Nova School of Business and Economics

Assignment 3

Authors:
Benjamin C. Herbert (45775)
Niclas Frederic Sturm (45914)
Thomas Dornigg (41727)
Vanderhulst Michel (45183)
Beatriz Vidal Rodríguez (40757)

Supervisors:
Prof. Paulo M. M. Rodrigues
TA. Daniel Belo

An assignment submitted for the course

2272 Financial Econometrics

March 7, 2022

Case statement

Consider the paper and the PredictorData2019.xls from Goyal. You are requested to build 3 forecasting models for the monthly Index log-returns: 1) an ARMA type model; 2) a VAR model with 3 variables; 3) a model including volatility. Use the last 3 years of data as the evaluation period of your forecasts and compute 1 step-ahead forecasts for each model.

To decide on the quality of the forecast performance of the models use the conventional statistics such as the root mean square forecast error and the mean absolute forecast error, as well as the equal predictability tests of Diebold-Mariano and the Clark and West.

Write a small report where you describe the models that you used in your forecasting exercise, explain how you have computed the forecasts and discuss the forecast quality of the models based on the statistics indicated above.

---

1 ARMA-model

In the following, the group will first examine the time series for stationarity by using the plot as shown in Figure 1 below. In a next step, we will test with the **A**ugmented **D**icky **F**uller (ADF) and the KPSS test to validate our analysis.

A time series process $\{Y_t \mid t \in T\}$ is weakly stationary (or second-order stationary) if:

1. The mean function is constant and finite - $\mu_t = \mathbb{E}[Y_t] = \mu < \infty$
2. The variance function is constant and finite - $\sigma_t^2 = \mathrm{Var}[Y_t] = \sigma^2 < \infty$
3. The autocovariance and autocorrelation functions only depend on the lag

As stated in the task, the computation of the monthly log-returns is required to model the following processes. Log returns have some more favourable properties for statistical analysis than the simple level returns R_t. Mathematically speaking, they can be formulated as:

$$r_t = \ln(1 + R_t) = \ln\left(\frac{S_t}{S_{t-1}}\right) = \ln(S_t) - \ln(S_{t-1}) \tag{1}$$

As for practical meaning, academics concentrate on log returns because they eliminate the non-stationary properties of the data set, making the financial data more stable. Also, it is important to mention that log-returns are independent and identically distributed (i.i.d.).

The starting point for the examination of the time series for stationarity is the plot in Figure 1 below. Observing the differenced log-returns in the graph it can be assumed that the series is most likely stationary. This is also clearly observable, when looking at the rolling standard-deviation and mean in the plot. Both lines have a constant variation in the beginning, which indicates a stationary behaviour. One exception from this stationary property can be observed around the 1930s, where log-returns, rolling mean and standard deviation show an upward movement respectively outbreak which most likely has its roots in the Great Depression, starting in August 1929. However, all in all it can be said that even though small breaks occur, a first-order differencing may not be required to transform the time-series.

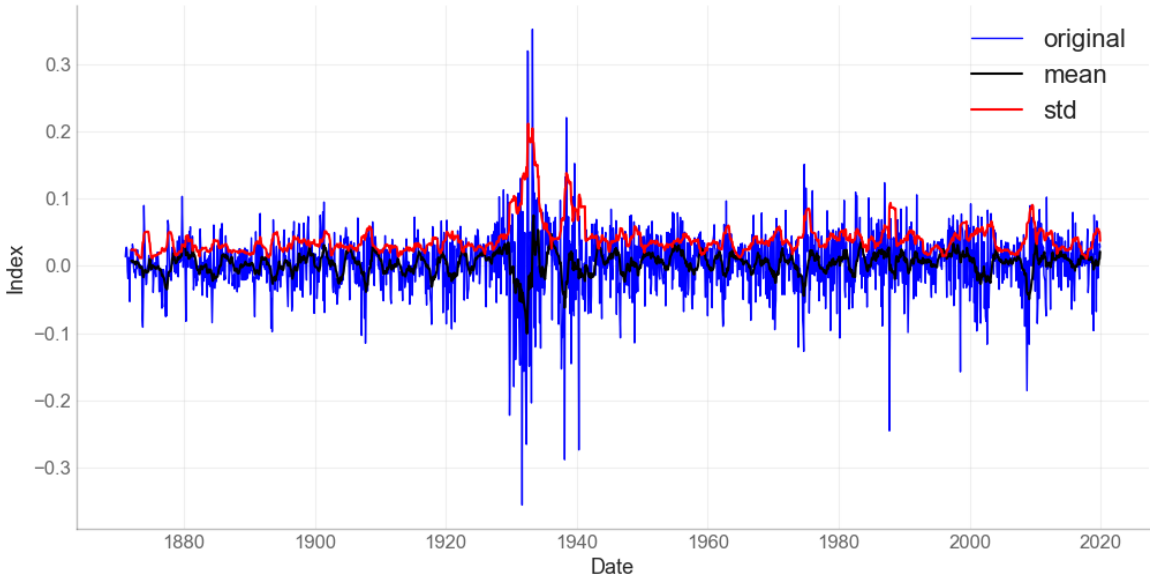

Figure 1: Rolling-mean and standard dev. for Index series

After a first visual conclusion, the group investigated the stationarity of the data by conducting two well-known tests (Augmented-Dickey-Fuller and KPSS), most commonly used in the academic literature. The Augmented Dickey-Fuller test allows for higher-order autoregressive processes by including Δy_{t-p} in the model where Δy_t can be written as

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_p \Delta y_{t-p} + \epsilon_t$$

The null hypothesis for the test is that the investigated data is non-stationary:

$$H_0 : \gamma = 0$$
$$H_1 : \gamma < 0$$

While the null-hypothesis of the ADF test investigates if the data is non-stationary, the null hypothesis for the KPSS test investigates if the data is stationary. As a result, one does not want to reject H_1. In Table below the group displayed the results for the Augmented Dickey-Fuller and KPSS test for the respective time-series (Note: a confidence level of 95% was applied while conducting the tests). By looking at the test results in Table 1, one can see that they are yielding the same outcome: both suggest that stationarity is present in the log-returns dataset.

