is the number of parameters estimated in the unrestricted model.

Φ_i is the Dickey and Fuller (1981) derived test statistic for the specific models being compared. Φ_3 is for testing between trend and drift, Φ_2 for testing between trend and random walk and finally Φ_1 is for testing between drift and random walk.

The different null hypotheses are defined as:

$$H_0 \text{ for } \Phi_3 : \gamma = a_2 = 0 \tag{11}$$

$$H_0 \text{ for } \Phi_2 : a_0 = \gamma = a_2 = 0 \tag{12}$$

$$H_0 \text{ for } \Phi_1 : a_0 = \gamma = 0 \tag{13}$$

The F-test tests the null hypothesis that the model only contains the parameters of the restricted model.

A further development of the DF-test was done by Dickey (1984) to include more lags in an autoregressive process. Using Enders (2015) notation the tests are calculated as:

$$\Delta y_t = \gamma y_{y-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \tag{14}$$

$$\Delta y_t = a_0 + \gamma y_{y-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \tag{15}$$

$$\Delta y_t = a_0 + a_2 t + \gamma y_{y-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \varepsilon_t \tag{16}$$

Where

$$\gamma = -(1 - \sum_{i=1}^{p} a_i) \text{ and } \beta_i = \sum_{j=i}^{p} a_j \tag{17}$$

And p is the lag order of the autoregressive process and a are the autoregressive parameters.

All three tests use the same test statistic as the corresponding (random walk, drift or trend) test statistic in the original Dickey-Fuller test. The F-statistics developed by Dickey and Fuller (1981) for parameter selection also apply to this specification. Adding too many lags reduce the power of the test while adding to few risks bias the test due to correlation between the error term and dependent variables. Enders (2015, p. 216) suggests using a Ljung-Box test and correlogram to examine the residuals when determining lag-length. He also suggests that using AIC, SBC, F-tests or t-tests can be used as guidance to find the correct number of lags. For non-seasonal data he suggests starting with a longer lag length and then reduce it while checking for significance in the last lag. The t-tests and F-tests are possible due to the fact that all the lags are t-distributed. According to Enders (2015) Monte Carlo experiments

have shown that starting with a low number of lags and then increase them while checking for significance is biased towards selecting a model with too few lags.

Enders (2015) points out that it is important to look for structural breaks as Dickey-Fuller tests become biased toward the non-rejection of a unit root when structural breaks are present. This was also shown by Perron (1989) using Monte Carlo simulations.

2.3.2 Ljung-Box test

The Ljung-Box test was proposed by Box and Ljung (1978) as a test of autocorrelation in a time series. The test measures the autocorrelation over m lags and the result is then compared against a predefined test statistics developed by Box and Ljung (1978).

The observed Q statistic is calculated as:

$$Q = n(n+2) \sum_{k=1}^{m} \frac{r_k^2}{n-k} \tag{18}$$

Where n is the total amount of observations, r^2 is the autocorrelation at lag k and m is the number of lags tested.

2.3.3 Structural breaks

There exists many tests for testing for structural breaks. Many, such as the commonly used Chow-test proposed by Chow (1960), has the drawback that it can only test for a single known structural break in the time series and are not specifically constructed for testing structural breaks in the variance. Many tests are also not able to test for endogenous breaks. An endogenous break is a break that takes place at a time that is not pre-specified by the researcher. (Enders, 2015)

To test for an endogenous break the cumulative sum of squares (CUSUM) methodology developed by Brown, Durbin and Evans (1975) can be used. The idea is to create a model of the mean and then see if the error structure changes or remain constant over time. If the model is correctly specified and the error structure does not change we assume that the underlying process is constant.

The CUSUM test of the mean is conducted by first estimating a model of the mean. The n number of observations in the estimation window will vary depending on the model selected. This model is then used to conduct one-step-ahead forecasts and the residuals are saved.

Assuming that the model is correctly specified and the data generating process of the mean has not changed the sum of the residuals should not change significantly as a function of time.

Using the notation of Enders (2015, p 106) a five 5% confidence interval the confidence boundary for the CUSUM test should be constructed as:

$$\pm 0.948[(T-n)^{0.5} + 2(N-n)(T-n)^{-0.5}] \tag{19}$$

Where T is the total number of observations in the dataset
n is the first observation that is forecasted
N is the total number of forecasts made in the dataset

Inclan and Tiao (1994) introduced a CUSUM test for finding structural breaks in the variance of time series. They derive the test statistic to be used and outline the procedure to detect single and multiple structural breaks in the variance. An assumption of this test is that the variance is i.i.d.

Kim et al (2000) further develop the Inclan and Tiao (1994) methodology by expanding and adapting the test to detect a change in the conditional variance where the conditional variance is described by a GARCH(1,1) model. This effectively means the process no longer has to be i.i.d. Instead of focusing directly on the observed variance the Kim et al (2000) test is focused on observing changes in the GARCH parameters ω, α and β over time. If the parameters change this is a clear indication that there is a structural break present in the data.

Kim et al (2000) specify their hypotheses as:

$$H_0 : \theta(\omega, \alpha, \beta) \text{ remain unchanged} \tag{20}$$
$$H_a : \theta(\omega, \alpha, \beta) \text{ are not constant over the measured period} \tag{21}$$

Kim et al (2000) show that the test statistic will be a straight line that moves around zero as long as the data generating process is unchanged. If the data generating process change the test statistic have a "kink" in the straight line and the line will move away from zero. This is the result of residuals after the kink no longer canceling out as they follow different processes.

Lee et al (2004) show that this test has severe size distortions and low power. In response to this Lee et al (2004) develop a CUSUM test for testing the squared residuals of a GARCH

model rather than the GARCH parameters themselves which improved the problem with size distortion. The null hypothesis is the same as in the Kim et al (2000) test.

The test statistic to be evaluated is the largest absolute value of the cumulative residuals. The theoretical test statistic is defined, using the notation of Lee et al (2004), as:

$$T_n = \frac{1}{\sqrt{n}\tau} \max_{1 \leq k \leq n} \left| \sum_{t=1}^{k} \varepsilon_t^2 - \left(\frac{k}{n}\right) \sum_{t=1}^{n} \varepsilon_t^2 \right| \tag{22}$$

Where n is the full sample size and k is the current observation to which the cumulative sum is calculated.

However, as the real errors are unobservable one has to estimate them using the squared prediction error divided by the conditional variance estimated:

$$\hat{\varepsilon}_t^2 = (y_t - \hat{c}_t)^2 / \hat{h}_t^2 \tag{23}$$

Where y_t is the observed return, $\hat{c}_t$ is the conditional return and $\hat{h}_t^2$ is the conditional variance estimated in the GARCH model.

Assuming that the model is correctly estimated and no structural break is present this should generate a ε with zero mean and unit variance.

Where $\tau^2 = \text{Var}(\varepsilon^2)$ meaning that τ^2 simply can be estimated as:

$$\hat{\tau}^2 = \frac{1}{n} \sum_{t=1}^{n} \hat{\varepsilon}_t^4 - \left(\frac{1}{n} \sum_{t=1}^{n} \hat{\varepsilon}_t\right)^2 \tag{24}$$

Lee et al (2004) prove that the limiting distribution of this CUSUM test can be described by the supremum of a Brownian bridge. Ploberger and Krämer (1992) show that the applicable test statistic in this case can be calculated as:

$$T_{critical} = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} exp(-2j^2 T^2) \tag{25}$$

Andreou and Ghysels (2002) find that CUSUM type test perform well when examining structural breaks in GARCH models.

2.4 Forecast evaluation

Forecast evaluation is characterized by three large decisions that the researcher has to make. The first one is over what period the evaluation should take place and if it should be in or out-of-sample. In general out-of-sample forecasting is the most interesting for the practitioner as the models designed are used to forecast the future rather than the next in sample data point. The focus of this study is on out-of-sample forecasting.

Second, the researcher must also decide on what reference value to use. In volatility forecasting potential candidates are squared returns, realized volatility or implied volatility all measured over the forecasted period. When selecting the reference variance I take three primary concerns into consideration: The type of return, the measure of the return and the frequency of the measure.

Third, the researcher must also decide on what loss function to use.

These three considerations are all elaborated on in this section.

2.4.1 Returns

Andersen and Bollerslev (1998) point out that a major problem with volatility modeling, specifically in relation GARCH modelling, is that the squared returns provide a noisy proxy for the true volatility. They also show that volatility can be predicted to a more satisfying degree using realized volatility, which they define as the sum of intraday squared returns.

$$RV = \sum_{i=1}^{m} r_i^2 \tag{26}$$

Where r is some return measure and m the number of observations per day. Ghysels and Valkanov (2012) also use the same definition of realized variance when using high frequency data to forecast daily conditional variance.

Hansen and Lunde (2005) also show that using squared returns for model evaluation rather than realized returns might distort the model selection resulting in the researcher picking a non-optimal model as the best performing model. This is in line with the findings of Andersen and Bollerslev (1998) and highlights the importance of using the least noisy estimate of the realized volatility as reference value when evaluating forecast performance.

2.4.2 Log returns

Log returns, also known as compounded returns, are defined as the following in accordance with Hansen and Lunde (2005), Meucchi (2010):

$$r_t = \log\left(\frac{S_{t+1}}{S_t}\right) \tag{27}$$

Where S is the asset price and t is the observation number.

This measure is similar to ordinary linear returns if the time step is small and the volatility is not too high. When the measured time span grow or volatility increase the two methods of measuring returns will start to differ more and more with the implication that the means will start to differ and the linear returns will start to deviate from zero in the case of security returns (Meucchi (2010); Hudson and Gregoriou (2015).

