ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Bachelor Thesis BSc^2 Econometrics/Economics

Adding economic variables to the HAR-RV model

Name Student: Ndubuisi Igwala

Student ID number: 452107

Supervisor: KP de Wit MSc
Second assessor: dr. X Xiao
Date final version: 05/07/2020

Abstract

We introduce various methods aimed at improving the HAR-RV model introduced by Corsi [5] through the addition of economic variables. These economic variables are taken from Welch and Goyal [25]. Through applying LASSO, principal components and partial least squares techniques we find models which significantly improve upon the in-sample forecasting performance of the HAR-RV model. The introduction of a LASSO-VAR proves to be unsuccessful. The LASSO-VAR model is not able to significantly beat the HAR-RV model out of sample. Furthermore, we perform a simulation to attempt to display the stylized facts of stock returns through using the HAR-RV model.

Keywords— HAR, LASSO, economic variables, VAR, Principal Components, Partial Least Squares

Contents

1 Introduction

Measuring and predicting volatility has been a well-researched topic for decades. This can be attributed to the fact that this field of research is greatly applicable in practice. Before volatility became an important topic, measuring and predicting the equity premium was already a well-established research topic. This arguably changed when Markowitz published his groundbreaking research on portfolio selection Markowitz [15]. He was the first who put a large emphasis on the risk side of selecting portfolios. A new field of research was born. Since then, major developments occurred in the field of volatility estimation. Some major examples are the GARCH model introduced by Engle [8] and the more economic CAPM model introduced by Sharpe [21] and Lintner [12] which actually builds upon the aforementioned paper of Markowitz. These models and developments have been of great use to investors who sought a way to incorporate risk into their decisions. The measuring of risk is so important that there is even a whole separate branch in finance called risk management.

This paper focuses on a small part of a wide array of available topics within this field. The Heterogeneous Autoregressive Realized Variance (HAR-RV) model as introduced by Corsi [5] will be extended with economic variables. The HAR-RV model is a cascade model in the sense that the model aims to predict realized volatility with dependent variables which have heterogeneous interval lengths. Those are in fact also realized volatility. The dependent variables are daily, weekly and monthly realized volatility. As an extension we use economic variables to further develop the HAR-RV model and possibly improve its accurateness. In order to do this, we use monthly data on economic variables previously used for research carried out by Welch and Goyal [25]. They discovered that some of the economic variables inherit predictive power on the risk premium. Since, normally, volatility is easier to predict than the risk premium we can expect positive results. We choose these economic variables, since these variables have already proven their worth in research aimed at forecasting and predicting. Furthermore, these variables are linked to volatility in different ways. This will be elaborated on in a later section. The aim of this research is thus ultimately to discover whether this addition of economic variables yields even more accurate forecasting results than the ordinary HAR-RV model. The addition of economic variables could make the HAR-RV model more comprehensible for ordinary investors. Moreover, it could give investors yet another model with which they could analyse risk.

Hence, the following research question is formulated: "To what extent do economic variables increase the predictive power of the HAR-RV model?" The predictive power will be measured with simple forecast comparison metrics. Both in-sample and out-of-sample performance will be compared. The HAR-RV model is used as a benchmark in order to easily evaluate whether

the addition of economic variables yields better results.

The academic relevance of this paper lies in the understanding of volatility in stock markets. The HAR-RV model provides great results, but its accurateness can be increased, and it can be made more intuitive. When proven to be successful, future researchers could aim to find better fitting economic variables. Alternatively, the models provided in this paper can be used as an improved benchmark over the basic HAR-RV model. Moreover, this paper aims to discover whether economic variables are a sound addition to the HAR-RV model and help to improve its accurateness. It has been shown by, for instance Christiansen, Schmeling, and Schrimpf [4], that economic variables aid in explaining stock volatility. We aim to find out whether the HAR-RV model benefits similarly or whether it is a model which does not improve through usage of economic variables.

This paper is socially relevant, since it helps investors to better understand and predict the volatility of the market. This makes them able to make better decisions and thus decrease their exposure to stock market risk. This could alleviate losses on days with high volatility. Ultimately, this could lead to the survival of financial firms during highly volatile times and preserve numerous jobs. Moreover, institutions such as pension funds would be able to better forecast volatility in the market. This is of great use, since these funds could make better decisions by doing this resulting in more certainty for taxpayers. Another example is the use of the aforementioned forecasts by the central bank of countries. They can better direct their policies in the case of better volatility forecasts. This benefits the economy as a whole and thus helps individuals which are not even employed in the financial world.

We first provide an extensive overview of the background of the subject and the related literature. Thereafter, we will elaborate on which data is used and how it is prepared for the research. Then, the used methods will be extensively explained. Logically, the results section follows. This paper is then concluded by a discussion of the results and a conclusion.

2 Literature review

2.1 Economic background

As mentioned before, measuring volatility has been one of the main focuses of research in the world of quantitative finance. Yet, also in financial economics, risk has been pivotal. An example of a model where volatility played a great part is the CAPM model introduced by Sharpe [21]. Although this has been empirically dis-proven in more recent times, it is still a widely acknowledged and used model. Its strength lies in the arguably simple linkage between

the expected return on a stock and the overall market. It provides a clear-cut formula which enabled investors to make predictions on the potential return of a stock. Risk comes into play when computing the β of a stock. This model is linked to the Sharpe ratio introduced by Sharpe [22]. This simple ratio between the expected return of a portfolio or stock in excess of the risk-free rate and the standard deviation of the portfolio provides investors a basic interpretation into the safeness of a certain return. The Sharpe ratio and the CAPM model lie at the basis of more advanced risk models which were developed later. Before the CAPM model was introduced to the world, Markowitz [15] introduced his work on portfolio selection. He developed a portfolio selection theory where a trade-off between return and risk was paramount. This work has been widely regarded as one of the most influential papers on portfolio optimization. However, there are some clear limitations to the research such as the high number of estimates needed to perform the calculations. More recently, improvements were made over the original portfolio theory. The introduction of downside risk and asymmetric distributions has led to empirically better results. An example of such a paper is the work of Vercher, Bermúdez, and Segura [24]. Since those major works a lot of research on possible influences on risk on the stock market has been carried out.

Now, we discuss how economic variables are potentially able to improve the aforementioned risk models in general. Economic variables have long been an interesting approach to forecasting certain dependent variables. They are used to predict similar macro variables such as GDP, but also to predict stock returns and the equity premium. These differing models serve different fields of work. Investors have benefited greatly from research on how economic variables can aid in predicting the equity premium. Welch and Goyal [25] analysed whether univariate forecasts using economic variables as independent variables yielded significant results in predicting the equity premium. The results were all but positive. Welch and Goyal concluded that the variables do not hold enough forecasting power on their own. Nevertheless, this work has been a building block for future research into economic variables. Researchers were interested in whether combining these univariate forecasts in some form yielded better results. This showed to be a promising field. For instance, Rapach, Strauss, and Zhou [20] find significantly better out-of-sample results through combining univariate forecasts. Since then, there have been studies aiming to find an optimal weighting to such a combination forecast. A promising approach to tackle this phenomenon is machine learning. Machine learning has, in recent years, played a big part in forecasting the equity premium. Gu, Kelly, and Xiu [11] used a vast array of available techniques to forecast the equity premium and subsequently compared their performance. Overall, machine learning techniques proved to do a good job in explaining the equity premium.

Finally, we discuss how our selection of economic variables is potentially able to improve forecasting daily realized volatility. We chose the economic variables used in Welch and Goyal [25], since it is a wide array of different variables. Economic theory laid out in Mele [16] suggests that variables which have to do with time-varying risk premia are great predictors of risk. This indicates that yield spreads, interest rates and equity valuation series are potentially great predictors when it comes to volatility forecasting. **B/M** and the yield and return variables which we use in this paper are good examples of such variables. Baskin [2] show that there is a significant relationship between dividend policy and stock market risk. They find that stock market risk is inversely related with dividend yields. Hence, the addition of **D/P**, **D/Y**, **E/P** and **D/E** potentially aids in explaining daily realized volatility. Furthermore, **SVAR** is a very good indicator of daily realized volatility since its definition is the squared daily returns on the SP500 index.

2.2 Econometric background

This paper, as mentioned earlier, builds on the research carried out by Corsi [5]. Hence, it is only logical to go over alternative extensions of his research. Corsi pointed out in his paper that there were clear extensions available for future research. He followed up on one of them by introducing jumps and heterogeneous leverage in Corsi and Reno [6]. The leverage effect states that positive returns bring about a different increase in the realized volatility than a negative return with the same size. He found that continuous volatility, leverage and jumps all contribute greatly to the accurateness of the predictions. Most important, he found great forecasting power for the negative return. This indicates that this model benefits greatly from the leverage effect, since the model responds differently to a positive jump in return than to an equally sized negative jump in return. Realized volatility is defined below:

$$RV_t = \sqrt{\sum_{j=0}^{M-1} r^2_{t-j\cdot\Delta'}} \tag{1}$$

where $\Delta = 1/M$, and $r_{t-j\cdot\Delta'} = p(t - j\cdot\Delta) - p(t-(j+1)\cdot\Delta)$ is the return at interval Δ. Here, t defines the time period for which the realized volatility is measured and j defines the time within this time period.