Test	Test Statistic	p-value	Result
Augmented Dickey-Fuller	-9.9964	0.000	stationary
KPSS	0.3142	0.100	stationary

Table 1: Stationarity Test-Results for log-returns

Using the results from above, the group estimated an ARMA model for the log-returns of the monthly index series. To determine the included lags in the final model, the group used the information criterion according to Schwarz [1] (BIC). The model was preferred, which minimised this criterion. For this task, the group used the renown Python package pmdarima, which is equivalent to R's auto.arima function.

$$BIC = log\frac{1}{T}\sum_{n=1}^{T}\epsilon_t^2 + \frac{k}{T}\,logT$$

Before fitting the optimal ARMA model, the group split the provided dataset into a training and test set. For this, we chose as time-range for the training set Jan. 2000 till Dec. 2016 and for the test set Jan. 2017 to Dec. 2019, which can be seen in Figure 2 below. There was a trade-off between data availability (taking a longer time-horizon) and accuracy of the model (focusing on more recent events), i.e. taking a shorter period of time. One driving factor for this decision was the highly dubious assertion that trends and dynamics from *in extremis* 140 years ago could still be driving asset prices nowadays. Decisions regarding this issue are notoriously difficult and would merit a complete deliberation of their own, but in our opinion, all the secular trends that drive the stock market today are reflected best at that point in time.

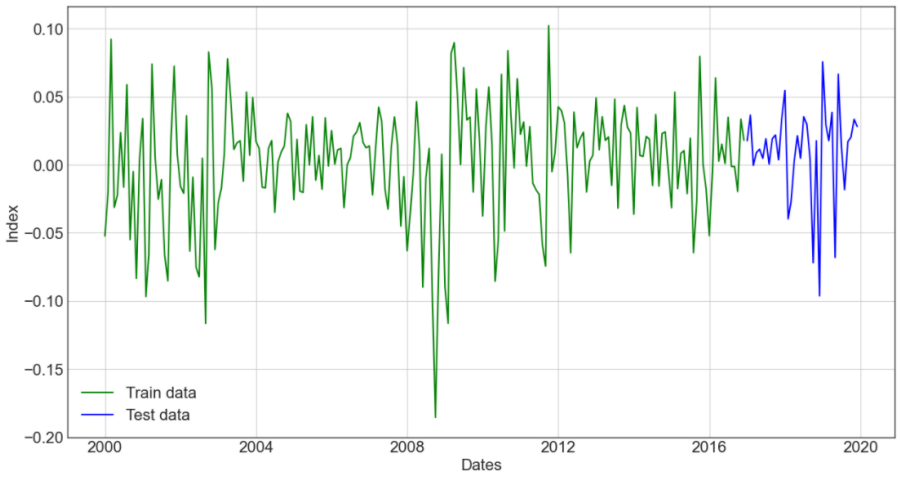

Figure 2: Train-Test split

The model is then estimated on in-sample and its forecasting performance is evaluated using some error measure on the holdout sample, which can be seen in a later part of this paper.

As one can see in Table 2, the group's estimated model is an **A**uto-**R**egressive-**M**oving-**A**verage model with specification ARMA(1,1).

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0021	0.003	0.631	0.528	-0.004	0.008
ar.L1.Index	-0.5437	0.259	-2.097	0.036	-1.052	-0.036
ma.L1.Index	0.6806	0.226	3.012	0.003	0.238	1.124

Table 2: Estimated parameters for ARMA(1, 1) model

In Table 2, the group presents the estimated parameters. It can be seen that, as expected, the log-return time series has no constant. In other words, the constant is not significant. On the other hand, both the AR and MA part of the model are significant.

In Figure 3 the group plotted the residuals $\hat{\epsilon}_t$ of the fitted model. If the model has been fitted sufficiently well, the residuals must correspond to a white noise process i.e., no auto-correlation which seems not be the case here.

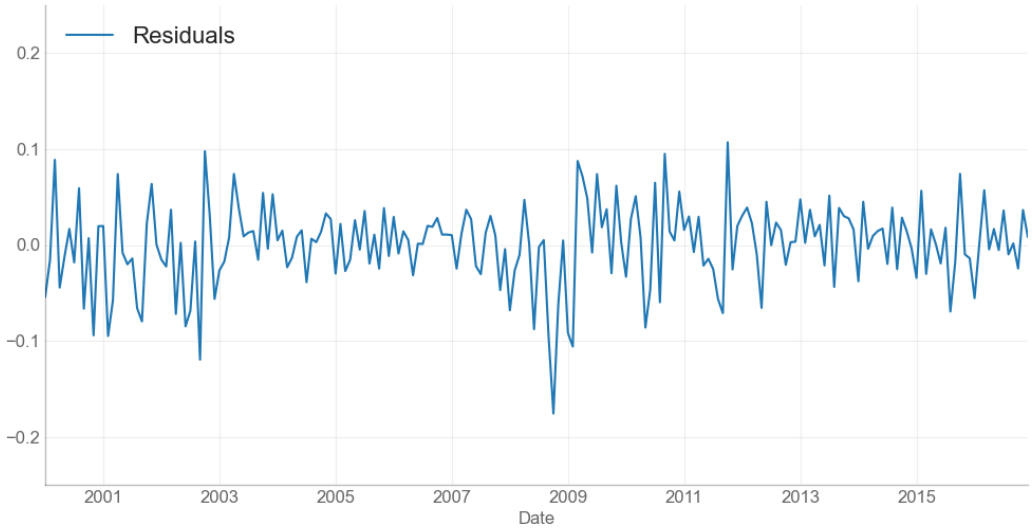

Figure 3: Residuals of fitted values from an ARMA(1,1)

To test the group's impression, a **Breusch-Godfrey** test, which investigates the presence of serial correlation of the residuals and a **Ljung-Box** test were performed. The Breusch-Godfrey test is especially good to use where lagged values of the dependent variables are used as independent variables in the model's representation, as it is the case in the ARMA model. One might use the Ljung-Box test on the residuals of the respective model to look for autocorrelation, ideally the residuals should be white noise.

Breusch-Godfrey tests the following hypothesis:

$$\begin{aligned} H_0 &: \text{ There is no serial correlation} \\ H_A &: \text{ There is serial correlation present} \end{aligned}$$

Ljung-Box tests the following hypothesis:

$$\begin{aligned} H_0 &: \text{ Data are independently distributed} \\ H_A &: \text{ Data exhibit serial correlation} \end{aligned}$$

Below, the group summarized the test results from the Breusch-Godfrey and Ljung-Box test. As one can see, both tests yield the same result: no serial correlation is present in the data.