Hudson and Gregoriou (2015) state several reasons why logarithmic returns are useful. The most important are that they become additive, they prevent security prices from becoming negative, provide better forecasts of future cumulative returns and they are normally distributed under the assumption of Brownian motion.

Hudson and Gregoriou (2015) further points out that the way of measuring returns does impact research and conclude that the theoretical mean calculated using log returns will be smaller than the mean calculated by simple returns. They write that increasing volatility will algebraically lower the expected return of log returns when compared to simple returns and that simple returns will provide a more meaningful and interpretable result of the final wealth in forecasting. They further write that either method is applicable for forecasting returns but that one needs to be careful when comparing studies using different ways of measuring returns and that the measure of return can have especially large impact on studies using high frequency data.

Andersen et. al (2003) show that using logarithmic returns generate better out of sample forecasts as extreme observations in the variance become smaller and thereby have a less distortive effect.

2.4.3 High frequency data

When using high frequency data some of the problems with time discretization and lack of information in between observations is reduced as the sampling frequency is increased. However, examining a financial time series using high frequency data also pose a new challenge in what is called micro structure effects. This is caused by the bid-ask spread where prices

are quoted in an interval with buyers and sellers giving different prices. This generates an uncertainty and a "noise" in the measurement of the prices (Barndorff-Nielsen and Shephard, 2002). The microstructure problem is increased with higher sampling frequency. Awartani et al (2009) find that microstructure noise is statistically significant in realized volatility measures sampled more frequently than 2-3 minutes when examining the Dow-Jones Industrial Index. Barndorff-Nielsen and Shephard (2002) found that the microstructure problem increase with volatility. Ait-sahila et al (2005) suggest that the optimal sampling frequency for a 1 day sample without accounting for the noise structure would be around 20 minutes. This is in line with the findings of Barndorff-Nielsen and Shephard (2002). Hansen and Lunde (2006) find that in a 20 minutes sample frequency the microstructure effects can be completely ignored.

Hansen and Lunde (2005) stress the importance of having a good estimate of the volatility in forecast evaluation as a noisy estimate can distort the model evaluation. This was also highlighted in Andersen and Bollerslev (1998) where they conclud that much of the critique against GARCH models poor out of sample performance had its roots in the noisy volatility estimate of squared returns. One has to be careful using previous estimates of optimal sampling frequency as Hansen and Lunde (2006) show that the properties of the noise has changed significantly over time.

Another downside to using high frequency data is that the datasets are large and hard to come by. In addition, large historical datasets are often prohibitively expensive for researchers to access when available.

2.4.4   The selected reference measure

I use the logarithmic value of the realized returns sampled at a 15 minute frequency as reference variance in the study. I chose this value as it strikes a good balance between high sample frequency and micro structure noise while incorporating the benefits of using logarithmic values. The fact that the standard trading day of 6 hours and 30 minutes easily can be divided into 15 minute segments also contributed to the selection. Using logarithmic returns also make the study more comparable to previous research.

2.4.5   Scoring the models

Variance forecast evaluation is not trivial. Hansen and Lunde (2005) highlights the difficulty of evaluating volatility models out of sample. They conclude that it is not obvious which loss function is more appropriate over others and as such decide to use six different loss functions defined as:

$$MSE_1 = n^{-1}\Sigma_{t=1}^n(\sigma_t - h_t)^2 \qquad MSE_2 = n^{-1}\Sigma_{t=1}^n(\sigma_t^2 - h_t^2)^2$$
$$QLIKE = n^{-1}\Sigma_{t=1}^n(log(h_t^2) + \sigma_t^2 h_t^{-2}) \qquad R^2LOG = n^{-1}\Sigma_{t=1}^n[log(\sigma^2 h_t^{-2})]^2 \qquad (28)$$
$$MAE_1 = n^{-1}\Sigma_{t=1}^n|\sigma_t - h_t| \qquad MAE_2 = n^{-1}\Sigma_{t=1}^n|\sigma_t^2 - h_t^2|$$

Where h_t is the realized volatility and σ_t is the forecasted volatility estimate conditional on the previous time period, n is the number of forecasts made and t is the time subscript for each forecast.

MSE_1 is commonly used to evaluate linear regression models but can also be used to evaluate variance models. MSE_2 is a variant of MSE_1 that is more sensitive to large outliers. A problem with this loss function, depending on forecasting methodology, is that it does not penalize models for creating negative variance forecasts which are obviously unrealistic as explained by Bollerslev et al (1994)

$QLIKE$ is what Bollerslev et al (1994) call the percentage of absolute errors and is the loss function implied by a Gaussian likelihood. This type of scoring penalizes negative forecast errors in variance more heavily. This can be valuable in financial forecasting where large unexpected volatility can be very costly

R^2LOG enhances the weight of the really good forecasts where the estimation error is close to zero.

MAE_1 and MAE_2 are more robust against outliers compared to MSE_1 and MSE_2, as explained by Bollerslev et al (1994).

However, as Bollerslev et al (1994) points out, the best loss function depends on what the estimated model is intended to do. In a financial setting this could for example mean that some profit based loss function would be optimal. The optimal model will ultimately be selected depending on what loss function the researcher decides to use. In this study I employ all of the above measures to illustrate the different characteristics of the examined models.

2.4.6 Diebold-Mariano test

The Diebold-Mariano (DM) test originally developed by Diebold and Mariano (1995) is used to compare the statistical significance of difference in forecast performance.

The test statistic is calculated as, using the notation of Diebold (2015):

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \tag{29}$$

Where the mean of the loss differential d_{12} between the compared models is calculated as:

$$\bar{d}_{12} = n^{-1} \sum_{t=1}^{n} L(e_{1t}) - L(e_{2t}) \tag{30}$$

Where n is the number of forecasts, t is a time subscript, $L()$ is a loss function and e_{1t} and e_{2t} are the forecast errors of the compared models. The parameter $\sigma_{\bar{d}_{12}}$ is the standard deviation of the mean loss differential $\bar{d}_{12}$. As the population of the test statistic has a standard normal distriubtion the examined DM statistic should be compared against an ordinary t-statistic.

The null hypothesis of the test is that the compared models have equal predictive ability while the alternative hypothesis is that one of the models is superior. The hypotheses are stated as:

$$H_0 : E(d_{12t}) = E[L(e_{1t}) - L(e_{2t})] = 0 \tag{31}$$
$$H_a : E(d_{12t}) = E[L(e_{1t}) - L(e_{2t})] \neq 0 \tag{32}$$

In this paper I compare all the estimated MIDAS models against the best performing GARCH(1,1) model. As it is not clear from the outset which model is superior in which measure I use two-sided confidence intervals.

3   Data

3.1   Theoretical data considerations

Poon and Granger (2003) do an overview of 93 papers published about volatility forecasting. They state that it is well documented that financial time series have "..fat tail distributions of risky asset returns, volatility clustering, asymmetry and mean reversion, and movements of volatilities across assets and financial markets".

Empirical work by Rydberg (2000) and others show that security returns tend to have heavier tails than the Gaussian distribution, that distributions sometimes are skewed, exhibit volatility clustering, there exists a leverage effect and returns show some seasonality. The heavier tail in practice means that extreme events are more common than what we would expect under a normal distribution. Rydberg (2000) shows that the skew of the return distribution is slightly negative meaning that large negative returns are more prevalent than large positive returns. One stated reason for this could be that investors are more influenced to negative information. She also states that the negative skew is prevalent in high frequency data. Volatility clustering, as the name alludes, refers to the property of larger changes in returns to appear in clusters in the data. This property is also called a long memory and is what makes the absolute returns and squared returns correlated over time. According to Rydberg (2000) the persistence of the long memory is stronger in absolute returns than it is in squared returns. The seasonality she refers to mainly stem from the fact that exchange traded securities only trade in a specified time window each day and that the exchange generally is closed in between which generates so called overnight effects in the data.

The leverage effect describes that volatility is negatively correlated with returns and was initially described by (Black 1976 and Christie 1982). The reason for this effect is that when asset prices decline the assets decrease but the loans do not, thus increasing the leverage. Leverage is generally perceived as risky meaning that as the leverage increase so does the risk. This is then directly translated to an increase in variance.

3.2   The collected data

The data used is the CRSP U.S. Mega cap Index collected from the Wharton Research Data Services (WRDS). The dataset is a second by second intraday dataset measuring the total return of the top 70% publicly investable US equity market as measured by market capitalization. The quotes that make up the index are collected from NYSE, NYSE MKT, NASDAQ, and NYSE Arca stock exchanges and include common stock, certificates and ADRs. The index measures total value and is adjusted for dividends. The data set collected stretch from 2012-12-07 to 2016-06-30 and the trading data is collected between 9:30 and

16:00 every trading day with the exception of 9 trading days where the exchange closed at 13:00. These dates are: 2012-12-24, 2013-07-03, 2013-11-29, 2013-12-24, 2014-07-03, 2014-11-28, 2014-12-24, 2015-11-27, and 2015-12-24. This means that they are spread out widely across the sample and should have a fairly small impact on the study as a whole. There are two alternatives with regards to how to deal with these missing values. The first is to drop them and proceed without them as including them would generate problems with the MIDAS forecasting. However, removing them will also obviously affect the autocorrelation function and I will therefore include them in the sample. When calculating the intraday high frequency returns I sample the data from these dates at a higher frequency to make sure that I have the same number of intraday observations as the other days. For example, in the 5 minute case the data would be sampled 78 times in an ordinary day. This means that for the shorter trading day I divide the 9:30 to 13:00 trading window into 78 segments and sample at this higher frequency of 2 minutes and 42 seconds between observations. Some observations are rounded and thus sampled at 2 minutes and 41 seconds to make sure that the trading day sums up to exactly 3 hours, 30 minutes.