The HAR-RV framework as laid out by Corsi is not the only part of quantitative finance which works with intraday data. Andersen et al. [1] use intraday data to measure and model return volatility and distribution. They establish a measurable link between realized volatility forecasting and the conditional covariance matrix.

Besides the HAR-RV model, there exists a plethora of alternative models to accurately predict risk. One famous example is the Value-at-Risk model invented by employees of JPMorgan in the wake of the 1987 market crash. This model aims to quantify the maximum loss that could be had with a certain probability. After the investment bank made their model publicly available, it became a widespread used model in the world of financial risk. Another major model where risk plays a significant role is the Black-Scholes model as invented by Black and Scholes [3]. The aim of this model is to compute the price of European options. It has been a pivotal paper in option pricing theory for many years.

Now, after briefly mentioning some important papers involving risk, let us focus more on realized volatility as used in this research. More standard econometric methods have failed to incorporate the stylized facts present in financial data. These stylized facts include asymmetry in the returns, fat tails of the return distribution and a slow decay of autocorrelation in the returns as explained by Malmsten, Teräsvirta, et al. [14]. Moreover, they find that the GARCH model, amongst others, fails to incorporate all aforementioned stylized facts. Another type of model used to produce the stylized facts are stochastic volatility models. Those are based on the assumption that the log price behaves as follows:

$$\partial p(t) = \mu(t)\partial t + \sigma(t)\partial W(t) \tag{2}$$

where $p(t)$ is the log price, $\mu(t)$ the finite variation process, $W(t)$ is a Brownian motion and $\sigma(t)$ is a stochastic process. Andersen et al. [1] showed that the modelling of realized volatility outperforms the use of GARCH and stochastic volatility models. One such way of modelling the realized volatility directly has been carried out by Corsi [5]. His HAR-RV model is based on the Heterogeneous Market Hypothesis by Müller et al. [18] which assumes that market participants analyse events and data with different time horizons. The implementation of differing timescales for realized volatility as explanatory variables is a consequence of this thinking. A benefit of this model is that it is parsimonious. It is simple to perform and understand this model. The model brings about a cascade pattern from low to high frequencies. Short-term traders care about long-term volatility, because it affects the size of risk in the future. When short-term traders act upon this information, they create short-term volatility. On the other hand, long-term traders do not care about short-term risk. Therefore, the cascade pattern is justified. The cascade model looks as follows:

$$\sigma_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \tilde{\omega}_{t+1d}^{(d)} \tag{3}$$

The factors are the past realized volatilities at different frequencies. The functional form of a time series model can be derived from the equation above. Note that:

$$\sigma_{t+1d}^{(d)} = RV_{t+1d}^{(d)} + \omega_{t+1d}^{(d)} \tag{4}$$

Briefly, the parsimoniousness and the ability to replicate financial data so well is the reason the HAR-RV model is a great model in theory to tackle the realized volatility issue. Now, $\sigma_{t+1d}^{(d)} = RV_{t+1d}^{(d)} + \omega_{t+1d}^{(d)}$ can be substituted into the cascade model to get a time series representation of the cascade model. The resulting model is discussed on more elaborately in the Methodology section.

3 Data

In this section we discuss the data used to carry out this research. Firstly, we will explain the use of the realized volatility data. Three time series of indices were collected from the Oxford-Man Institute of Quantitative Finance. Namely, the AEX, FTSE and S&P500. The sample period ranges from the 1st of January 2000 through May 2020. It is important to note which daily realized variance metric is used, since there are multiple alternatives available on the Oxford-Man database. We use the rk two scale metric for this paper. This metric is chosen, since it most closely resembles the way in which daily realized volatility is computed in the paper of Corsi [5]. Corsi employed the two scales estimator proposed by Zhang, Mykland, and Aıt-Sahalia [26]. We transform the daily variance into annualized daily volatility through the following formula:

$$RV_t^{(d)} = (RVar_t^{(d)} * 252)^{\frac{1}{2}} * 100 \tag{5}$$

where we annualize the realized variance by multiplying it with 252 (trading days in a year). Subsequently we take the square root to get the realized daily volatility and multiply it with a 100 in order to obtain a percentage.

Now, we elaborate on the economic variables which are used as an extension in this paper. Since Welch and Goyal [25] has shown to have had reasonable results in the field of equity premium prediction, the variables he considered will be considered for this paper. Since realized volatility is expected to be easier to forecast than the equity premium, it is reasonable to assume that these variables hold predictive power in this setting. The aim is to see whether the addition of these variables result in a better functioning HAR-RV model. In order to fit the existing framework of this paper the monthly data needs some handling. The monthly observations are used for each daily realized volatility such that the monthly value is taken on for each day in

the respective month. This is done since daily data on economic variables was hard to come by. Moreover, the values of these economic variables do not change a lot in a short span of time. This makes this way of handling the data not as bad as it initially looks. Yet, it is sub-optimal. A list and short description of the economic variables is shown below:

- **D/P**: logarithm of the ratio of dividends on the S&P 500 index to the price of the index itself where dividends are measured as a one-year moving sum.
- **D/Y**: logarithm of the ratio of dividend to lagged stock prices.
- **E/P**: logarithm of the ratio of earnings to price where earnings are computed through a one-year moving sum.
- **D/E**: logarithm of the ratio of dividend to earnings.
- **SVAR**: stock variance of the S&P 500 index which is calculated through summing the squared daily returns.
- **B/M**: Dow Jones Industrial Average index book-to-market ratio.
- **NTIS**: net equity expansion, which is defined as the ratio of net issues by NYSE-listed stocks to the end-of-year market capitalization of all stocks on the New York Stock Exchange (twelve-month moving sums).
- **TBL**: interest rate on a secondary market three-month US Treasury bill as an annual percentage.
- **LTY**: yield on long-term government bonds as an annual percentage.
- **LTR**: return on long-term government bonds as an annual percentage.
- **TMS**: term spread, which is defined as the difference between long-term government bond yield and the rate on US Treasury bills as an annual percentage.
- **DFY**: default yield spread, which is defined as the difference between corporate yield which have been rated AAA and BAA by Moody's as an annual percentage.
- **DFR**: default return spread, which is defined as the difference between long-term government bond yields and long-term corporate bond yields as a percentage.
- **INFL**: Consumer Price Index (CPI) inflation as a percentage. The inflation is calculated with a lag of one month to incorporate information delay.

Table 1

This table reports descriptive statistics for the 14 economic variables used to extend the HAR(3) model. The data has been retrieved from Amit Goyal's web page at `http://www.hec.unil.ch/agoyal/` *The sample period ranges from January 2000 through December 2019. All values are rounded to three decimal points.*

Variable	Mean	Standard deviation	Minimum	Maximum
D/P	-3.988	0.191	-4.524	-3.281
D/Y	-3.985	0.193	-4.531	-3.295
E/P	-3.140	0.388	-4.837	-2.566
D/E	-0.848	0.455	-1.244	1.380
SVAR	0.296	0.533	0.015	5.809
B/M	0.281	0.066	0.121	0.441
NTIS	0.004	0.018	-0.029	0.058
TBL	-1.645	1.787	-6.170	-0.010
LTY	-4.036	1.208	-6.660	-1.750
LTR	0.621	3.124	-11.240	14.430
TMS	2.416	1.315	-0.590	4.530
DFY	1.044	0.426	0.550	3.380
DFR	0.049	1.892	-9.750	7.370
INFL	-0.178	0.373	-1.222	1.915

The descriptive statistics of these economic variables are given in Table 1 above. The series of particular interest to us are the time series with a high standard deviation. The high standard deviation implies there is a lot of variability in these variables. If there is consistent effect of certain economic variables on the daily realized volatility, these variables contribute to the variation in our dependent variable. The return on long term government bonds is the most volatile economic variable.

Now, let us specify the data used to compute the daily, weekly and monthly realized variance. The original data is gathered from the online Oxford-Man database. It comprises daily realized variance estimates for the AEX, FTSE and SP500 indices. The daily realized variances are two tick realized variances in order to comply with the original work of Corsi [5]. Subsequently, data transformations were performed to create three time series of annualized daily realized volatility as explained at the beginning of this section. The data is sampled from January 2000 through December 2019.