Test	Test Statistic	p-value	Result
Breusch-Godfrey	8.99	0.8313	no-serial correlation
Ljung-Box	0.01	0.9427	independently distributed

Table 3: Autocorrelation Test-Results

Additionally to test for autocorrelation, the group also tested for the normality of the residuals with a QQ-plot and a Jarque-Bera test, as seen in Figure 4. A Quantile-Quantile or in short QQ plot, is a method to compare two distributions. In more formal terms, it is a

technique to compare whether two sets of sample points are from or follow the same distribution. If the distributions are identical, then the quantiles should be approximately equal and they should all lie on the red line.

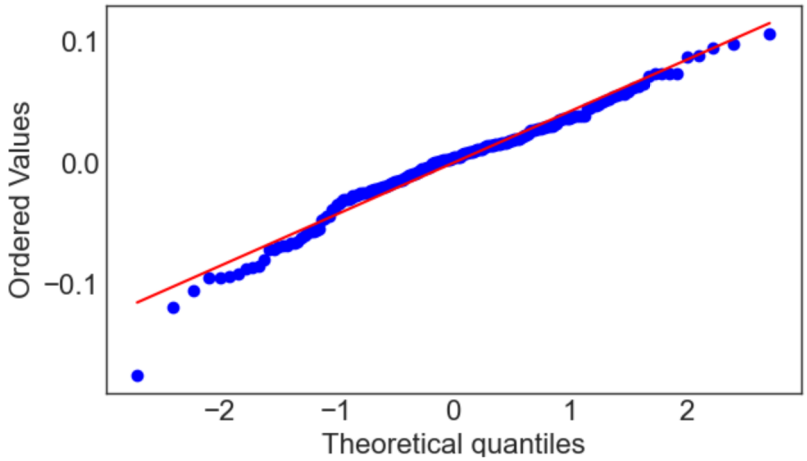

Figure 4: QQ-plot of residuals from ARMA (1,1)

The test statistic for the JB-test is 8.99 and the corresponding p-value is 0.00. Since the p-value is less than 0.05, we fail to reject the null hypothesis. We have sufficient evidence to say that this data has skewness and kurtosis that is significantly different from a normal distribution, meaning that the residuals are not normally distributed, which can also be confirmed by the fat-tails of the QQ-plot from above.

Rolling ARMA Forecast

In order to evaulate the performance of the fitted ARMA (1,1) model, the group used a rolling-window forecast for the test-set and evaluated the outcome with traditional error metrics such as the mean-absolute error (MAE), root mean-squared error (RMSE) and Diebold & Mariano. Note that the results for the error metrics are shown in section 4 of this paper.

A rolling window forecast was chosen, given the following reasoning from Gabriel Vasconcelos: *Naturally, if you do only one (or just a few) forecasting tests, your results will have no robustness and in the next forecast the results may change drastically. Another possibility is to estimate the model in, let's say, half of the sample, and use the estimated model to forecast the other half. This is better than a single forecast but it does not account for possible changes in the structure of the data over the time because you have only one estimation of the model. The most accurate way to compare models is using rolling windows. Suppose you have, for example, 200 observations of a time-series. First you estimate the model with the first 100 observations to forecast the observation 101. Then you include the observation 101 in the estimation sample and estimate the model again to forecast the observation 102. The process is repeated until you have a forecast for all 100 out-of-sample observations. This procedure is also called expanding window. If you drop the first observation in each iteration to keep the window size always the same then you have a fixed rolling window estimation. In the end you will have 100 forecasts for each model and you can calculate RMSE, MAE and formal tests such as Diebold & Mariano* [2].

To illustrate this rationale, consider the following visualization:

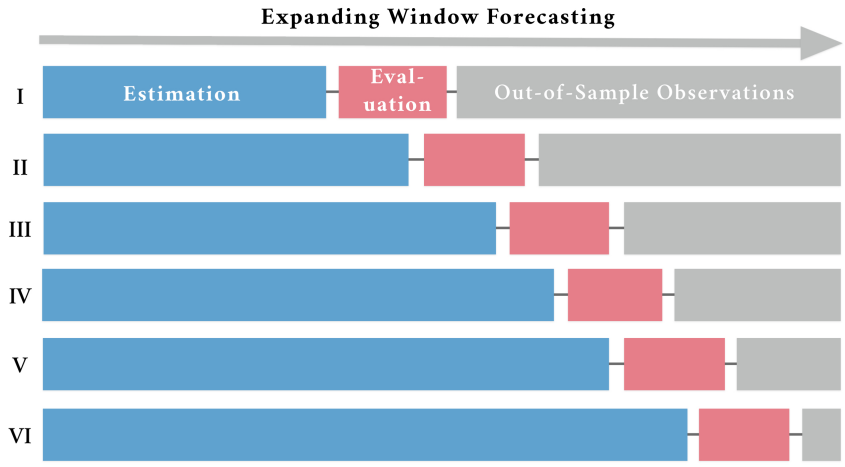

Figure 5: Rolling window forecasts

Below in Figure 6, the group plotted the rolling-window forecast against the original test-set. As one can see, the fitted ARMA (1,1) model does not posses a good fitting accuracy.

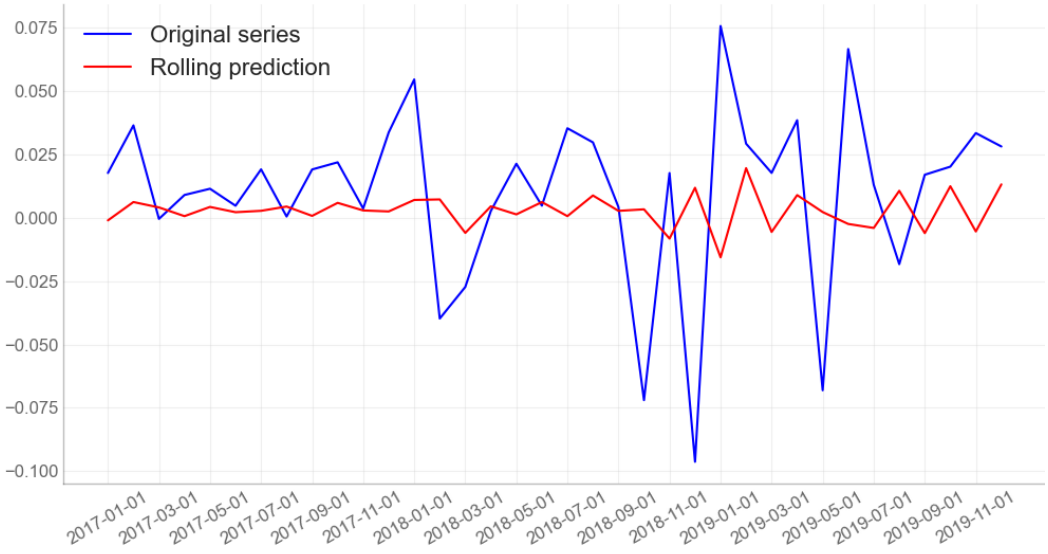

Figure 6: Rolling window forecast for test-set

2 VAR model

Another model that the group implemented for the purpose of this analysis was the VAR (*Vector Autoregression*) model, which entails the simultaneous estimation of a system of equations. Consider below one such system.