The dataset contains 897 daily observations. From here on in the data section I will refer to observations rather than dates as they are of little importance for the statistical tests.

I begin by plotting the asset value over time. As can be seen in figure 3 the observed period saw a large increase in index price. The process appears to have some positive trend. If the trend is stochastic, as I suspect it to be due to the nature of financial time series as explained by Enders (2015), differencing should be able to solve this problem. There are also some very large swings in asset prices towards the later part of the time series which possibly could indicate a structural break in the data. This will be examined further in the results section.
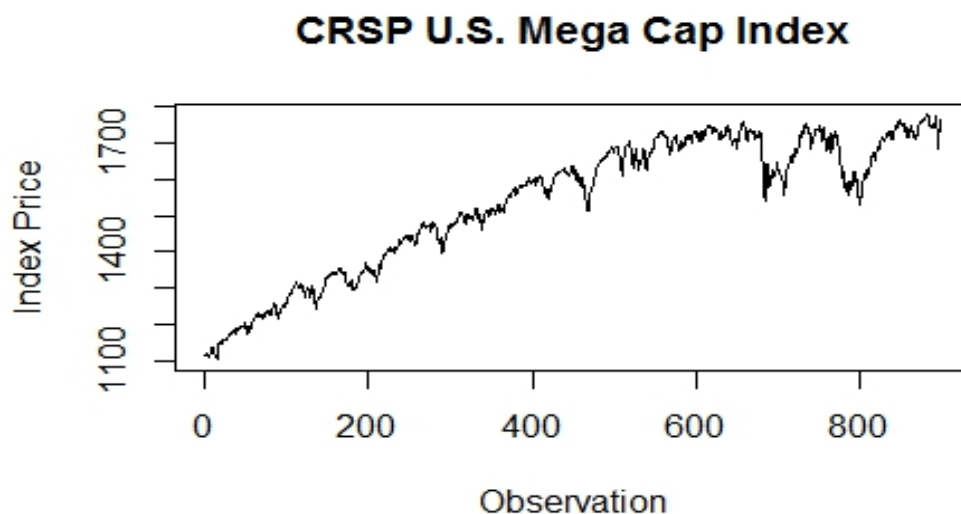
Figure 3: The CRSP mega cap returns

I then create three new datasets using the intraday data. All observations between 09:30 and 16:00 in an interval of 5, 15 and 30 minutes are collected and split of into the new datasets. A fourth time series including the daily observations is also kept.

Using a lagged vector I calculate the log returns for all four time series using the methodology described earlier. At this point I also calculate the realized variance using squared intraday returns sampled at 15 minutes. These will be used as my reference variance when evaluating the different models.

Next I look at the daily logarithmic returns. They are plotted as a time series in figure 4.
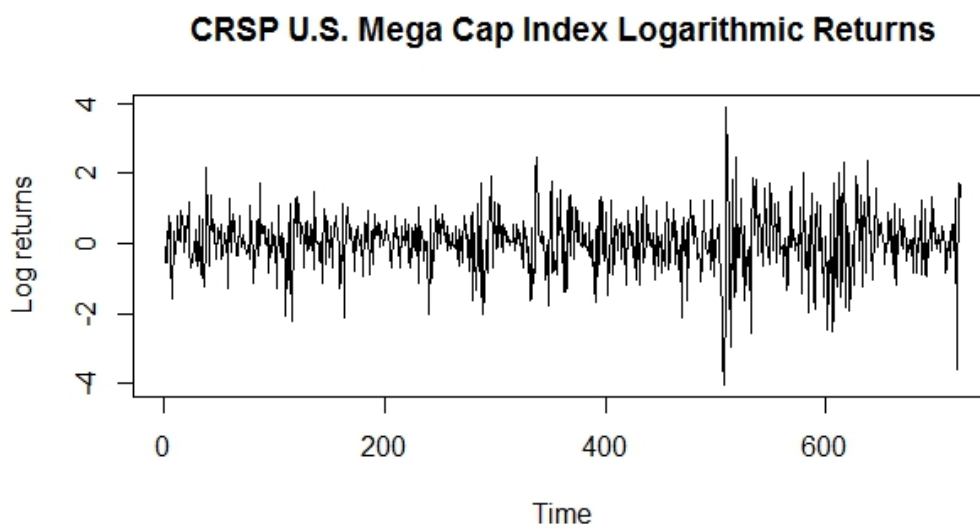
CRSP U.S. Mega Cap Index Logarithmic Returns

Figure 4: Logarithmic returns plotted over time

The daily log-returns look a lot more like a stationary time series than the original asset prices. The mean appears to be approximately zero and no linear trend can be visually identified in the data. Perhaps there is a structural break in the time series with the later part having higher variance but this will have to be tested before any conclusions can be drawn.

4 Results

4.1 Results of data processing

I begin by testing the data series for the presence of a unit root using the ADF methodology. The model is estimated using trend with drift and 10 lags. I reduce the lags until all the lags are significant which reduces the model to a total of 3 lags. Next I calculate the values of ϕ_2 and ϕ_3 to determine if I should proceed using the random walk, drift or trend specification. With this sample the significant values at the 1% level are:

$$\phi_{2, 1\%} = 6.09 \tag{33}$$

$$\phi_{3, 1\%} = 8.27 \tag{34}$$

The observed values are:

$$\phi_{2, \text{ Obeserved}} = 88.55 \tag{35}$$

$$\phi_{3, \text{ Obeserved}} = 132.82 \tag{36}$$

I am thus able to reject the null hypothesis of both no trend and no drift and I continue proceed to test for τ_3 using the trend model.

The observed $\tau_{3, \text{ Obeserved}}$ is -16.32 which is to be compared against the critical $\tau_{3, 1\%} = -3.96$.

The test is highly significant and I am able to reject the null hypothesis of a unit root being present. The process appears to be stationary.

I create a QQ-plot which can be seen in figure 5 to check the normality of the data. I do this in order to determine what type of error structure I should use in my GARCH modeling.
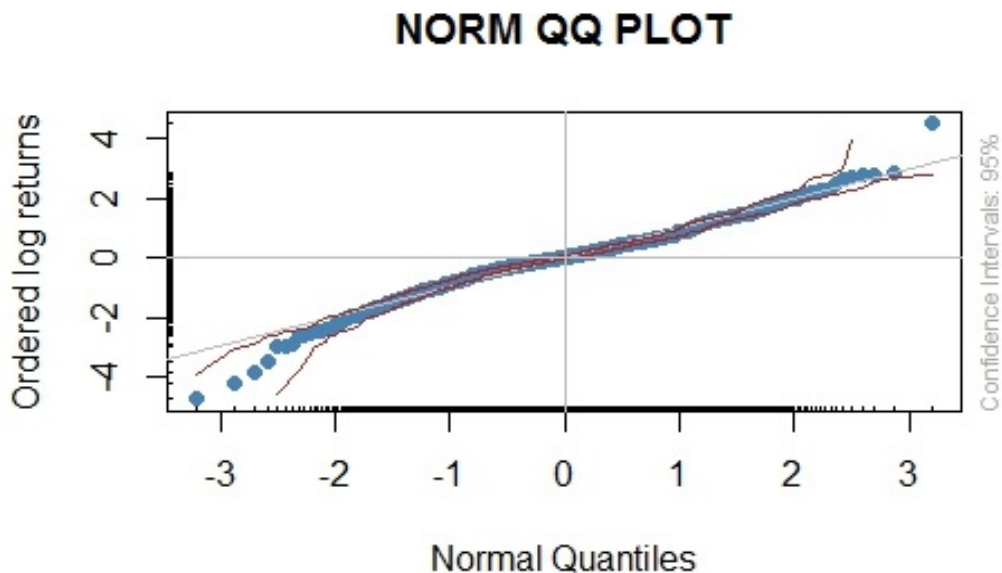

Figure 5: QQ-plot of the logarithmic returns

The daily log returns appear to be approximately normal but with somewhat heavy tails and a little skew to the right. This suggests that using a skewed t-distribution might be valuable when estimating a GARCH model for this dataset.

28
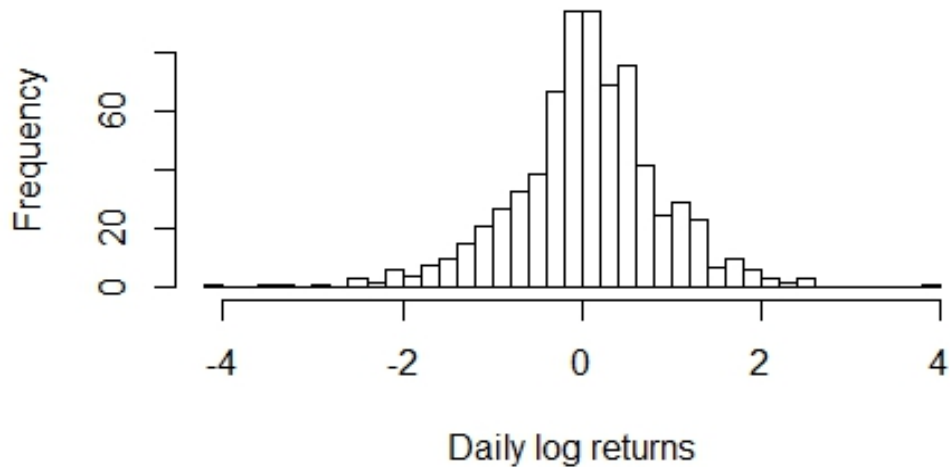
Figure 6: Histogram of the logarithmic returns

Looking at the histogram in figure 6 it is hard to tell if the data is skewed. The mean is 0.0517 and the median is 0.068 indicating that the data is slightly skewed to the left. The non-zero mean also indicates that an intercept might prove useful when modeling the mean.