4 Methodology

This section elaborates on the methods with which the research question is answered. The methodology section is twofold, since we first give a brief overview of the HAR model introduced by Corsi [5]. Subsequently, the implementation of the economic variables will be discussed. Corsi [5] used the following definition for the realized variance:

$$RV_t^{(d)} = \sqrt{\sum_{j=0}^{M-1} r^2_{t-j\cdot\Delta'}} \tag{6}$$

where $\Delta = 1d/M$, and $r_{t-j\cdot\Delta'} = p(t - j\cdot\Delta) - p(t-(j+1)\cdot\Delta)$ is the intraday return at interval Δ. Here, t defines the day and j defines the time within the day. Corsi [5] used three different time windows to construct his cascade model. Namely, one day, one week and one month. For a single day, he calculated tick-by-tick realized volatility estimates according to the two scales estimator method proposed in Zhang, Mykland, and Aıt-Sahalia [26]. This estimator combines two distinct estimators; an estimator which computes the realized volatility through summing the squared returns for each tick and an estimator which averages the realized volatility over the day through introducing clusters. When a realized volatility estimate is made for a week or for a month the daily estimates are simply aggregated. For instance, the weekly realized volatility at time t looks as follows:

$$RV_t^{(w)} = \frac{1}{5}(RV_t^{(d)} + RV_{t-1d}^{(d)} + \cdots + RV_{t-4d}^{(d)}). \tag{7}$$

Here, the superscript (w) refers indicates the usage of a weekly realized variance. Similarly, (m) would indicate a monthly aggregation of the daily realized variances (which contains 22 trading days). Now, the HAR-RV model can be discussed Corsi [5]. The model is specified below:

$$RV_{t+1d}^{(d)} = c + \beta^{(d)}RV_t^{(d)} + \beta^{(w)}RV_t^{(w)} + \beta^{(m)}RV_t^{(m)} + \omega_{t+1d} \tag{8}$$

The hallmark of this model is its heterogeneity in the interval lengths. From the model it can be understood that the daily realized volatility depends on a constant and the daily, weekly and monthly realized volatility computed the prior day. Moreover, an error term is added. This model can be considered an HAR(3)-RV model, since it contains three differing time intervals.

The basic model has now been specified. Subsequently, the addition of economic variables as an extension to the model needs elaboration. The addition of economic variables is done in three ways. Firstly, principal components are used to reduce the dimensionality and capture the maximum amount of variance of the economic variables with a few factors. Secondly, a LASSO penalty term according to Tibshirani [23] is be added to the ordinary HAR-RV model with the economic variables. This serves the same purpose as the principal components technique since it shrinks the amount of total economic variables considered. Both techniques find the most appropriate variables while both techniques simultaneously prevent overfitting. These models seem suitable, since this paper deals with arguably many independent variables. Besides these

methods, the technique of partial least squares is also used to set up a model. This method is similar to principal component analysis and will be elaborated on further in a following section. Ma, Wahab, and Zhang [13] show that partial least squares regression performs better than PCR in forecasting stock volatility. Hence, we find it to be a valid model to include in our analysis.

4.1 In-sample

4.1.1 HAR-PCR

Firstly, we introduce the technique of principal component analysis as invented by Pearson [19]. In order to evaluate whether incorporating principal components into our baseline model aids in predictability, all variables are considered in the analysis. Hence, every economic variable and the three realized volatility variables are used to create the principal components. Before constructing the principal components, the data set needs to be defined. The data matrix $\mathbf{X}$ contains p n-dimensional vectors $\mathbf{x}_1, ..., \mathbf{x}_p$ For instance, the third column represents $\mathbf{x}3$ with observations on the third variable. The goal of principal component analysis is to find a linear combination of the columns of $\mathbf{X}$ such that maximum variance is attained. Finding this maximum variance boils down to finding the solution to the following equation:

$$\Omega \cdot a = \lambda \cdot a \tag{9}$$

where a is a vector containing the weights for the linear combination which needs to be an eigenvector. λ is the corresponding eigenvalue, while Ω is the sample covariance matrix.

Now, the largest eigenvalue λ_1 corresponds to eigenvector a_1. Moreover, the variance of $\mathbf{Xa}$ is equal to λ. The linear combinations $\mathbf{Xa}_k = \sum_{j=1}^{p} a_{jk}\mathbf{x}_j$ are commonly called the principal components of the dataset. This technique will be put to work in a simple way for this research. The matrix $\mathbf{X}$ will consist of all economic variables and the three time series according to the three different time horizons (daily, weekly and monthly). The in-sample predictions are then made with the principal components of the data matrix X. The in-sample period runs from January 2000 through May 2020. In order to determine the number of principal components we take a look at the percentage of variance explained by n principal components. This graph can be found in the appendix. The number of used components for each time series will be decided on by looking at the increase of percentage of the variance explained when an extra component is added. We elaborate on this further on in the paper. These numbers of components are also used for the partial least squares model which is introduced in the next section. We provide

graphs in the appendix to illustrate the amount of variance captured by the number of principal components.

4.1.2 HAR-PLR

Partial least squares regression is a statistical method which works particularly well when variables are multicollinear. Since this is the case with our three realized volatility variables with different time horizons, this method is suitable for this research. We use the partial least squares regression technique according to Gerlach, Kowalski, and Wold [10]. PLS functions similarly to PCR. Both techniques aim to maximize the variance through linear combinations of the variables. However, PCR only does this through the variance in the $\mathbf{X}$ matrix. PLS tries to find a linear combination of the variables in $\mathbf{X}$ that explain maximum variance in the response variable $\mathbf{y}$. First, latent variables $\mathbf{t}$ and $\mathbf{u}$ are created. These are linear combinations of the original economic variables. The latent variables are defined by:

$$(v_h, w_h) = argmax(cov[\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h]) \tag{10}$$

under the constraints:

$$||\mathbf{a}_h|| = 1, ||\mathbf{b}_h|| = 1 \tag{11}$$

and

$$\mathbf{a}_h^T \mathbf{X}^T \mathbf{X} \mathbf{a}_i = 0 \tag{12}$$

for $1 \leq h < i$. Now, $\mathbf{t}_i = \mathbf{X}\mathbf{v}_i$ and $\mathbf{u}_i = \mathbf{Y}\mathbf{w}_i$. $\mathbf{U}$ and $\mathbf{T}$ denote matrices compromised of the vectors $\mathbf{u}_i$ and $\mathbf{t}_i$. Subsequently, $\mathbf{U}$ is regressed on $\mathbf{T}$. The regression formula can be rewritten to show the relationship between $\mathbf{X}$ and $\mathbf{Y}$:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{V}(\mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{Y} \tag{13}$$

where $\mathbf{V}$ is a matrix containing the weighting vector $\mathbf{v}_i$. This equation can be transformed easily to obtain a formula for PLS regression coefficients:

$$\hat{\mathbf{B}} = \mathbf{V}(\mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{Y} \tag{14}$$

The one-day ahead in-sample predictions are made in a similar way as those made by principal components. Once again, the data matrix $\mathbf{X}$ consists of all economic variables plus the three time series for a certain index. Predictions are subsequently made using $\hat{\mathbf{B}}$ to predict the daily realized volatility. The in-sample period runs from January 2000 through May 2020.

4.1.3 HAR-LASSO

The LASSO penalty is easily incorporated into the model. We consider the following model including all economic variables:

$$RV_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \gamma_j' z_{j;t} + \omega_{t+1d} \tag{15}$$

Here, γ_j' is the k-dimensional vector of the coefficients for the economic variables. Through incorporating the LASSO penalty, the penalty function of the regression changes. The penalty function including the LASSO term looks as follows:

$$\hat{\zeta}^{LASSO} = \sum_{i=1}^{n} (RV_{t+1d;i}^{(d)} - c - \beta^{(d)} RV_{t;i}^{(d)} - \beta^{(w)} RV_{t;i}^{(w)} - \beta^{(m)} RV_{t;i}^{(m)} - \gamma_j' z_{j;t;i})^2 + \lambda \sum_{j=1}^{k} |\gamma_j| \tag{16}$$

where the hyperparameter $\lambda \geq 0$. A larger λ corresponds to a larger weight on the LASSO penalty in the regression. In the case where $\lambda = 0$, the penalty function collapses to the OLS penalty function. The value of λ is chosen to be 0.1.

4.2 Out-of-sample LASSO-VAR

To analyse the out-of-sample performance of our models we make one day, one week and two weeks ahead forecasts with a rolling window of 1000 observations. Subsequently, we calculate the same metrics as calculated for the in-sample forecasts. We follow a different approach for the out-of-sample forecasts we intend to make in this research. Similarly to the in-sample section, the proposed models are compared to ordinary AR-models and the general HAR-model. Since the aforementioned extensions of the HAR-model are not well suited for multi-step ahead forecasting, we follow a different approach for this part of the comparison. The aforementioned in-sample methods are not well suited, since they do not use an AR(p) type structure where the value on a forecasted day cannot be deduced from the values on the days before. For instance, principal components technique cannot create principal components when the economic variables are not updated based on the previous day. Therefore, we will introduce a VAR-model including the daily realized volatility and some selected economic variables. In order to prevent overfitting, not all economic variables are added to the VAR model. We first perform a LASSO regression where the daily realized volatility series are regressed on all economic variables. Subsequently, the variables which are given a non-zero value are included in the VAR model. This is

done for every step of the rolling window. This will result in performing the following regression:

$$RV^{(d)}_{t-1000:t-1} = c + \gamma'_j z_{j;t-1000:t-1} + \omega_{t-1000:t-1} \tag{17}$$

The penalty function of this regression looks as follows:

$$\hat{\zeta}^{LASSO} = \sum_{i=1}^{n} (RV^{(d)}_{t-1000:t-1;i} - c - \gamma'_j z_{j;t-1000:t-1;i})^2 + \lambda \sum_{j=1}^{k} |\gamma_j|. \tag{18}$$

The values of λ we test are 1, 1.5 and 2. The lag order is determined by the partial autocorrelation structure of the daily realized volatility series. These autocorrelation structures are given in the Appendix in Figure 10, 11 and 12. We see that it is most appropriate to include 6 or 7 lags to all LASSO-VAR models. However, in order to reduce computational time, we restrict all LASSO-VAR models to have 5 lags. Subsequently, we create a VAR model with which we forecast one day, one week or two weeks ahead. This model looks as follows:

$$RV^d_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_j y_{t-j} + \epsilon_t \tag{19}$$

where y_t is the vector with all observations of the included economic variables at time t. j is the chosen number of lags. However, it will not be greater than 5 in order to reduce computation time.