$$y_{1t} = \beta_{10} + \beta_{11} y_{1t-1} + \ldots + \beta_{1k} y_{1t-k} + \alpha_{11} y_{2t-1} + \ldots \alpha_{1k} y_{2t-k} + u_{1t}. \tag{2}$$
$$y_{2t} = \beta_{20} + \beta_{21} y_{2t-1} + \ldots + \beta_{2k} y_{2t-k} + \alpha_{21} y_{1t-1} + \ldots \alpha_{2k} y_{1t-k} + u_{2t}. \tag{3}$$

We consider u_{kt} to be *White Noise*. Other error specifications are possible here, including contemporaneous correlation. However, this model can also be written in Matrix/Vector-Notation:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \alpha_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} + \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{4}$$

The major challenge in the implementation of a VAR model here was to select - apart from the log-returns as a dependent variable, as described in Section 1 - two other variables that hold predictive power. Such decision might be informed by economic theory, a judgement derived from that source however of highly subjective quality, as the period considered needs be considered carefully. Thus, we decided to instruct a *Machine Learning* algorithm - in this case a Random Forest - to compute variable importances for all of the variables in the *Goyal* dataset. This *automatic* approach has several advantages, such as the possibility to work with large amounts of data. In recent literature Medeiros, Vasconcelos and Zilberman (2019) used Random Forest to extract variables for the purpose of inflation forecasting. Random Forests are best understand in algorithmic terms. They are based on entities called "Decision Trees", which can be used for regression tasks like the one at hand as well as classification. A decision trees splits in a way that a particular metric, such as variance in each leaf, is minimized. They are intriguing in their setup, yet exhibit several drawbacks, such as high variance, which makes them a highly imperfect estimator.

A Random Forest is then constructed by fitting multiple (usually hundreds) of Decision Trees on bootstrapped samples of the original dataset. Predictions are validated on the so-called *Out-of-Bag*-Sample (OOB). To derive from these predicitions the variable importance, we need to introduce the concept of *Permutation Importance*. Here, observations of a variable are randomly permuted. Depending on whether predictions improve (less important variable) or get worse (more important variable), we can compute overall variable importances by summing the importance scores of the lags of each variable, up to a certain maximum lag. In more mathematical terms, consider the formula used for this type of computation (Genuer, Poggi, Tuleau-Malot 2010) [3]:

$$VI(X_j) = \frac{1}{n} \sum_t (\tilde{\epsilon}_{t,j}^{OOB} - \epsilon_{t,j}^{OOB}). \tag{5}$$

Here n denotes the number of trees in the Random Forest, $\tilde{\epsilon}_{t,j}^{OOB}$ the error of the Random

Forest with the variables, whose observations are being permuted and finally $\epsilon_{t,j}^{OOB}$ characterized the prediction error without permutation.

For a Random Forest to function properly with Time Series, we need to make sure that all variables are considered stationary, otherwise the random permutation that is needed for the calculation of variable importance, yields biased results. We tested the Variable Selection algorithm as proposed for both differenced and not differenced variables and in fact they yield comparable results, yet employing differences improves algorithmic stability and constitutes a best practice.

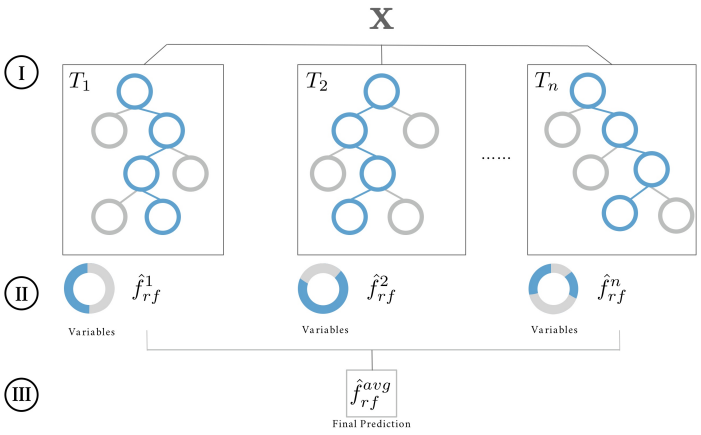

Figure 7: Basic structure of a Random Forest

The following table contains the four most significant variables as computed with the Random Forest Variable Selection methodology as described above.

Variable Name	$VI(X)$
b/m	0.000244009
tbl	0.0001561154
D12	0.0001333248
ntis	0.0001515955

Table 4: Random Forest Variable Selection Results

One can already in this excerpt identify an "Elbow Structure", in that the variable importance is sharply decreasing. The two most import variables are the absolute change *Book-to-Market* value and the absolute change in the *3-month Treasury Bill Secondary Market Rate*. As the final VAR should contain three variable and one of them is the log-returns variable, these two will be incorporated. The two remaining variables are the month-to-month difference in dividends (D12) and the month-to-month difference in net equity expansion (ntis).

Interestingly, the lagged values of the log-returns are among the least important variables, with their precise $VI(X)$ being close to zero. This means that randomly permutating the

lagged values of log-returns has very close to no effect at all on prediction accuracy. In fact, some variables such as inflation have a negative permutation importance, meaning that by randomly permutating the observations predictions actually get **better**. Before we used these three variables in our model, we conducted further analysis on the stationarity of the three series. For the log-returns, we already established their stationarity. Cointegrating relationships were not applicable here as the log-returns series is already I(0), which precludes a possible relationship with the other two variables.

To conclude the VAR modeling, we selected the optimal lag length for the VAR model using the Bayes Information Criterion (BIC) as described above. We found that the best model was a $VAR(2)$, which was then subsequently fit on the data. The rolling window forecast is contrasted with the actual returns in the graph below.