There is a huge outlier where the asset prices fell dramatically between 09:30 and 10:00 on the 24th of august 2015. The outlier is due to what is called the "flash crash" of 2015. The effect of this crash is less prominent in the daily data than in the high frequency data and corresponds to observation 682.

I continue by examining the autocorrelation in the data by plotting the autocorrelation function (ACF) and partial autocorrelation function (PACF) in figure 7.

Figure 7: ACF and PACF of the logarithmic returns

Neither ACF nor PACF show any indication of autocorrelation which is expected for the return series given the assumption of returns following a random walk. One observation in the ACF and PACF is 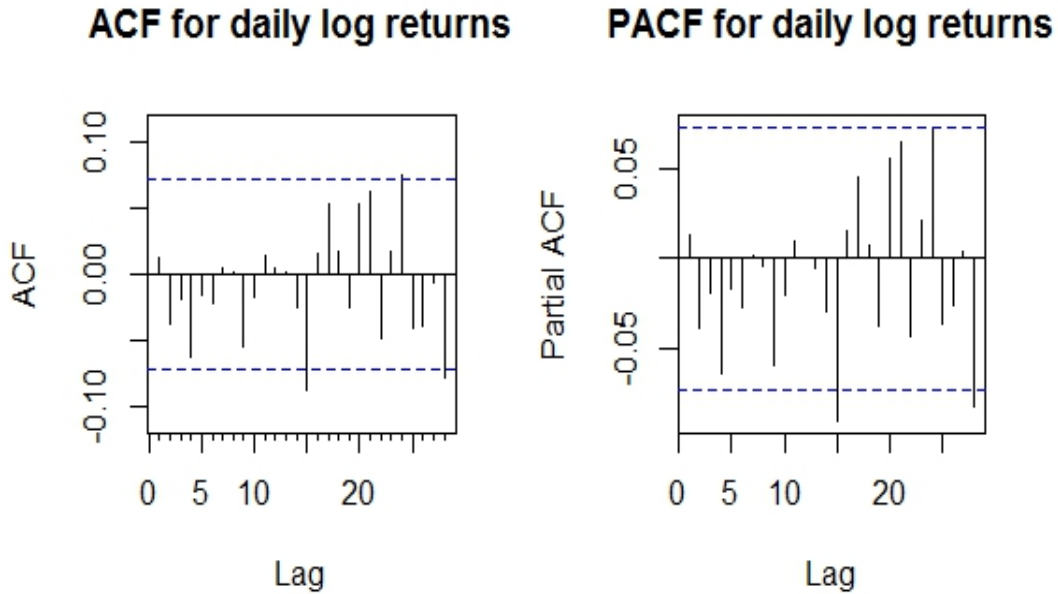outside the confidence interval but this is likely due to chance given that the lags before the observation are insignificant. This suggests that the mean process will be modeled well with only an intercept.

I use auto.arima function in R to confirm my choice. In doing this I am aware that the model can become over fitted but I think that it is good to see it agrees with my model selection. Auto.arima also picks a model with only an intercept (meaning that it is the model with lowest AIC).

Next I model the mean process with only an intercept; which is equal to an ARMA(0,0) process. The estimated model has an intercept of 0.0517.

I check the ACF and PACF in the ARMA(0,0) residuals and it again shows no indication of autocorrelation or that they have changed. The ACF and PACF plots have not changed since I have not directly touched the autocorrelation process but only slightly adjusted the mean of the process through adding the intercept.

I now run three Ljung-Box tests to confirm there is no autocorrelation left. I test 12, 8 and 4 lags. They are all insignificant which confirms what was seen in the diagnostic plots. The result of the Ljung-Box test can be seen in table 1.

Lags	Q statistic	P-value
4	4.376	0.375
8	4.947	0.7633
12	7.632	0.8132

Table 1: Significance of the Ljung-Box test

4.1.1 Checking for structural breaks in the mean

Using the ARMA(0,0) model that appear to be correctly specified I will check for structural breaks in the mean using a CUSUM test. I start by splitting of the first 30 observations in the dataset and use these as my estimation window. The data in the estimation window is used to estimate the parameters of an ARMA(0,0) process with an intercept of 0.1798.

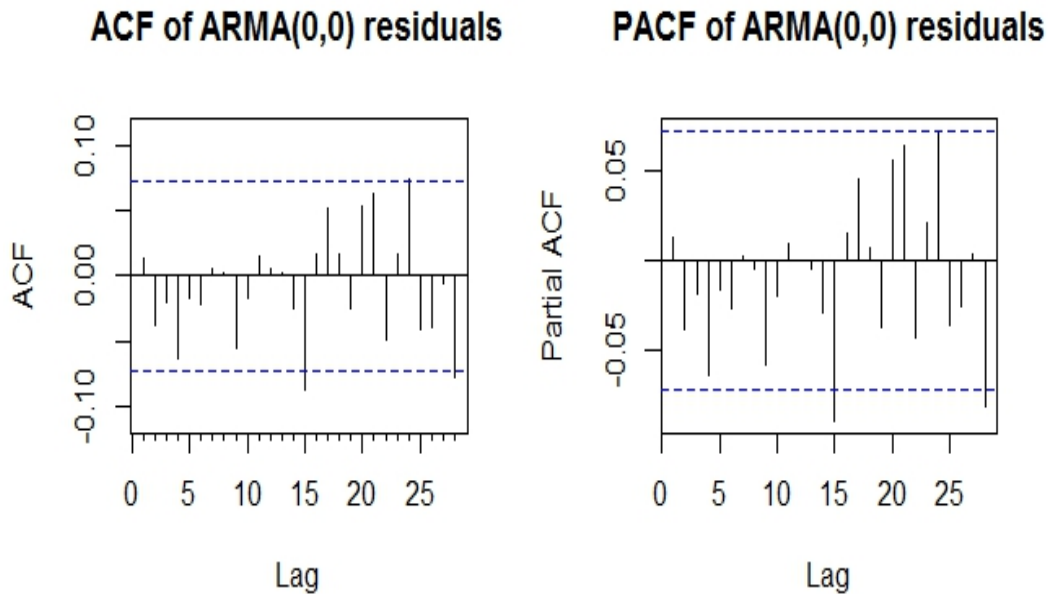Next I plot ACF and PACF on the model residuals in figure 8.

Figure 8: ACF and PACF of the ARMA(0,0) residuals

The ACF and PACF show no autocorrelation and the model appears to be correctly fitted to the data.

Next I conduct a Ljung-box test that generates the same results, plotted in table 2.

Lags	Q statistic	P-value
4	1.914	0.7515
8	6.324	0.611
12	10.217	0.597

Table 2: Significance of the Ljung-Box test on the ARMA(0,0) residuals

I now use the fitted model to create one step ahead forecasts of the asset returns. Using the actual returns and the fitted values I calculate an estimation error. Next I calculate the cumulative sum of the errors and divide it by the standard deviation of the errors to obtain the CUSUM test statistic. Using the formula provided by Enders (2015) I calculate the corresponding 5% confidence boundaries. The result is plotted below in figure 9, with the black series being the cumulative sum of residuals and the red lines being the confidence intervals.

Figure 9: CUSUM test of the ARMA(0,0) residuals to check for a structural break in the mean. The black line is the cumulative residuals is the red lines are the 95% confidence intervals of break being present.

The CUSUM test clearly indicates that there is a structural break in the mean present. I examine a few different breakpoints and conclude that splitting the data between point 174 and 175 generates the best results. Consequently I split the dataset into two parts. I discard the first 174 observations as the sample does not contain enough data points to adequately

fit a GARCH model, hence no comparison can be made using this sample. This leaves 723 observations.

I redo the all the above tests up to the CUSUM test for the remaining data and all the characteristics of the dataset are unchanged.