4.3 Model comparison

After having estimated the two aforementioned models, they are compared to the original HAR-RV model by calculating the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) displayed below:

$$MAE = \frac{1}{N} \sum_{t=1}^{N} |RV_t - \hat{RV_t}|, \quad RMSE = \frac{1}{N} \sqrt{\sum_{t=1}^{N} (RV_t - \hat{RV_t})^2}, \tag{20}$$

where $\hat{RV_t}$ is the estimated daily realized volatility. The in-sample one-day ahead forecasts are made through using the LASSO selected variables for our regression as well as the PLR and PCR models. These variables are selected through fitting the data and minimizing the loss function in equation (12). This way, certain variables are left out of the predictive regression so that we do not end up with a so called "kitchen-sink" regression. This leads to a more parsimonious model. It must be noted that a kitchen-sink model could produce reasonable results, since we test this LASSO regression in-sample. The risk of overfitting is thus greatly reduced.

Moreover, we also calculate the Diebold-Mariano test statistic according to Diebold and Mariano [7]. This test tests whether a particular forecast significantly outperforms another forecast in terms of predictive power. We compare our three in-sample and three out-of-sample LASSO-VAR extension models to the general HAR(3) model to find out whether there is a significant improvement. Through analysing this metric, it is easier to make valid conclusions concerning the performance of the forecasting models.

Finally, we perform a Mincer-Zarnowitz regression for all our forecasts as originally proposed by Mincer and Zarnowitz [17]. The R^2 of these regressions will be shown in the results tables together with the MAE and RMSE. The Mincer-Zarnowitz regression looks as follows:

$$RV^d_{t+h} = \alpha + \beta RV^d_{t+h|t} + \epsilon_{t+h|t} \tag{21}$$

where the actual series is regressed on a constant and the forecasted series. A forecasted series is said to predict well when α is close to zero and β is close to one.

4.4 Simulation

In this section the detailed nature of the general HAR-RV model will be illustrated. We prove that the HAR-RV model produces rich dynamics which closely reflect the real time series of AEX realized volatility. We stay close to the simulation carried out by Corsi [5] through using the same parameters and model specification. However, since we just have daily estimates of realized variance instead of tick data, we slightly alter the model. Moreover, the results will not be as similar to each other as displayed in the paper of Corsi [5]. The loss of high frequency data is to blame for this inconveniency. The simulated model is specified below:

$$r^{(d)}_t = \sigma^{(d)}_t \epsilon_t \tag{22}$$

$$\sigma^{(d)}_{t+1d} = c + \beta^{(d)} RV^{(d)}_t + \beta^w RV^{(w)}_t + \beta^m RV^{(m)}_t + \omega^{(d)}_{t+1d} \tag{23}$$

where $\beta^{(d)} = 0.36$, $\beta^{(d)} = 0.28$ and β^m 0.28.

5 Results

First, we will provide a table in which the HAR(3) estimation is stated for the three time series at hand. This Table 2 is given below.

We see very similar regression results for all three time series. The coefficients for the daily and weekly realized volatility are similar. Moreover, the general conclusion is that the monthly

Table 2

In-sample estimation results of the least squares regression of HAR(3) model for AEX (5194 daily observations), FTSE(5140 daily observations) and S&P500(5114 daily observations) time series. The sample data ranges from January 2000 to May 2020. Reported in parentheses are the t-statistics based on standard errors computed with Newey–West correction for serial correlation of order 5.

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

	AEX	FTSE	SP500
c	0.799	0.869	0.683
	(4.875)	(3.934)	(3.715)
β_d	0.473	0.326	0.482
	(11.586)	(7.093)	(7.806)
β_m	0.413	0.495	0.390
	(8.658)	(7.320)	(4.859)
β_m	0.059	0.113	0.076
	(1.847)	(2.830)	(1.936)

realized volatility does not greatly impact the realized volatility of the next day. However, it must be noted that the coefficients for the monthly realized variance for the AEX and SP500 series are not significant at a 5% level. Here, we assume that the t-distribution is approximated by the normal distribution, since the sample size is large. Hence, variables with a t-statistic above 1.96 are considered significant.

Now, we will look at the goodness of fit for an AR(22) and the HAR(3) model in Table 3 below. We note a few interesting results from the table above. First of all, when we look at the

Table 3

Results of the F-test for multiple hypothesis testing between the unrestricted AR(22) model and the restricted HAR(3) model (1% critical values in parentheses) and their respective Akaike information criterion (AIC) and Bayesian information criterion (BIC). The sample period runs from January 2000 through May 2020 (5194 observations for AEX, 5140 observations for FTSE and 5114 observations for SP500).

	AEX	FTSE	SP500
F-test	6.805	3.487	5.149
	(1.908)	(1.908)	(1.908)
	AIC		
AR(22)	29609	30641	29210
HAR(3)	29697	30767	29351
	BIC		
AR(22)	29753	30785	29353
HAR(3)	29723	30793	29378

AIC, all AR(22) models have a better fit according to this metric. This is somewhat surprising since it is expected that the HAR(3) model is able to outperform a simpler AR(22) model.

Roughly the same margin can be seen across all three time series. The BIC tells a different story. The BIC is smaller or equal for all models. This can be explained by the fact that the BIC metric punishes more complex models heavier. The AR(22) model contains considerably more variables than the HAR(3) model. Hence the HAR(3) model, which only has 4 variables, is preferred according to the BIC. The F-statistics tell us that the null hypothesis that both models fit the data evenly well can be rejected, since the F-statistics are considerably larger than the critical values. The HAR(3) and the AR(22) model fit the model with a different accuracy. This is true for all three realized volatility series.

5.1 Simulation

In this section we will compare the simulated time series to the actual FTSE index. Firstly, we will visually inspect the time series for the returns of the FTSE index and our simulated time series. The time series are shown below in Figure 1 and 2. We note that the magnitude of the returns is similar. However, we note some stark contrasts. The real return series displays more clustering in volatility, whereas the volatility of the simulated series is much more constant. This can also be seen in Figure 3 and 4. Once again, we see a similar magnitude of the values. Yet, once more the real daily realized volatility displays more variation in the realized volatility. Clearly, the simulation somewhat failed to create an autocorrelation function which is expected for a major index. Figure 5 and 6 are the proof of this conclusion. We see that the real series displays a partial autocorrelation structure which is not repeated for the simulated index. Therefore, it is hard for the simulated realized volatility to contain signs of volatility clustering. These differences are likely due to the issue of lack of tick data. The simulated realized volatility series is simulated daily in order to approach the HAR(3) model used in this paper.

5.2 In-sample

In this section we will state the results for the in-sample forecasts. In Table 4 below we compare the performance of several AR-models as well as the regular HAR-model. On top of that, we add the HAR-LASSO, HAR-PCR and HAR-PLR models. The number of components used for the PCR and PLR models is determined by looking at Figure 7, 8 and 9. We incorporate k components when the $k+1-k < 0.01$. This resulted in including five components for the AEX series and 8 components for both the FTSE and SP500 series. The results are given below in Table 4. We note that the ordinary AR-models perform the worst. These models have the highest RMSE and MAE values out of all six alternatives. This result is the same across all three time series. The general HAR model performs better than the aforementioned models for