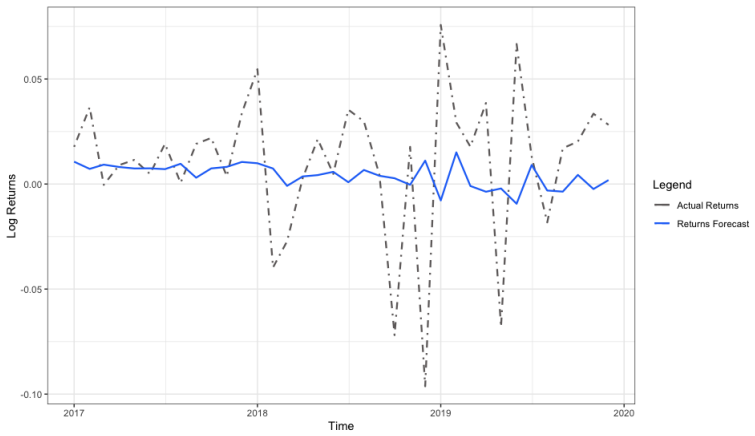

Figure 8: Returns forecast VAR

It seems that model seems to roughly follow the overall time trend, yet the more significant "eruptions" in the actual log-returns were not predicted by the model. In a sense, it does not model the volatility of the log-returns particularly well. Forecasting metrics and statistical tests are reported in tables 6, 7 and 8.

3 Volatility model

3.1 Fitting of the volatility model

To model and forecasting with the GARCH framework, we assume that the mean process can be described by the ARMA(1,1) model outlined in section 1.

$$r_t = \mu_0 + u_t$$

We're using three different conditional variance models in this section. Each is estimated with two assumptions on the error term u_t. First with the assumption of normal distribution and the second with a students-t distribution.

GARCH(1,1)-normal distributed errors

$$h_t = \alpha_0 + \alpha_i u_{t-1}^2 \beta_i h_{t-1}$$

In this framework we need to impose the following restrictions:

$$\begin{aligned} \alpha_0 &> 0 \\ \alpha_1 &\geq 0 \\ \beta_1 &\geq 0 \end{aligned}$$

GJR-GARCH(1,1,1)-normal distributed error

$$h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + + \beta_1 h_{t-1} + \gamma u_{t-1}^2 I_{t-1}$$

where

$$I_{t-1} = \begin{cases} 1 & : \quad u_{t-1} \geq 0 \\ 0 & : \quad u_{t-1} < 0 \end{cases}$$

Restrictions for GJR-GARCH models:

$$\begin{aligned} \alpha_0 &> 0 \\ \alpha_1 &\geq 0 \\ \beta_1 &\geq 0 \\ \alpha_1 + \gamma_1 &\geq 0 \end{aligned}$$

Additional Explanatory variable on the conditional variance.

$$h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + + \beta_1 h_{t-1} + \psi_1 X_t$$

As described above the two different assumptions were imposed on the error term u_t. Normal distributed errors:

$$u_t \sim (0, h_t = E_{t-1}(u_t^2))$$

Student-t distributed errors:

$$u_t \sim St(0, h_t, \nu)$$

In order to compute the starting value of h_1 we're following the proceeds described in the lecture.

3.2 Forecasting the volatility

Forecasting in the GARCH framework Since we are using a rolling window forecast the maximum steps which are forecasted are t+1.

GARCH(1,1)

$$h_{T+1|T} = \alpha_0 + \alpha_1 u_T^2 + + \beta_1 h_T$$

GJR-GARCH

$$h_{T+1|T} = \alpha_0 + \alpha_1 u_T^2 + + \beta_1 h_T + \gamma u_T^2 I_T$$

Explanatory variable

$$h_{T+1|T} = \alpha_0 + \alpha_1 u_T^2 + + \beta_1 h_T + \psi_1 X_{T+1}$$

To evaluate the the forecast accuracy of the **volatility** we use two different approaches which are popular in the financial econometrics literature. The first is a slight adaptions of the Mincer-Zarnowitz regression, which is reduces the impact of large returns:
Mincer-Zarnowitz regression:

$$r_t^2 = \delta_0 = \delta_1 \hat{h}_t + e_t$$

Adaption of the MC regression:

$$|r_t| = \delta_0 + \delta_1 \sqrt{\hat{h}_t} + e_t$$
$$log(r_t^2) = \delta_0 + \delta_1 \hat{h}_t + e_t$$

where e_t is the disturbance term. The related test is:

$$\begin{aligned} H_0 \ : \quad & \delta_0 = 0 \quad \text{and} \quad \delta_1 = 1 \\ H_1 \ : \quad & \delta_0 \neq 0 \quad \text{and} \quad \delta_1 \neq 1 \end{aligned}$$

For all of the six model H_0 **cannot** be rejected at a significance level $\alpha > 0.0001$.

Besides the regression we use the QLIKE loss function to evaluate the volatility forecast. Since this function can become negative for very small returns we use an equivalent alternative specification which is always positive:

$$QLIKE = \frac{r_t^2}{\hat{h}_t} - log\left(\frac{r_t^2}{\hat{h}_t}\right) - 1$$

The advantages and popularity of the QLIKE criterion arises from the fact that it penalises underestimating volatility more than overestimating it.

The model which minimizes the $QLIKE-function$ is the $GARCH(1,1)$ model with normal distributed error terms. The obtained rolling window forecast with re-estimation of the parameters after each step (visualisation in appendix 12) is shown in figure 9.