Next I redo the CUSUM test to check for a structural break in the mean. The result is no longer significant meaning that no structural break in the mean can be found. The test is plotted in figure 10.
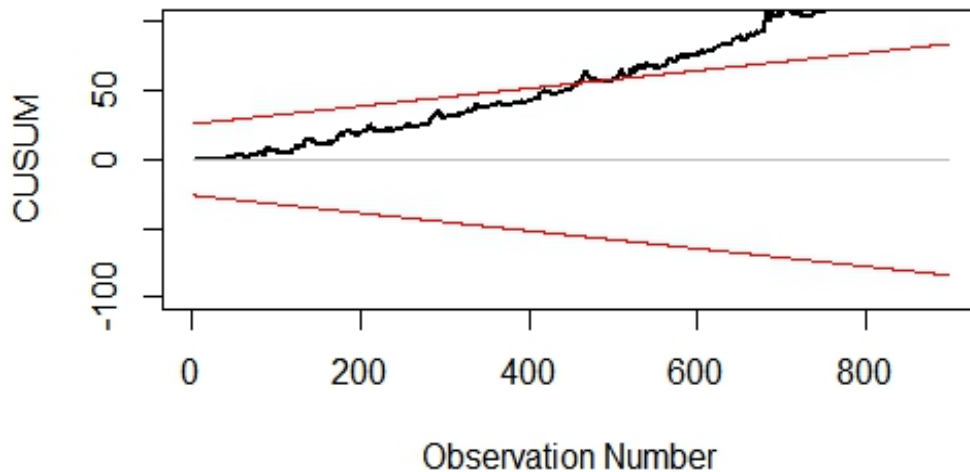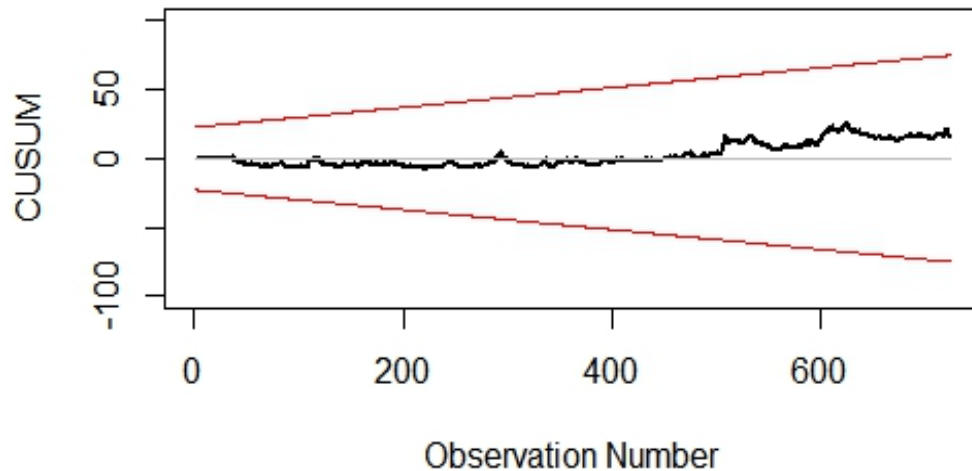

Figure 10: CUSUM test of the ARMA(0,0) residuals in the reduced sample to check for a structural break in the mean. The black line is the cumulative residuals and the red lines are the 95% confidence intervals of break being present.

4.1.2 Checking for structural breaks in the variance

I start by estimating an ARMA(0,0) model using all the remaining data. I then square the residuals and check for any GARCH effects. The GARCH effects are examined by looking for significance in the ACF and PACF of the squared residuals in figure 11.

ACF of squared residuals   **PACF of squared residuals**
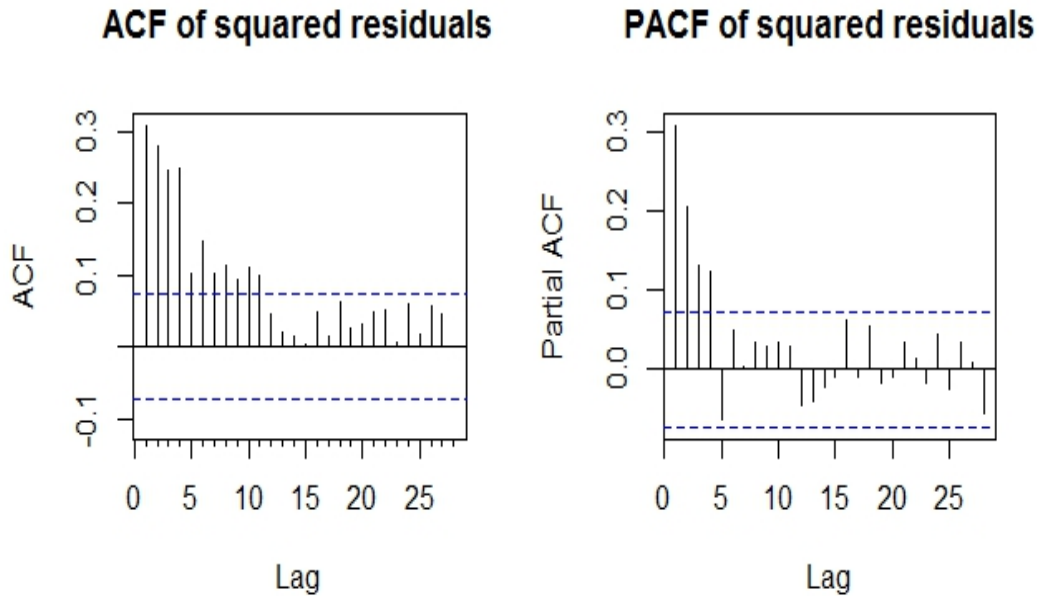
Figure 11: ACF and PACF of the squared residuals. No GARCH effect is present.

There appears to be a significant GARCH effect in the squared residuals. I now try to model these GARCH effects with a GARCH(1,1) model. I use the GARCH(1,1) as this is the most parsimonious model and the model most easily modeled.

I collect the conditional squared residuals of the mean process and the conditional variance of the ARMA(0,0)-GARCH(1,1) process using a skewed t-distribution as the conditional distribution. Next I subtract the conditional variance from the squared residuals and check if there is any unaccounted GARCH effect left in the sample by checking ACF and PACF in figure 12.
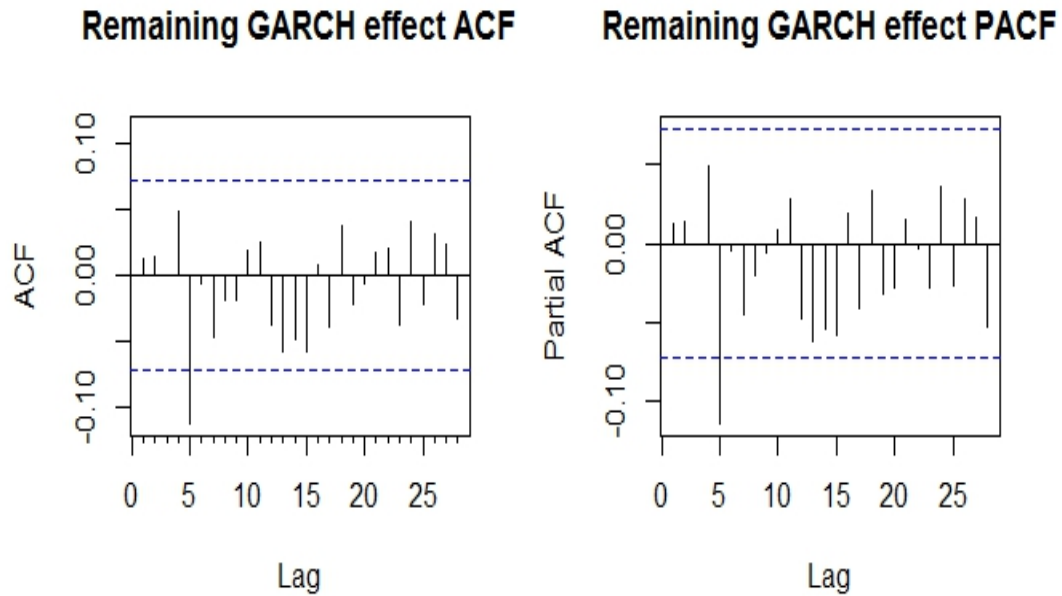
Figure 12: ACF and PACF of the squared residuals, also known as the GARCH effect.

The procedure appears to have removed most of the significant GARCH effects. At lag 5 some autocorrelation remains and I try a few different model specifications but it does not improve upon the selected model.

Next I divide the squared residuals by the conditional variance. From here on I will call these the estimation error.

Using the estimation error I calculate the Lee et al (2004) test statistic and corresponding p-value of the hypothesis of no structural break. The CUSUM of the GARCH(1,1) residuals is plotted in figure 13.
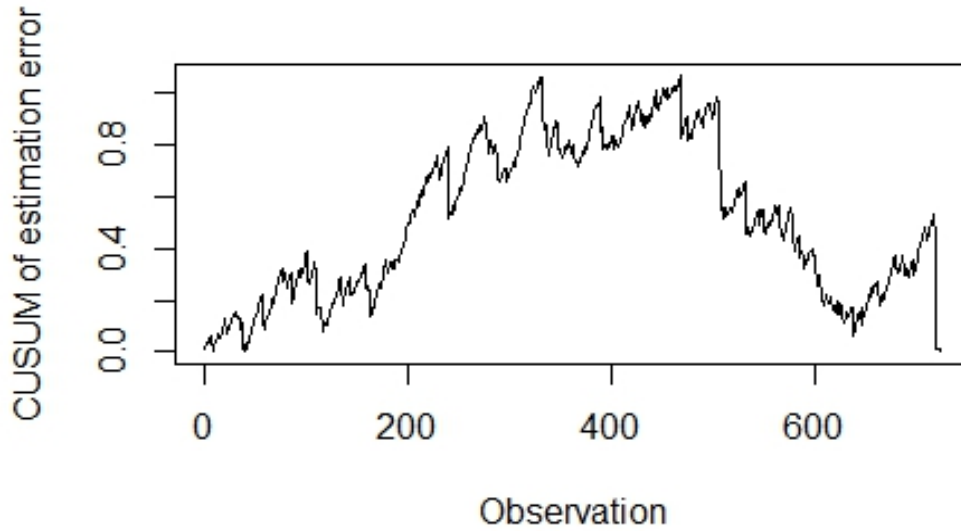
Figure 13: The cumulative sum of the GARCH(1,1) residuals.

The observed test statistic and corresponding confidence level of the Lee et al (2004) test for a structural break in the variance is listed in table 3.

Observed T	1.062
P-value	0.209
Suggested break point	461

Table 3: The result of the Lee et al (2004) test for structural breaks in the mean

The result is not significant at the 10% level. I therefore conclude that the null hypothesis of no structural break cannot be rejected, meaning that there is no evidence of structural break in the variance and no further action is needed.