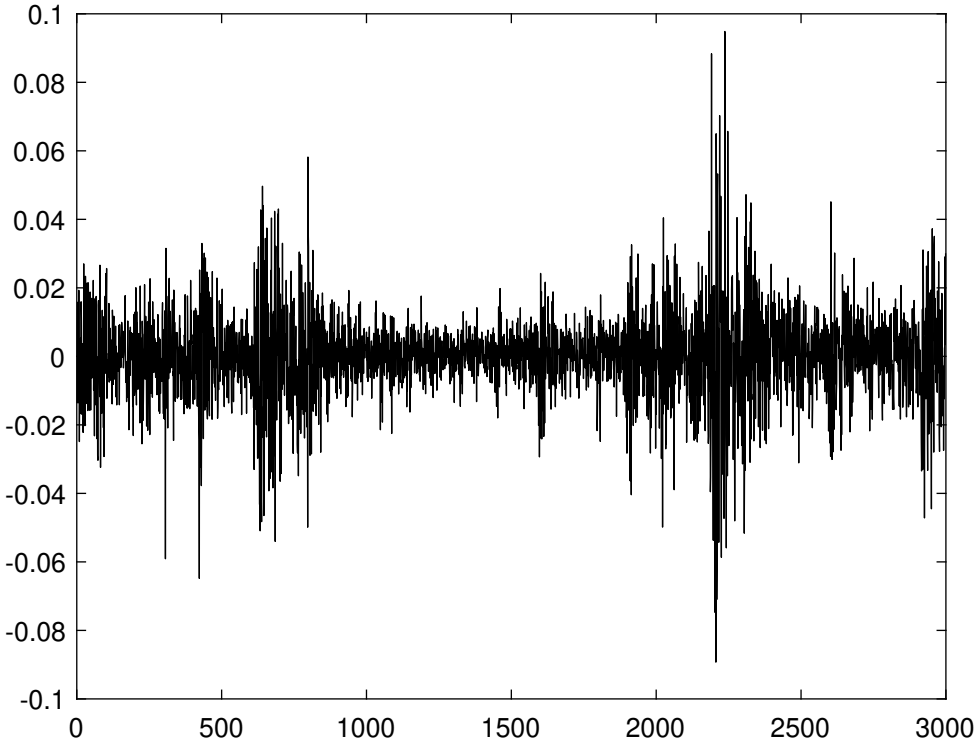

Figure 1. Daily return series FTSE

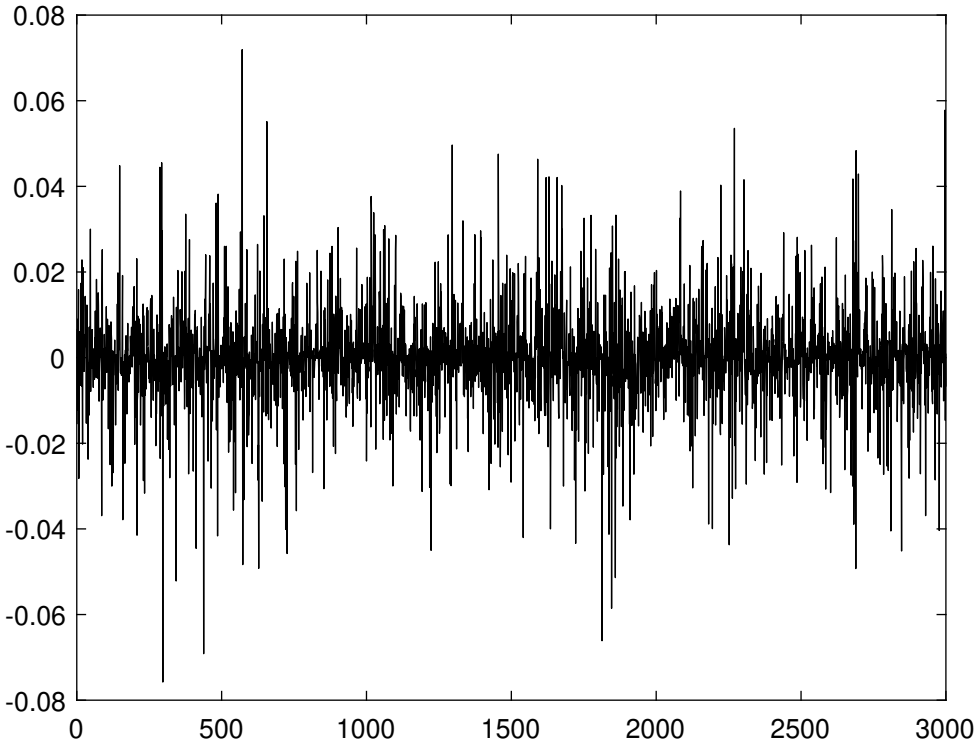

Figure 2. Daily return series simulation

all series. For instance, the RMSE of the HAR model for the AEX series is 4.219 compared to the value of 4.391 for the AR(3) model. All our three extensions outperform the general

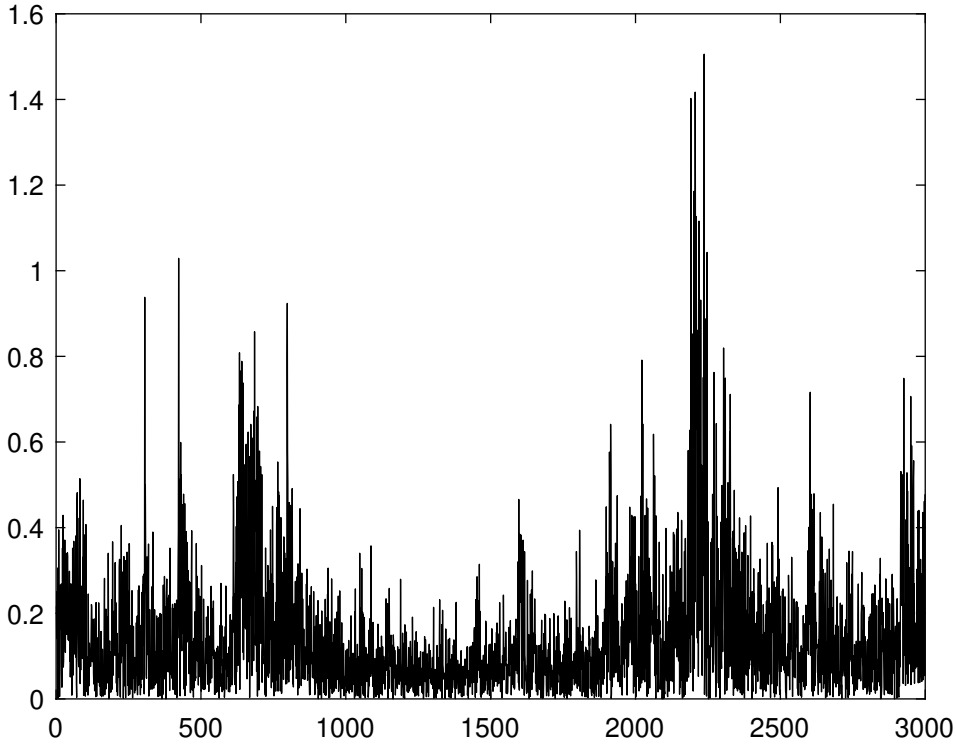

Figure 3. Daily realized volatility FTSE

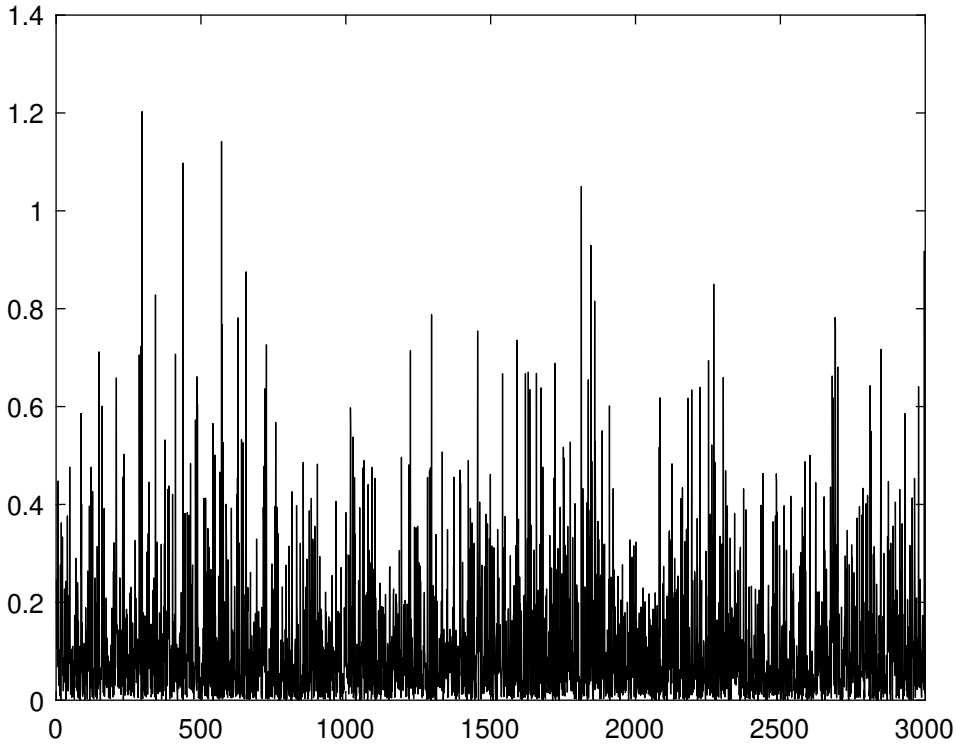

Figure 4. Daily realized volatility simulation

HAR-model in-sample. When looking at the RMSE metric, the HAR-PLR model performs the best out of the three extensions. This result holds across all series. When looking at the MAE