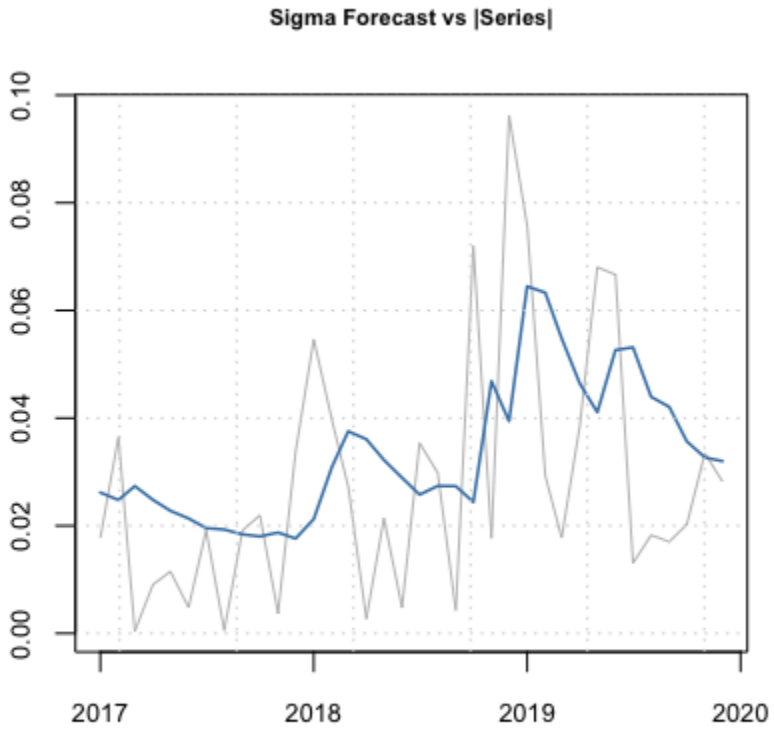

Figure 9: Volatility forecast GARCH(1,1)

In this model the shocks have the a symmetric impact on the volatility. The news impact curve of the initial model estimated on the training period is shown in figure 10

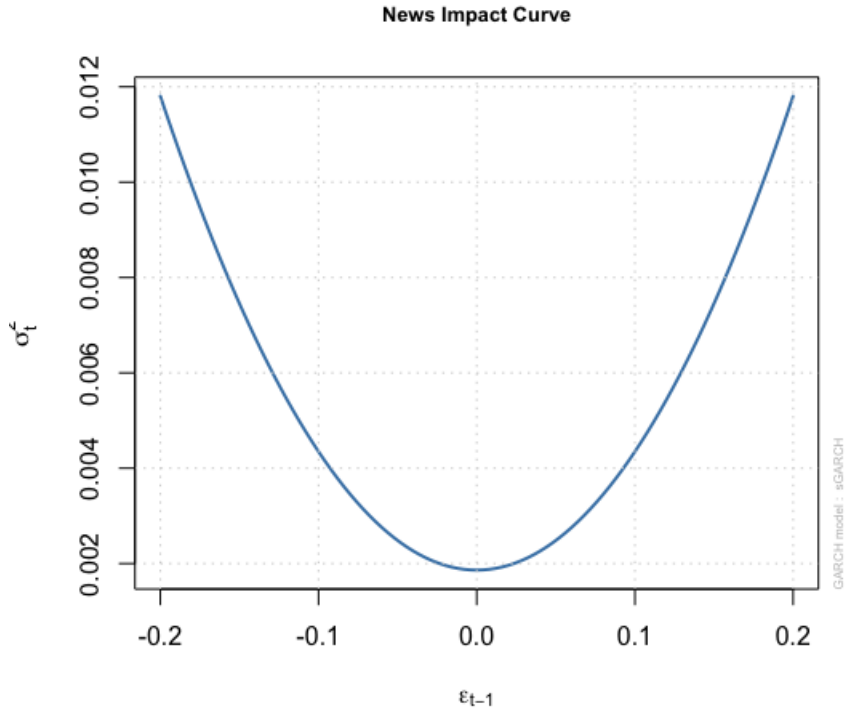

Figure 10: News impact curve GARCH(1,1)

For comparing reasons the News Impact curve of the $GJR-GARCH(1,1,1)$ is shown in the appendix **??**. In the GJR-GARCH framework negative shocks have an significantly higher impact on the volatility.

3.3 Return r_t forecast

With the volatility set-up we're also able to forecast the returns r_t for the period from January 2017 till December 2020. The error matrix is shown in table 5. Again we were using a rolling window forecast with an one step ahead forecast. To strengthen our model and forecast the parameters were re-estimated after each period like we did for the volatility forecast in the previous subsection. The most accurate forecast prediction was estimated by the $GARCH(1,1)$ model with standard normal distributed error term $u \sim N$. The forecast values against the realized returns are plotted in 11.

Model	RMSE	MAE
GARCH(1,1) $u \sim N$	0.0347	0.0250
GARCH(1,1,1) $u \sim st-t$	0.0353	0.0258
GJR-GARCH(1,1,1) $u \sim N$	0.0355	0.0248
GJR-GARCH(1,1,1) $u \sim st-t$	0.0353	0.0254
GARCH(1,1) + (book/market) $u \sim N$	0.0352	0.0251
GJR-GARCH(1,1,1) + (book/market) $u \sim N$	0.0353	0.02748

Table 5: Rolling-window error metrics

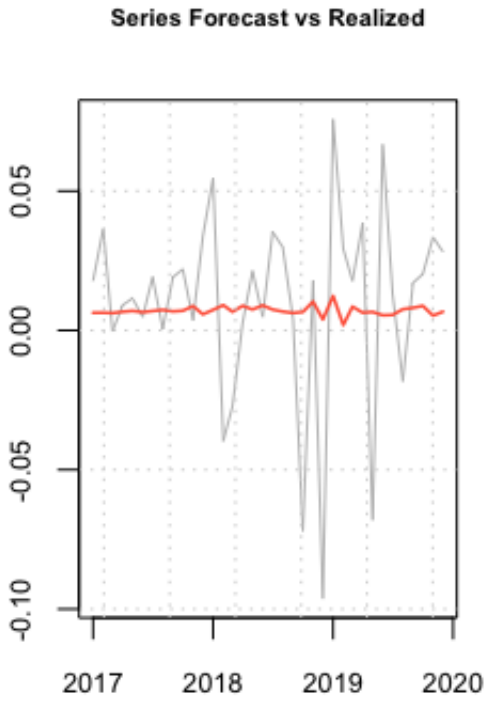

Figure 11: Return Forecast GARCH(1,1)

4 Forecast Evaluation

4.1 Error-based evaluation

Apart from the statistical tests described below, we evaluated the forecasts based on two well-established metrics in Time Series Forecasting: RMSFE (The Root Mean Square Forecasting

Error) and the MAFE (Mean Absolute Forecasting Error). In our case, the observation-wise forecast error can be written as:

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T} \tag{6}$$

T denotes our information set. We estimate a model on the data from y_0 to y_T. After that we conduct a rolling forecast on the data from y_T onwards, with a one-step-ahead forecast at each iteration. The aforementioned metrics are computed based on these errors e_t for each forecast.

$$\sqrt{\frac{1}{n}\sum_{t=1}^{\tau}(y_t - \hat{y}_t)^2} \tag{7}$$

$$\frac{1}{n}\sum_{t=1}^{\tau}|y_t - \hat{y}_t| \tag{8}$$

In both equations the Greek letter τ denotes the end of the evaluation data, to distinguish it from the running prediction index t. As the time series in question is the same across all models considered, no issues regarding the scale-dependence of this error metrics should arise. For a discussion of this issue see Hyndman, Athanasopoulos (2018) [4].