4.2 Results of model estimation

The evaluated GARCH models can be found in table 4.

GARCH(1,1)		MSE_1	MSE_2	$QLIKE$	R^2LOG	MAE_1	MAE_2
	Normal	2.85	964.8	9.50	3.60	0.74	4.34
	t	2.82	963.4	9.71	3.61	0.74	4.34
	Skewed t	2.81	963.0	9.46	3.58	0.74	4.34

Table 4: The scoring of the GARCH(1,1) models

The skewed t-distribution generate the smallest errors among the GARCH models but the difference is relatively small and the models perform relatively similar. It is worth noting that the $QLIKE$ statistic is especially high when compared against the MIDAS models below. The high $QLIKE$ loss function indicates that the model is underestimating the risk (volatility) of the assets compared to the MIDAS model. The high R^2LOG function indicates that it has fewer really good estimates compared to the MIDAS regression. The GARCH model on the other hand scores comparatively well in MSE_1 and MSE_2 indicating that the model on average is more accurate than most of the MIDAS models. The skewed t-distribution generates the best results out of the three tested specifications. This is expected given the observed data distribution presented earlier.

Two parameter exponential Almon distribution	Lag	Frequency	MSE_1	MSE_2	$QLIKE$	R^2LOG	MAE_1	MAE_2
	1 Day	5 min	9.53	21065.6	5.95**	3.82	0.91	13.65
		15 min	5.99	6378.9	4.70**	3.18	0.81	9.29
		30 min	3.47	1141.2	4.40*	3.33	0.79	5.43
	5 Day	5 min	10.16	24465.7	6.00**	3.83	0.92	14.42
		15 min	4.95	3633.2	5.89***	3.50	0.82	7.87
		30 min	3.47	1142.4	4.43*	3.33	0.79	5.43
	20 Day	5 min	10.64	27247.2	6.25**	3.86	0.92	15.02
		15 min	2.81	948.5	5.68***	3.33	0.72	4.20
		30 min	3.43	1124.4	4.59*	3.35	0.79	5.34
	50 Day	5 min	13.03	43034.0	6.74*	3.92	0.96	17.91
		15 min	5.27	4514.2	5.97***	3.45	0.83	8.41
		30 min	3.46	1138.7	4.88*	3.41	0.80	5.38

Significance level: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 5: The scoring of the MIDAS models using exponential Almon distrbuted lag functions

The MIDAS models using the two parameter exponential Almon distributed lag structure performed better than the GARCH model when looking at the $QLIKE$ loss function indicating that the MIDAS models do not underestimate the variance to the same extent as the GARCH model. In MSE_1, MSE_2, MAE_1 and MAE_2 the GARCH models performed better than 8 out of the 12 tested MIDAS models using exponential Almon distributed lag structures. The MIDAS model with exponential Almon lag appears to work best with the data sampled at 30 minutes in this sample with the exception of the 20 day lag length sampled in intervals of 15 minutes that perform very well. Other than this 15 minute data appears to perform somewhat worse than the 30 minute data. In both cases the 20 day estimation window performs best. The 5 minute data performs worst out of the three sample frequencies. In fact, the 5 minute data does not improve on the GARCH models in any of the measured areas. All in all sample frequency appears to be much more important than the length of the lag structure when it comes to predictive performance.

	Lag	Frequency	MSE_1	MSE_2	$QLIKE$	R^2LOG	MAE_1	MAE_2
	1 Day	5 min	16.40	70993.6	5.96**	3.67	0.97	21.86
		15 min	3.12	955.9	6.09	4.01	0.77	4.27
		30 min	3.09	956.6	6.10	3.96	0.74	4.11
	5 Day	5 min	10.16	24465.7	6.00**	3.83	0.92	14.42
		15 min	4.86	3621.9	5.31**	3.69	0.85	7.94
		30 min	4.06	1488.2	6.36***	4.04	0.83	5.83
	20 Day	5 min	10.64	27247.2	6.25**	3.86	0.92	15.02
		15 min	4.89	3683.8	5.51**	3.71	0.85	7.98
		30 min	4.06	1494.3	6.61***	4.07*	0.84	5.84
	50 Day	5 min	13.03	43034.0	6.74*	3.92	0.96	17.91
		15 min	5.22	4503.3	5.89**	3.77	0.87	8.49
		30 min	4.10	1535.5	7.08**	4.13*	0.85*	5.91

(Left rotated label spanning all rows: Two parameter beta distribution)

Significance level: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 6: The scoring of the MIDAS models using beta distrbuted lag functions

The normalized beta lag distribution performs similar to the exponential Almon distributed lags. The main difference being that the best forecasts are done in the 1 day estimation window using data sampled at 30 minutes and that the 20 day lag length sampled at 15 minutes does not perform as well.

Located in Appendix in tables 7 and 8 are the DM statistics where the MIDAS forecast errors are compared against the GARCH(1,1) model with skewed t-distribution across all loss functions. Only the $QLIKE$ loss function is significantly better among the MIDAS models when measured over the whole sample of forecast errors. In no other loss function

are the MIDAS models predictive ability significantly different from the GARCH model.

To extend the analysis I drop the first 50 forecast errors and recalculate the DM statistic. I do this to create a sample that is less affected by the initial jump in variance caused by the flash crash. The new DM statistics can be found in table 9 and 10 located in Appendix. When measuring across this new calmer period it is clear that the GARCH model performed significantly better than the MIDAS models in MSE_1 and MSE_2 but that it still performed worse when using the $QLIKE$ loss function.

Plotted below in figure 14 are the forecast errors from the GARCH model using the skewed t-distribution and the 20 day exponential Almon distributed MIDAS lag model sampled at the 15 minute frequency for comparison.
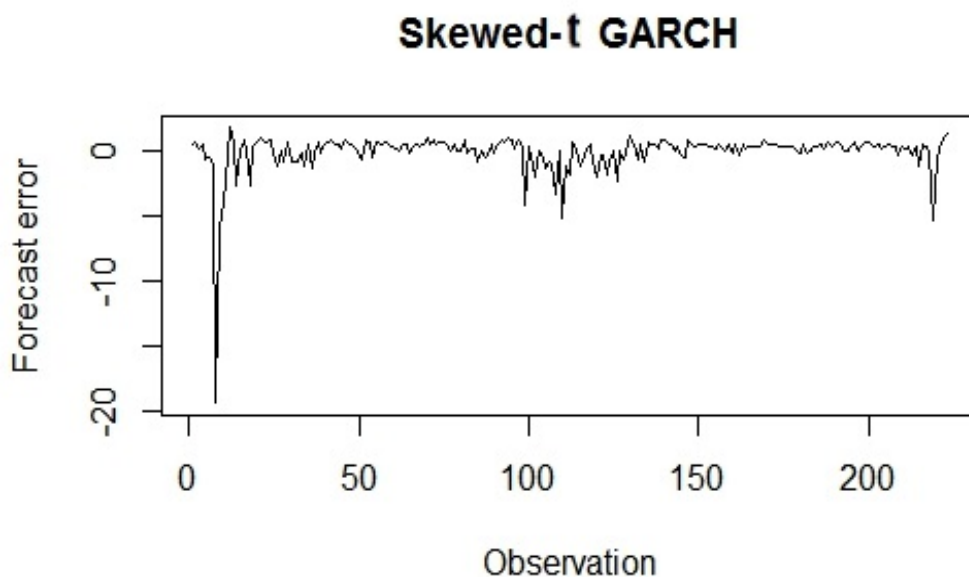
Figure 14: Forecast error generated by the GARCH(1,1) model estimated using a skewed t distribution

Much of the total error in the GARCH model comes from the first few observations surrounding the flash crash. When measuring in MSE_1 about 72% of the total squared errors arise in the first 15 observations. The GARCH models poor ability to account for the persistence of the volatility shock is what makes the GARCH model perform so much worse compared to the other models in the $QLIKE$ scoring.

In comparison the MIDAS model is much quicker to adjust to the shock and has a shorter period of negative forecasting errors following the shock. This is well illustrated by the

forecast errors produced by the 20-day lag exponential Almon model using data sampled at 15 minutes plotted in figure 15

20-Day Exponential Almon at 15min
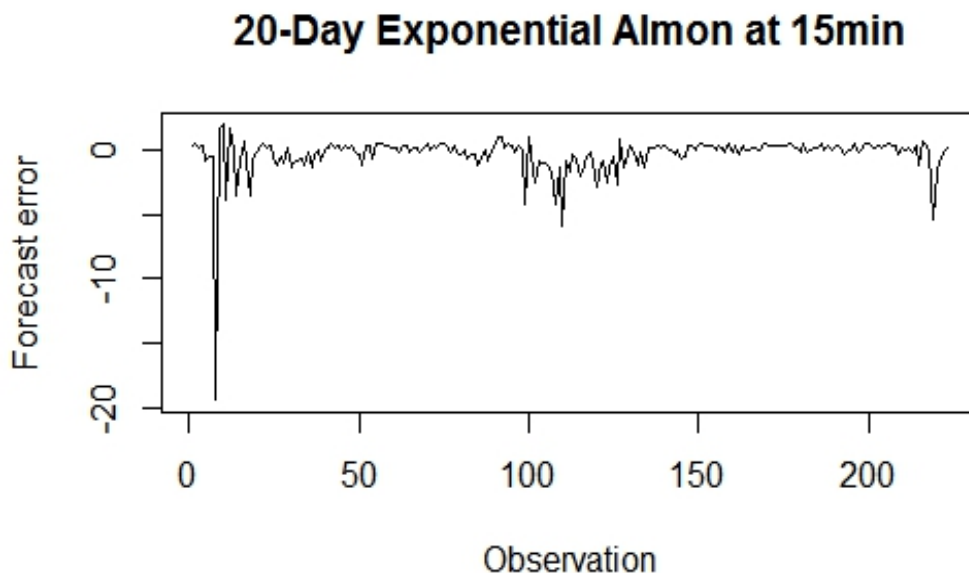
Figure 15: Forecast error generated by the MIDAS model estimated using exponential Almon distributed lags

The 15 minute realized variance and the squared returns plotted around the volatility shock can be found in figure 16 located in appendix.