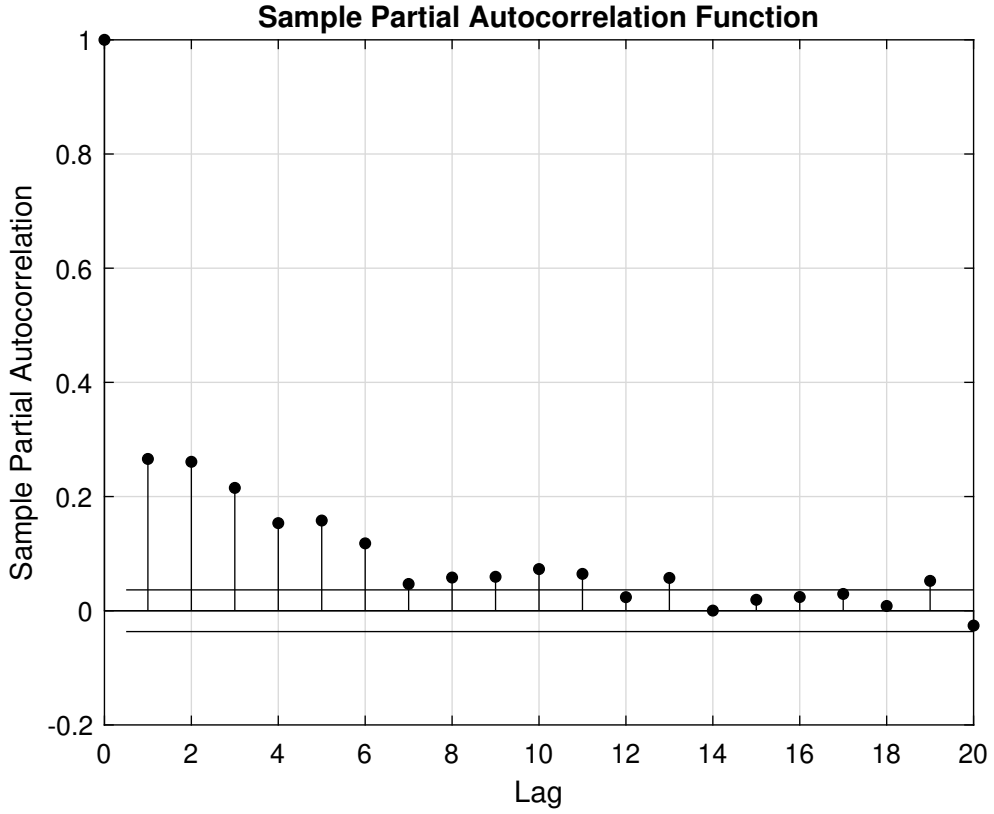

Figure 5. Partial autocorrelation FTSE

Table 4

*Comparison of the in-sample performances of the one-day-ahead forecasts of AR(1), AR(3), HAR(3), HAR-LASSO, HAR-PCA and HAR-PLR models for AEX, FTSE and S&P500 index data. The sample data ranges from January 2020 through May 2020. Performance measures are the root mean square error (RMSE), the mean absolute error (MAE), the R^2 of the Mincer–Zarnowitz regressions and the Diebold-Mariano test statistic as introduced by Diebold and Mariano [7]. Critical values of Diebold-Mariano test statistics is 1.96 when considering $\alpha = 0.05$. Hence, the null hypothesis of equal predictive accuracy is rejected when the D-M test statistic is larger than 1.96. A * indicates a model which has significantly better predictive accuracy than the general HAR(3) model.*

	AEX				FTSE				SP500			
	RMSE	MAE	R^2	D-M	RMSE	MAE	R^2	D-M	RMSE	MAE	R^2	D-M
AR(1)	4.810	3.058	0.83	6.604	5.762	3.286	0.76	9.806	4.775	2.832	0.82	6.604
AR(3)	4.391	2.696	0.85	2.691	5.042	2.713	0.81	4.773	4.363	2.554	0.84	2.691
HAR(3)	4.219	2.593	0.93		4.824	2.595	0.90		4.265	2.542	0.92	
HAR-LASSO	4.185	2.556	0.88	0.996	4.660	2.456	0.87	1.445	4.135	2.429	0.89	0.765
HAR-PCR	4.110	2.541	0.93	1.345	4.665	2.515	0.90	3.018*	4.106	2.477	0.92	1.446
HAR-PLR	4.091	2.534	0.93	2.335*	4.635	2.508	0.90	1.994*	4.090	2.468	0.92	1.710

metric to determine the best model, a different conclusion has to be made. According to this metric, the HAR-LASSO more accurately predicts the realized volatility than all other models for the FTSE index and SP500 index series. The HAR-PLR model produces the most accurate realized volatility forecasts according to both metrics. The R^2's of all models are similar and do not provide much additional information apart from the notion that the actual realized

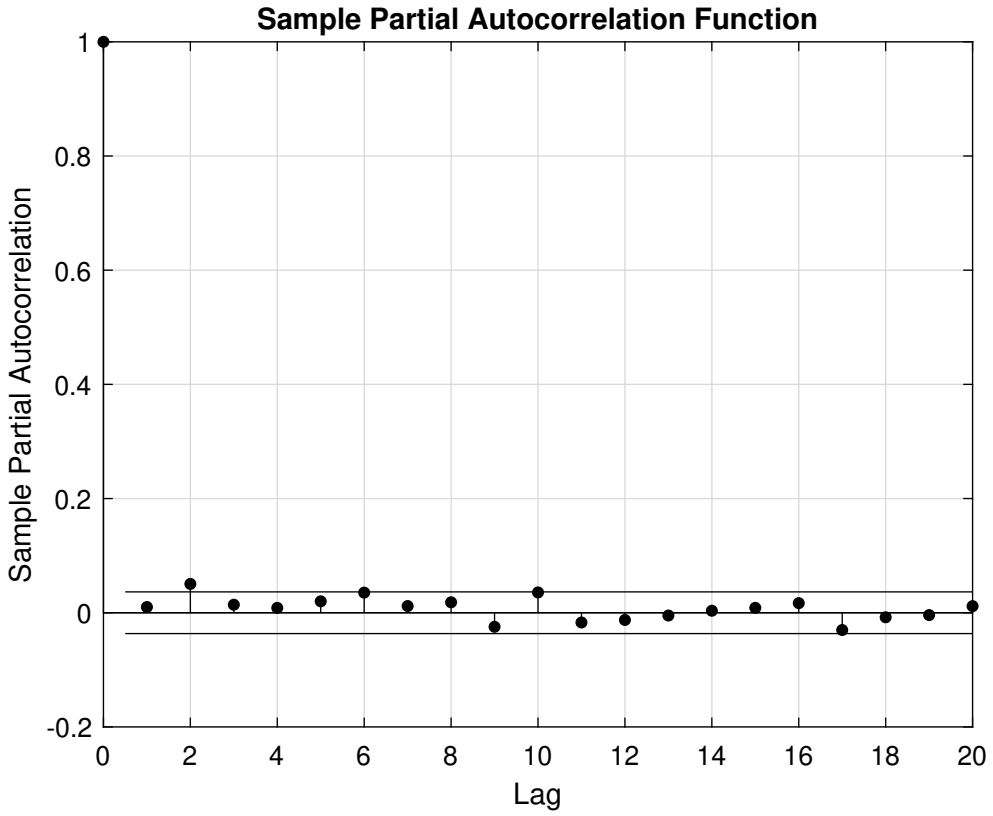

Figure 6. Partial autocorrelation simulation

volatility is explained by the forecasted realized volatility slightly worse when the forecasting window is increased. Now, we take a look at the Diebold-Mariano test statistics. We note that for the AEX daily realized volatility series, the HAR-PLR model performs significantly better than the HAR(3) model. This cannot be said for the SP500 series. Finally, the HAR-PCR and the HAR-PLR model perform significantly better than the HAR(3) model. We thus conclude that the HAR(3) model is beaten in-sample by the HAR-PCR and HAR-PLR models for one or more time series.

5.3 Out-of-sample

This section compares the out-of-sample performance of the models brought up by Corsi [5] and the proposed VAR-model. In order to test the out-of-sample performance, three time frames are looked at. Namely, one-day ahead, one week ahead and two weeks ahead. The results are given below in Table 5. We note some clear messages from Table 5. For the AEX series, the HAR(3) model performs best for the 1 day, 1 week and 2 weeks ahead forecasts. This is as expected, since the HAR(3) model also performed best in-sample. Furthermore, a gradual decline in forecasting performance can be noticed when extending the forecasting window. All three models suffer the same fate in this regard. The conclusions drawn for the AEX time series can be extended to

Table 5

*Comparison of the out-of-sample performances of the 1-day-, 1-week-, and 2-week-ahead forecasts of the AR(1), AR(3) and HAR(3) models for AEX index data, FTSE index data, and SP500 index data. The AR(1), AR(3), and HAR(3) are daily re-estimated on a moving window of 1000 observations. The sample data ranges from January 2020 through December 2019. Performance measures are the root mean square error (RMSE), the mean absolute error (MAE), the R^2 of the Mincer–Zarnowitz regressions and the Diebold-Mariano test statistic as introduced by Diebold and Mariano [7]. Critical values of Diebold-Mariano test statistics is 1.96 when considering $\alpha = 0.05$. Hence, the null hypothesis of equal predictive accuracy is rejected when the D-M test statistic is larger than 1.96. A * indicates a model which has significantly better predictive accuracy than the general HAR(3) model.*