4.2 Comparing forecasts quality

Formal comparisons typically check if the differences between different forecasts are statistically significant or due to sampling variability. Luger develops the main difficulties encountered when formal testing [5]:

1. Forecast errors are generally not mean-zero or normally distributed
2. Multi-step forecasts are serially correlated and heteroscedastic
3. Competing forecasts tend to be contemporaneously correlated
4. The economic loss function may be asymmetric and not correspond to the usual statistical measures, such as absolute or squared forecast error.

Then, forecast errors are defined as

$$e_{it} = y_t - \hat{y}_{it} \tag{9}$$

With index i corresponding to the model forecast. In our case, *i=1, 2, 3* (i.e., ARMA, VAR, GARCH). The loss associated to each forecast is a function of it, denoted by $g(e_{it})$. In this assignment, we use the square function [6].

We define the loss differential as $d_t = g(e_{1t}) - g(e_{2t})$ and will test if the forecasts have the same level of accuracy if and only if d_t has null expectation for all t [6]. The null will thus be that the forecasts have the same accuracy.

$$H_0 : E(d_t) = 0$$
$$H_1 : E(d_t) \neq 0$$

In other words, the errors will be on average equally costly. If we reject the null, we choose the model that yields the smallest loss.

4.2.1 Diebold-Mariano (1995) (DM)

This well-known parametric test is given by

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \longrightarrow \mathcal{N}(0,1) \tag{10}$$

Where $\hat{V}(\bar{d})$ is the estimated asymptotic variance. Luger tells us that whenever an optimal forecast is produced, then the resulting h-step forecast errors follow a moving-average (MA) process of order $(h-1)$. Therefore, DM estimated $V(\bar{d})$ using the truncated kernel with $(h-1)$ bandwith for h-step forecasts. We have that

$$\hat{V}(\bar{d}) = \frac{1}{T}\left[\hat{\gamma}_0 + 2\sum_{k=1}^{h-1}\hat{\gamma}_k\right] \tag{11}$$

Where $\hat{\gamma}$ is the kth estimated autocovariance of d_t

$$\hat{\gamma} = \frac{1}{T}\sum_{t=k+1}^{T}(d_t - \bar{d})(d_{t-k} - \bar{d}) \tag{12}$$

Drawbacks of Diebold-Mariano: the normal distribution can be a very poor approximation of the DM test's finite-sample null distribution. Indeed, Diebold and Mariano showed that the DM test can have the wrong size, rejecting the null too often depending on the degree of serial correlation among the forecast errors and the sample size T [5].

4.2.2 Clark-West (2006, 2007) (CW)

Clark and West proposed a test statistic to evaluate the performance of forecasting models by testing whether the adjusted mean squared prediction error (MSPE) difference between 2 models is zero. These 2 models that are compared are: a parsimonious null model (model 1) that generates the data and a larger model (model 2) that nests the null model, which includes the variables to be predicted (and which would be reduced to the null model if some of its parameters are set to zero).

In these models h represents the h-step ahead forecasts the researcher is interested in predict. The period t forecasts of y_{t+h} from both models are denoted as $\hat{y}_{1t,t+h}$ and $\hat{y}_{2t,t+h}$ with their corresponding period $t+h$ forecast errors $y_{t+h} - \hat{y}_{1t,t+h}$ and $y_{t+h} - \hat{y}_{2t,t+h}$. The sample MSPEs $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are computed as sample averages of $(y_{t+h} - \hat{y}_{1t,t+h})^2$ and $(y_{t+h} - \hat{y}_{2t,t+h})^2$. If we consider P the number of predictions used to compute the models, we have the following:

$$\hat{\sigma}_1^2 = P^-1 \sum (y_{t+h} - \hat{y}_{1t,t+h})^2 \tag{13}$$

$$\hat{\sigma}_2^2 = P^-1 \sum (y_{t+h} - \hat{y}_{2t,t+h})^2 \tag{14}$$

$$\hat{\sigma}_2^2 - adj = P^-1 \sum (y_{t+h} - \hat{y}_{2t,t+h})^2 - P^-1 \sum (\hat{y}_{1t,t+h})^2 - \hat{y}_{2t,t+h})^2 \tag{15}$$

Where adj refers to the sample average of $(\hat{y}_{1t,t+h})^2 - \hat{y}_{2t,t+h})^2$.

Hence, to perform the CW test we would examine the difference of $\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj)$ instead of $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$, being the null hypothesis that both models have the same MSPE and the alternative, that the larger model (model 2) has a smaller MSPE than the first model[7]:

$$H_0 : \hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj) = 0$$
$$H_1 : \hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj) > 0$$

4.3 Empirical Results

The following table contains the out-of-sample rolling window forecasts for each of the three models.

Model	RMSFE	MAFE
ARMA	0.0355	0.259
$VAR(2)$	0.0369	0.0263
GARCH	0.0347	0.0250

Table 6: Rolling-window error metrics

In terms of static forecast evaluation metrics, the GARCH model comes out on top with both the lowest RMSFE and MAFE. It is followed by the ARMA model. The VAR(2) produces the highest errors on the out-of-sample data. This finding is interesting, as the ARMA and GARCH are relatively parsimonious models compared to the VAR, which has a much higher number of parameters.

In Table 7 the group summarized the results for the Diebold & Mariano test for all the fitted models from above. As one can see, the null hypothesis was rejected for the ARMA vs. VAR and ARMA vs. GARCH model, meaning that the model's forecasting errors don't have the same accuracy. Only for the GARCH vs. VAR model the null hypothesis was rejected, i.e. both models forecasting have the same accuracy.

Model	Diebold-Mariano Results	p-value
ARMA/VAR	Rejection of H_0 (different accuracy)	0.0106
ARMA/GARCH	Rejection of H_0 (different accuracy)	0.0000
GARCH/VAR	Non-rejection of H_0 (same accuracy)	0.1776

Table 7: Diebold-Mariano test results

Similar, yet not exactly the same results are derived from employing the Clark-West test. The only difference is that the Clark-West test does not reject the H_0 of equal predictability for the comparison between the ARMA and VAR model, where the Diebold-Mariano test issues a forceful rejection.