5 Conclusions

The strength of the GARCH model lies in its ability to produce a small mean square error. However, the nature of the GARCH model also makes it prone to underestimate the persistence of a temporary volatility shock. Since the GARCH model relies on squared returns for estimating the variance it is able to capture intraday volatility which puts it at a disadvantage in cases where the intraday volatility is high but the difference in open and close price is small.

This can have important economic consequences in for example value at risk estimations. However, as long as the practitioner is aware of this it is a weakness that is fairly easy to compensate for given that the poor estimates only happen after the fact of a large volatility

40

spike which gives time to adjust. Although slightly different studies the general result that the MIDAS model performs better in periods of high volatility is in line with the findings of Emre Alper et al (2012) and Çelik and Ergin (2013).

Another important weakness of the GARCH model is the requirement of a long estimation window. First, this can produce problems in datasets with limited data. Second, it is also much less flexible to potential parameter changes that can be difficult to detect in real time. This is especially troubling when forecasting volatility as it is a historic measure that only can be observed after the fact, meaning that a structural break only can be found after some time has passed. A model using a shorter estimation window is much more able to adjust to parameter changes.

The GARCH problem of requiring a long estimation window of course has to be set in relation to the MIDAS models problem of requiring a data series that is sampled at a higher frequency. Although the high cost might not be a very large obstacle for practitioners at financial insinuations it is considerable for the academic researcher.

A difficulty in the MIDAS methodology is picking the correct lag length. The result of this study is somewhat of a mixed message. In this study the beta distributed lag function performed well with shorter estimation window (1-day) while the exponential Almon distributed lag function performed better with longer estimation windows (5 and 20 days). It is hard to know if this is general property of the MIDAS model or if it originates from some sample property.

I believe this study has shown that MIDAS modeling of variance using data sampled at a frequency of 30 minutes can produce fairly good forecasts using a wide range of lag structures. At the same time the 15 minute 20-day exponential Almon model does also show how superior a model can perform if the lag-length and lag distribution is correctly specified.

Further, the poor performance of the models using data sampled at the 5 minute frequency indicates that this is too high a sampling frequency and that there is microstructure noise present or that the lag structure follow some specification that cannot be modeled by two parameter exponential Almon or beta distributions.

The sample frequency appears to be the most important factor when deciding on what model to use in MIDAS modeling as this appears to dictate the predictive performance more than any other factor.

In this sample the 20 day lagged and 5 day lagged models performed better forecasts than the longer 50 day estimation window. This is in contrast with what other researchers such as Ghysels et al (2006) and of Emre Alper et al (2012) have found. This could possibly be due to the large spike in variance following the flash crash that favors models with shorter estimation windows that are quicker to adapt to changing circumstances.

6   Further research

A natural extension of this study would be to include absolute returns to see how they perform in the MIDAS setting. This would be interesting given the that several sources state that absolute returns have more explanatory power than the squared returns used in this study. Forsberg and Ghysels (2007) show that "absolute return-based volatility measures have the most desirable population prediction properties" in a MIDAS setting. Further, Ghysels et al. (2006) find that intra-daily absolute returns are the best predictors over all horizons and thereby contain the most information to be utilized by the MIDAS model.

Another interesting research idea would be to further compare model properties using a Monte Carlo study with simulated shocks in the variance. This would help to further illustrate the abilities and disadvantages of the different approaches in several settings.

7   References

Articles

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.),Proceedings of the Second International Symposium on Information Theory (pp. 267–281). Budapest: Akademiai Kiado.

Andersen, Bollerslev, Diebold, & Ebens. (2001). The distribution of realized stock return volatility. Journal of Financial Economics, 61(1), 43-76.

Andersen, T., Bollerslev, T., Diebold, F., & Labys, P. (2003). Modeling and Forecasting Realized Volatility. Econometrica, 71(2), 579-625.

Andersen, Bollerslev, & Meddahi. (2011). Realized volatility forecasting and market microstructure noise. Journal of Econometrics, 160(1), 220-234.

Andersen, Torben G., & Bollerslev, Tim. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts.(Symposium on Forecasting and Empirical Methods in Macroeconomics and Finance). International Economic Review, 39(4), 885.

Andreou, E., Ghysels, E., Franses, Philip Hans, & Mcaleer, Michael. (2002). Detecting multiple breaks in financial market volatility dynamics. Journal of Applied Econometrics, 17(5), 579-600.

Andreou, Ghysels, & Kourtellos. (2010). Regression models with mixed sampling frequen-

cies. Journal of Econometrics, 158(2), 246-261.

Armesto, M., Engemann, Kristie M, & Owyang, Michael T. (2010). Forecasting with mixed frequencies. Review // Federal Reserve Bank of St. Louis, 92(6), 521-536

At-Sahalia, Y. (2005). How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. The Review of Financial Studies, 18(2), 351-416.

Awartani, B., Corradi, V., & Distaso, W. (2009). Assessing Market Microstructure Effects via Realized Volatility Measures with an Application to the Dow Jones Industrial Average Stocks. Journal of Business & Economic Statistics, 27(2), 251-265.

Barndorff-Nielsen, O., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(2), 253-280.

Brown, R., Durbin, J., & Evans, J. (1975). Techniques for Testing the Constancy of Regression Relationships over Time. Journal of the Royal Statistical Society. Series B (Methodological), 37(2), 149-192.

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

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. The Review of Economics and Statistics, 69(3), 542-547.

Black, F., (1976). Studies of stock price volatility changes. In: Proceedings of the 1976 Meeting of the American Statistical Association. pp. 171–181.

Çelik, & Ergin. (2014). Volatility forecasting using high frequency data: Evidence from stock markets. Economic Modelling, 36, 176-190.

Christie, A. A., (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. Journal of Financial Economics 10, 407–432.

Chow, G. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. Econometrica, 28(3), 591-605.

Day, T., & Lewis, C. (1992). Stock Market Volatility and the Information Content of Stock Index Options. Journal of Econometrics, 51(1,2), 267.

Dickey, D. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. Biometrika, 71(3), 599-607.

Dickey, D., & Fuller, W. (1979). "Distribution of the Estimators for Autoregressive Time

Series With a Unit Root" (T&M). Journal of the American Statistical Association, 74(365), 427.

Dickey, D., & Fuller, Wayne A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica : Journal of the Econometric Society, an Internat. Society for the Advancement of Economic Theory in Its Relation to Statistics and Mathematics, 49(4), 1057-1072.

Diebold, F., & Mariano, R. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3), 253-263.

Diebold, F. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. Journal of Business & Economic Statistics, 33(1), 1.

Emre Alper, Fendoglu, & Saltoglu. (2012). MIDAS volatility forecast performance under market stress: Evidence from emerging stock markets. Economics Letters, 117(2), 528-532.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica : Journal of the Econometric Society, an Internat. Society for the Advancement of Economic Theory in Its Relation to Statistics and Mathematics, 50(4), 987-1007.

Engle, R., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. Econometric Reviews, 5(1), 1-50.

Engle, R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. Journal of Economic Perspectives, 15(4), 157-168.

Fama, E. (1971). Efficient capital markets: A review of theory and empirical work. Frontiers of Quantitative Economics, 309-361.

Foroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(1), 57-82.

Forsberg, L., & Ghysels, E. (2007). Why Do Absolute Returns Predict Volatility So Well? Journal of Financial Econometrics, 5(1), 31-67.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models. IDEAS Working Paper Series from RePEc, IDEAS Working Paper Series from RePEc, 2004.

Ghysels, Santa-Clara, & Valkanov. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. Journal of Econometrics, 131(1), 59-95.

44

Ghysels, E.; Kvedaras, V.; Zemlys, (2016) V. Mixed Frequency Data Sampling Regression Models: The R Package midasr. Journal of Statistical Software, Volume 72, Issue 4.

Ghysels, E., Valkanov, R. and Serrano, A.R., Multi-Period Forecasts of Volatility: Direct, Iterated, and Mixed-Data Approaches (February 16, 2009). EFA 2009 Bergen Meetings Paper. Available at SSRN: https://ssrn.com/abstract=1344742 or http://dx.doi.org/10.2139/ssrn.1344742

Hansen, P., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics, 20(7), 873-889.

Hansen, P., & Lunde, A. (2006). Realized Variance and Market Microstructure Noise. Journal of Business & Economic Statistics, 24(2), 127-161.

Hudson, & Gregoriou. (2015). Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. International Review of Financial Analysis, 38, 151-162.

Hwang, S., & Valls Pereira, P. (2006). Small sample properties of GARCH estimates and persistence. The European Journal of Finance, 12(6-7), 473-494.

Inclán, C., & Tiao, G. (1994). Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance. Journal of the American Statistical Association, 89(427), 913-923.

Kim, S., Cho, S., & Lee, S. (2000). On the Cusum test for parameter changes in garch(1,1) Models. Communications in Statistics - Theory and Methods, 29(2), 445-462.