	1 day				1 week				2 weeks			
	RMSE	MAE	R^2	D-M	RMSE	MAE	R^2	D-M	RMSE	MAE	R^2	D-M
					AEX							
AR(1)	4.118	2.640	0.98	5.833	6.005	4.183	0.96	6.012	7.218	5.172	0.91	4.769
AR(3)	3.918	2.448	0.98	2.751	5.237	3.386	0.98	3.691	6.160	4.125	0.96	3.062
HAR(3)	3.843	2.372	0.98		4.959	3.046	0.97		5.623	3.526	0.95	
VAR-LASSO $\lambda = 1$	3.972	2.480	0.97	2.309	4.904	3.135	0.95	0.831	5.550	3.582	0.92	0.505
VAR-LASSO $\lambda = 1.5$	3.882	2.438	0.98	1.465	4.950	3.145	0.95	0.142	5.661	3.618	0.93	0.306
VAR-LASSO $\lambda = 2$	3.896	2.444	0.98	1.915	5.025	3.184	0.95	0.926	5.747	3.684	0.93	0.915
					FTSE							
AR(1)	5.053	2.794	0.98	4.405	7.125	4.531	0.95	5.324	8.045	5.277	0.89	4.239
AR(3)	4.717	2.522	0.98	1.424	5.978	3.433	0.97	3.526	6.678	4.090	0.96	3.109
HAR(3)	4.588	2.406	0.98		5.605	2.987	0.96		6.027	3.349	0.95	
VAR-LASSO $\lambda = 1$	4.832	2.521	0.96	1.396	5.905	3.094	0.91	0.985	6.734	3.587	0.86	1.067
VAR-LASSO $\lambda = 1.5$	4.823	2.511	0.96	1.340	5.897	3.082	0.91	0.960	6.741	3.576	0.86	1.081
VAR-LASSO $\lambda = 2$	4.831	2.513	0.96	1.388	5.922	3.082	0.92	1.052	6.782	3.583	0.87	1.152
					SP500							
AR(1)	4.592	2.618	0.98	2.827	6.742	4.239	0.97	3.519	7.947	5.138	0.92	3.183
AR(3)	4.321	2.481	0.98	0.571	5.901	3.509	0.96	1.389	6.725	4.145	0.95	1.087
HAR(3)	4.289	2.436	0.97		5.706	3.247	0.94		6.441	3.692	0.92	
VAR-LASSO $\lambda = 1$	4.622	2.545	0.96	2.385	6.359	3.530	0.90	2.595	7.589	4.196	0.85	1.969
VAR-LASSO $\lambda = 1.5$	4.628	2.543	0.96	2.426	6.340	3.503	0.91	2.444	7.536	4.118	0.85	1.862
VAR-LASSO $\lambda = 2$	4.621	2.541	0.96	2.379	6.364	3.510	0.91	2.471	7.566	4.114	0.85	1.870

the FTSE and SP500 realized volatility time series. There is no significant difference in result for these time series. We see that, in terms of RMSE, the HAR(3) model performs best when forecasting the AEX realized volatility time series one day ahead. The worst performance can be attributed to the AR(1) model when forecasting the FTSE realized volatility time series two weeks ahead with a RMSE of 8.045. The Diebold-Mariano test statistics confirm that none of our proposed models are able to beat the HAR(3) out-of-sample. We thus conclude that our models are no improvement over the HAR(3) out-of-sample.

6 Conclusion

In this paper we sought to answer the following research question: 'To what extent do economic variables increase the predictive power of the HAR-RV model?'. We answered this question through adding economic variables as described in Welch and Goyal [25] to the HAR-RV model.

Before analysing the in-sample and out-of-sample performance of our models, we simulated

the general HAR-RV model to attempt show its dynamics. However, the simulation did not accomplish to reflect the stylized facts present in the FTSE index time series. This can be arguably attributed to the fact that our simulation used daily data instead of tick data.

To analyse the in-sample performance of our models we performed one-day ahead forecasts for the AEX index, FTSE index and the SP500 index series. We find that the HAR-PCA and HAR-PLR model significantly beat the HAR(3) model as proposed by Corsi [5] in terms of predictive power. According to the RMSE metric, the HAR-PLR is the undistinguished best model to perform in-sample forecasts. The HAR-LASSO model is the best performer when looking at the MAE for the FTSE and SP500 series. This is an interesting takeaway from the analysis. This anomaly indicates the HAR-PLR model does not predict values which are as far off from the actual value as the HAR-LASSO model. On the other hand, the HAR-LASSO model performs best when errors with a large magnitude are not punished harder. Hence, we conclude that adding economic variables to the general HAR(3) model as proposed by Corsi [5] significantly improved in-sample forecasting performance. This can be due to the fact that these economic variables hold valuable information on volatility with which the HAR model can be extended. We conclude that realized volatility is dependent on external economic metrics to a certain extent. These equity ratios and yield variables impact the daily realized volatility of our tested indices. We can thus say that these factors cannot be overlooked when trying to optimize in-sample realized volatility forecasting.

Next, we looked at out-of-sample forecasting performance. We came up with the following models: HAR-LASSO, HAR-PCR and HAR-PLR. We performed one day, one week and two weeks ahead forecasts. The three aforementioned models arguably have the same forecasting accuracy as the AR(3) model. There are occasions where the AR(3) model gets beaten by these models. Yet, this also occurs the other way around. We conclude that the three extension models cannot beat the general HAR(3) model out-of-sample. However, according to the Diebold-Mariano test statistics, they occasionally share the same forecasting power as the HAR(3) model. Hence, the addition of economic variables did not contribute to a better out-of-sample forecasting performance in the current setting.

Although we found some promising results, there are some limitations to our research. Firstly, the use of monthly data on economic variables hampers the predicting and forecasting ability of our models. Clearly, there is a loss of potentially useful information. Considering this, it is already surprising that the ordinary HAR-RV is beaten in-sample. A second point is the fact that three time series are arguably not a valid representation for the whole market. It could well be that our extensions fail to outperform the general HAR-RV model in-sample

for other time series. A more comprehensive analysis would thus be a sound extension on this work, since the used indices are based in Western developed economies. It would be interesting to gauge the performance of our models in developing economies, since their financial markets are generally less efficient. A third limitation is the handling of the LASSO-VAR model. Due to computational time, we restricted the lag order to 5. However, it is a reasonable assumption that a higher lag order would lead to better forecasts. Looking at the partial autocorrelation stuctures of the time series strengthens this argument. Furthermore, we did not test whether the economic variables time series are stationary. Therefore, a future extension could be the introduction of a cointegration analysis which could lead to the discovery of spurious relations. Introducing Vector Error Correction Models (VECM) initially introduced by Engle and Granger [9] could resolve this issue.

References

[1] Torben G Andersen et al. "Modeling and forecasting realized volatility". In: *Econometrica* 71.2 (2003), pp. 579–625.

[2] Jonathan Baskin. "Dividend policy and the volatility of common stocks". In: *Journal of portfolio Management* 15.3 (1989), p. 19.

[3] Fischer Black and Myron Scholes. "The pricing of options and corporate liabilities". In: *Journal of political economy* 81.3 (1973), pp. 637–654.

[4] Charlotte Christiansen, Maik Schmeling, and Andreas Schrimpf. "A comprehensive look at financial volatility prediction by economic variables". In: *Journal of Applied Econometrics* 27.6 (2012), pp. 956–977.

[5] Fulvio Corsi. "A simple approximate long-memory model of realized volatility". In: *Journal of Financial Econometrics* 7.2 (2009), pp. 174–196.

[6] Fulvio Corsi and Roberto Reno. "HAR volatility modelling with heterogeneous leverage and jumps". In: *Available at SSRN 1316953* (2009).

[7] Francis X Diebold and Robert S Mariano. "Comparing predictive accuracy". In: *Journal of Business & economic statistics* 20.1 (2002), pp. 134–144.

[8] Robert F Engle. "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the Econometric Society* (1982), pp. 987–1007.

[9] Robert F Engle and Clive WJ Granger. "Co-integration and error correction: representation, estimation, and testing". In: *Econometrica: journal of the Econometric Society* (1987), pp. 251–276.

[10] Robert W Gerlach, Bruce R Kowalski, and Herman OA Wold. *Partial least squares path modelling with latent variables.* Tech. rep. WASHINGTON UNIV SEATTLE LAB FOR CHEMOMETRICS, 1979.

[11] Shihao Gu, Bryan Kelly, and Dacheng Xiu. "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5 (2020), pp. 2223–2273.

[12] John Lintner. *The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, in ≪Review of Economics and Statistics≫, rist.* 1965.

[13] Feng Ma, MIM Wahab, and Yaojie Zhang. "Forecasting the US stock volatility: An aligned jump index from G7 stock markets". In: *Pacific-Basin Finance Journal* 54 (2019), pp. 132–146.

[14] Hans Malmsten, Timo Teräsvirta, et al. “Stylized facts of financial time series and three popular models of volatility”. In: *SSE/EFI Working Paper Series in Economics and Finance* 563 (2004), pp. 1–44.

[15] H Markowitz. “Portfolio selection, 1952 J”. In: *Financ* 7 (), p. 77.

[16] Antonio Mele. “Asymmetric stock market volatility and the cyclical behavior of expected returns”. In: *Journal of financial economics* 86.2 (2007), pp. 446–478.

[17] Jacob A Mincer and Victor Zarnowitz. “The evaluation of economic forecasts”. In: *Economic forecasts and expectations: Analysis of forecasting behavior and performance*. NBER, 1969, pp. 3–46.

[18] Ulrich A Müller et al. “Fractals and intrinsic time: A challenge to econometricians”. In: *Unpublished manuscript, Olsen & Associates, Zürich* (1993), p. 130.