Model	Clark-West Results	p-value
ARMA/VAR	Non-rejection of H_0 (same accuracy)	0.736
ARMA/GARCH	Rejection of H_0	0.046
GARCH/VAR	Non-rejection of H_0 (same accuracy)	0.823

Table 8: Clark-West test results

Taken together, these results yield somewhat ambiguous results. While we have some evidence that the ARMA and GARCH have a significantly different predictive power and through the Clark-West test a further qualification that the GARCH models performs significantly better than ARMA, other such pairwise relationships could not be established with statistical power. We should add that the out-of-sample dataset consisting of the last three years (i.e. 36 observations) might be too short. Increasing the size of the test set could potentially yield more statistically robust results.

References

[1] Gideon Schwarz et al. "Estimating the dimension of a model". In: *Annals of statistics* 6.2 (1978), pp. 461–464.

[2] Gabriel Vasconcelos. *Formal ways to compare forecasting models: Rolling windows*. https://www.r-bloggers.com/2017/11/formal-ways-to-compare-forecasting-models-rolling-windows/. Accessed on 2021-04-08. Nov. 2017.

[3] Tuleau-Malot Genuer Poggi. "Variable selection using random forests". In: *Pattern Recognition Letters* 31 (2010), pp. 2225–2235.

[4] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. Vol. 2nd edition. OTexts: Melbourne, Australia, 2018.

[5] Richard Luger. "Exact Tests of Equal Forecast Accuracy with an Application to the Term Structure of Interest Rates". In: *Bank of Canada Working Paper* (2004).

[6] Umberto Triacca. "Lesson19: Comparing predictive accuracy of two forecasts: the Diebold-Mariano Test". In: *Dipartimento di Ingegneria e Scienze dell Informazione e Matematica, Universita dell Alquila* (2020).

[7] Todd E. Clark and Kenneth D. West. "Approximately normal tests for equal predictive accuracy in nested models". In: *Journal of Econometrics* 138.1 (2007). 50th Anniversary Econometric Institute, pp. 291–311. DOI: https://doi.org/10.1016/j.jeconom.2006.05.023.

A Parameter refit

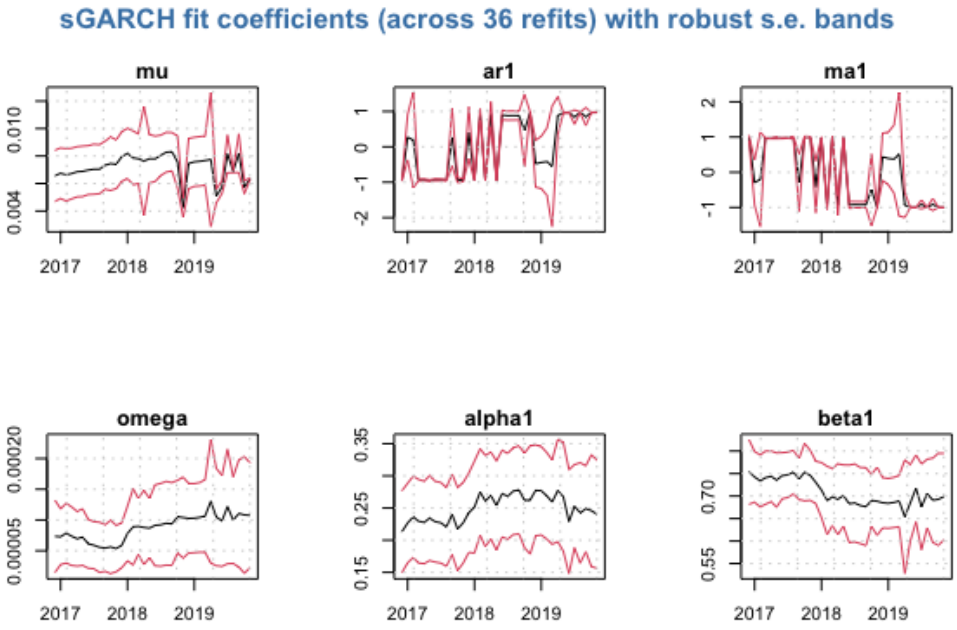

Figure 12: Parameter refit GARCH(1,1)

B News impact curve

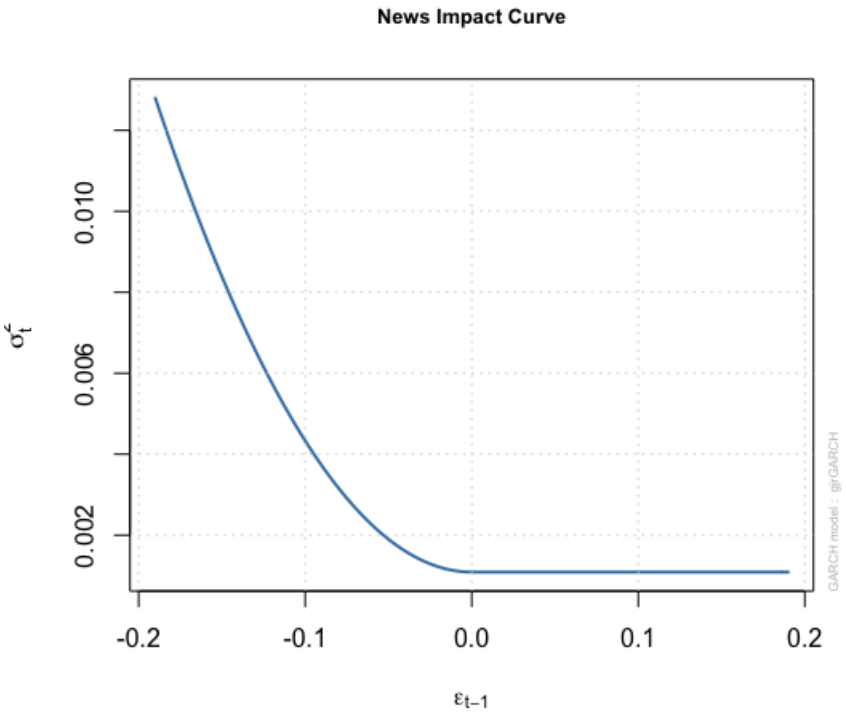

Figure 13: News impact curve GJR-GARCH(1,1)