Lee, S., Tokutsu, Y., & Maekawa, K. (2004). The Cusum Test for Parameter Change in Regression Models with ARCH Errors. JOURNAL OF THE JAPAN STATISTICAL SOCIETY, 34(2), 173-188.

Ljung, G., & Box, G. (1978). On a Measure of Lack of Fit in Time Series Models. Biometrika, 65(2), 297-303.

Meucci, A. (May 1, 2010). Quant Nugget 2: Linear vs. Compounded Returns – Common Pitfalls in Portfolio Management GARP Risk Professional, pp. 49-51, April 2010 . Available at SSRN: https://ssrn.com/abstract=1586656

Nelson, & Plosser. (1982). Trends and random walks in macroeconmic time series: Some evidence and implications. Journal of Monetary Economics, 10(2), 139-162.

Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. Econometrica : Journal of the Econometric Society, an Internat. Society for the Advancement of Economic Theory in Its Relation to Statistics and Mathematics, 57(6), 1361-1401.

Ploberger, Werner, & Kramer, Walter. (1992). The CUSUM test with OLS residuals. (ordinary least squares). Econometrica, 60(2), 271.

Poon, S., & Granger, C. (2003). Forecasting Volatility in Financial Markets: A Review. Journal of Economic Literature, 41(2), 478-539.

Rydberg, T. (2000). Realistic Statistical Modelling of Financial Data. International Statistical Review, 68(3), 233-258.

Rydberg, T., & Shephard, Neil G. (1998). Dynamics of trade-by-trade price movements : Decomposition and models (Working paper series / Centre for Analytical Finance 21). Aarhus.

Schwarz, G. Statist. 6 (1978). Estimating the Dimension of a Model. The Annals of Statistics no. 2, 461-464. doi:10.1214/aos/1176344136. http://projecteuclid.org/euclid.aos/1176344136.

Books

Bauwens, L. (2012) (Authored by Ghysels et al). Forecasting Volatility with MIDAS. In Wiley Handbooks in Financial Engineering and Econometrics (pp. 383-401). Hoboken, NJ, USA: John Wiley & Sons.

Bollerslev, Engle, & Nelson. (1994). Chapter 49 Arch models. Handbook of Econometrics, 4, 2959-3038.

Enders, W. (2015). Applied econometric time series (Applied econometric time series). Hoboken, NJ: Wiley. 184-192,
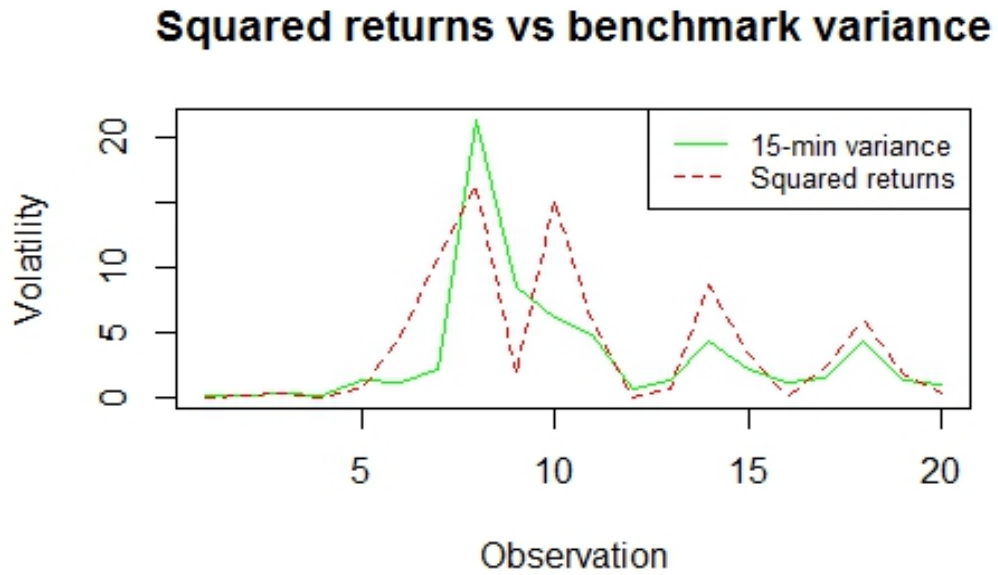
8 Appendix

Figure 16: The 15-minute realized variance and the squared returns plotted around the volatility shock

	Lag	Frequency	MSE_1	MSE_2	$QLIKE$	R^2LOG	MAE_1	MAE_2
Two parameter exponential Almon distribution	1 Day	5 min	1.05	1.00	-2.98	1.06	1.14	1.01
		15 min	1.15	1.01	-2.94	-1.79	0.74	1.07
		30 min	1.68	1.07	-2.43	-1.03	1.06	1.34
	5 Day	5 min	1.05	1.00	-2.94	1.11	1.14	1.01
		15 min	1.26	1.02	-3.89	-0.36	1.16	1.12
		30 min	1.68	1.07	-2.42	-1.00	1.07	1.34
	20 Day	5 min	1.04	1.00	-2.89	1.25	1.15	1.01
		15 min	-0.01	-0.91	-4.19	-1.29	-0.72	-0.79
		30 min	1.65	1.06	-2.38	-0.96	0.97	1.25
	50 Day	5 min	1.03	1.00	-2.58	1.56	1.16	1.01
		15 min	1.19	1.02	-4.31	-0.60	1.13	1.11
		30 min	1.63	1.06	-2.26	-0.72	1.08	1.25

Table 7: The t-statistic from the DM tests when comparing the exponential Almon distributed lag functions against the GARCH model with skewed t-distribution using the full sample. Positive values indicate that GARCH performed better while negative values indicate that the compared MIDAS model performed better.

	Lag	Frequency	MSE_1	MSE_2	$QLIKE$	R^2LOG	MAE_1	MAE_2
Two parameter beta distribution	1 Day	5 min	1.02	1.00	-2.72	0.44	1.04	1.01
		15 min	1.13	-0.30	-1.71	1.60	0.78	-0.22
		30 min	0.96	-0.27	-1.96	1.46	0.11	-0.74
	5 Day	5 min	1.05	1.00	-2.94	1.11	1.14	1.01
		15 min	1.21	1.02	-2.97	0.46	1.60	1.15
		30 min	1.70	1.03	-3.44	1.89	1.91	1.09
	20 Day	5 min	1.04	1.00	-2.89	1.25	1.15	1.01
		15 min	1.20	1.02	-2.95	0.57	1.63	1.15
		30 min	1.69	1.03	-3.42	2.04	1.96	1.09
	50 Day	5 min	1.03	1.00	-2.58	1.56	1.16	1.01
		15 min	1.16	1.02	-2.77	0.80	1.63	1.13
		30 min	1.65	1.03	-3.09	2.37	2.08	1.10

Table 8: The t-statistic from the DM tests when comparing the beta distributed lag functions against the GARCH model with skewed t-distribution using the full sample. Positive values indicate that GARCH performed better while negative values indicate that the compared MIDAS model performed better.

	Lag	Frequency	MSE_1	MSE_2	$QLIKE$	R^2LOG	MAE_1	MAE_2
Two parameter exponential Almon distribution	1 Day	5 min	2.64	2.38	-2.52	0.29	0.54	0.41
		15 min	1.56	2.11	-2.62	-3.22	-1.97	-1.06
		30 min	1.69	1.11	-2.21	-1.76	0.76	1.45
	5 Day	5 min	2.64	2.37	-2.49	0.34	0.57	0.43
		15 min	1.96	2.21	-3.44	-1.81	-0.66	-0.15
		30 min	1.68	1.10	-2.20	-1.74	0.77	1.45
	20 Day	5 min	2.64	2.36	-2.44	0.52	0.69	0.50
		15 min	2.38	2.28	-3.70	-1.10	0.05	0.59
		30 min	1.67	1.10	-2.17	-1.64	0.83	1.46
	50 Day	5 min	2.68	2.31	-2.16	1.00	1.03	0.72
		15 min	2.00	2.19	-3.89	-1.71	-0.69	-0.37
		30 min	1.66	1.09	-2.06	-1.36	0.99	1.47

Table 9: The t-statistic from the DM tests when comparing the exponential Almon distributed lag functions against the GARCH model with skewed t-distribution after the first 50 forecast errors have been removed. Positive values indicate that GARCH performed better while negative values indicate that the compared MIDAS model performed better.

	Lag	Frequency	MSE_1	MSE_2	$QLIKE$	R^2LOG	MAE_1	MAE_2
Two parameter beta distribution	1 Day	5 min	2.57	2.44	-2.41	-0.08	0.27	0.31
		15 min	3.01	2.50	-1.39	1.09	1.66	1.91
		30 min	3.00	2.48	-1.54	1.32	1.31	1.35
	5 Day	5 min	2.64	2.37	-2.49	0.34	0.57	0.43
		15 min	2.71	2.32	-2.53	0.09	1.11	1.55
		30 min	2.79	2.30	-2.77	0.94	1.27	1.29
	20 Day	5 min	2.64	2.36	-2.44	0.52	0.69	0.50
		15 min	2.71	2.30	-2.49	0.23	1.21	1.61
		30 min	2.79	2.28	-2.71	1.14	1.39	1.37
	50 Day	5 min	2.68	2.31	-2.16	1.00	1.03	0.72
		15 min	2.73	2.27	-2.31	0.60	1.49	1.79
		30 min	2.83	2.25	-2.34	1.63	1.77	1.62

Table 10: The t-statistic from the DM tests when comparing the beta distributed lag functions against the GARCH model with skewed t-distribution after the first 50 forecast errors have been removed. Positive values indicate that GARCH performed better while negative values indicate that the compared MIDAS model performed better.