[19] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[20] David E Rapach, Jack K Strauss, and Guofu Zhou. “Out-of-sample equity premium prediction: Combination forecasts and links to the real economy”. In: *The Review of Financial Studies* 23.2 (2010), pp. 821–862.

[21] William F Sharpe. “Capital asset prices: A theory of market equilibrium under conditions of risk”. In: *The journal of finance* 19.3 (1964), pp. 425–442.

[22] William F Sharpe. “Mutual fund performance”. In: *The Journal of business* 39.1 (1966), pp. 119–138.

[23] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[24] Enriqueta Vercher, José D Bermúdez, and José Vicente Segura. “Fuzzy portfolio optimization under downside risk measures”. In: *Fuzzy sets and systems* 158.7 (2007), pp. 769–782.

[25] Ivo Welch and Amit Goyal. “A comprehensive look at the empirical performance of equity premium prediction”. In: *The Review of Financial Studies* 21.4 (2008), pp. 1455–1508.

[26] Lan Zhang, Per A Mykland, and Yacine Aıt-Sahalia. “A tale of two time scales: Determining integrated volatility with noisy high-frequency data”. In: *Journal of the American Statistical Association* 100.472 (2005), pp. 1394–1411.

7 Appendix

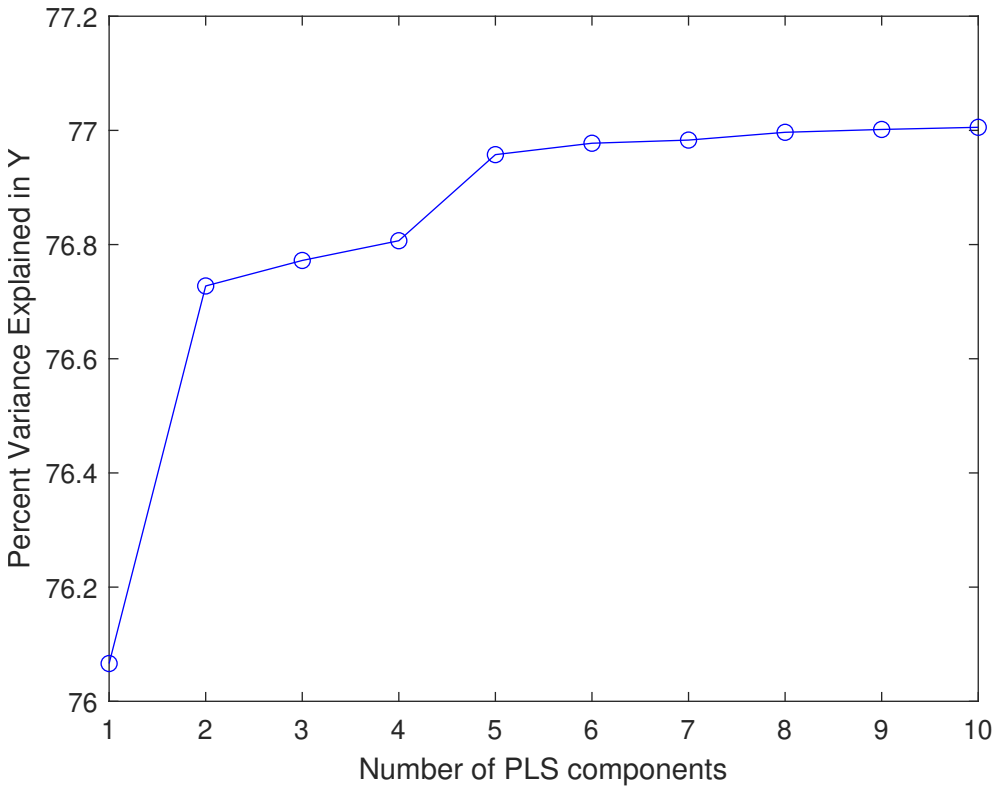

Figure 7. Principal components AEX daily realized volatility

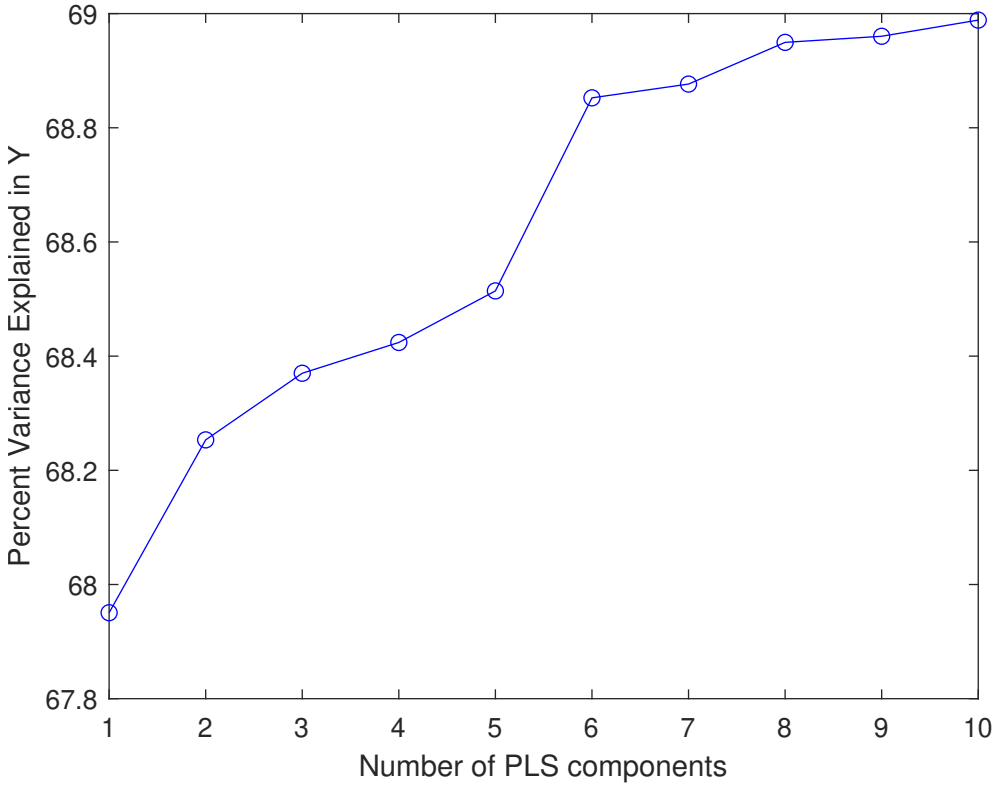

Figure 8. Principal components FTSE daily realized volatility

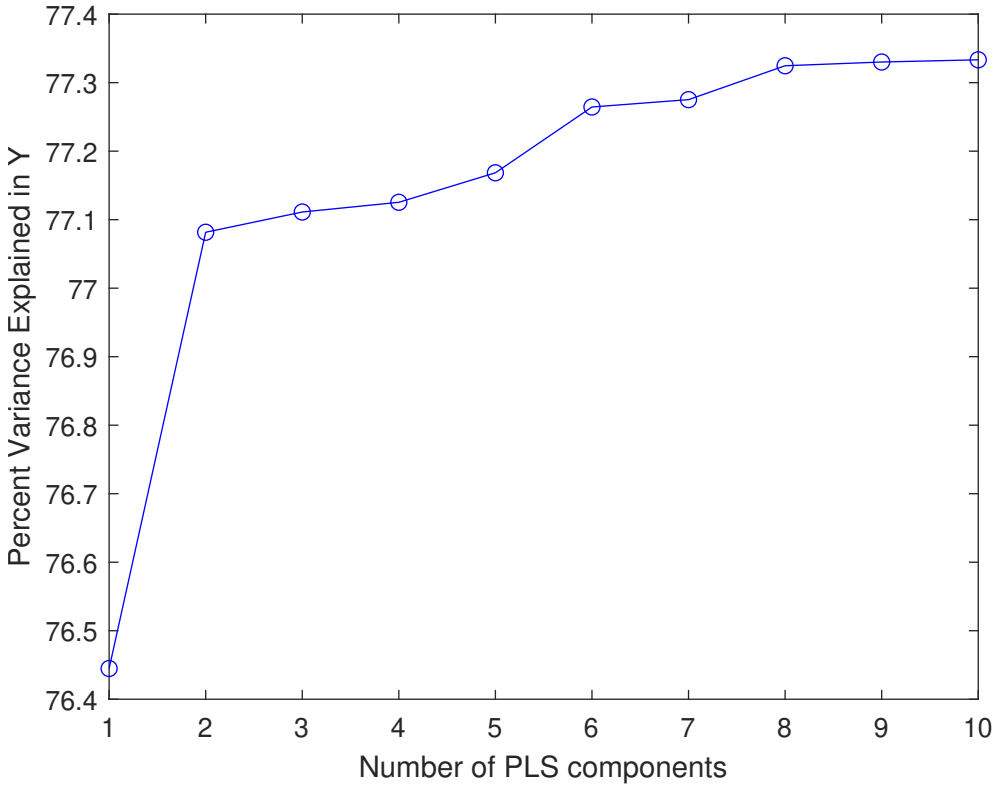

Figure 9. Principal components SP500 daily realized volatility

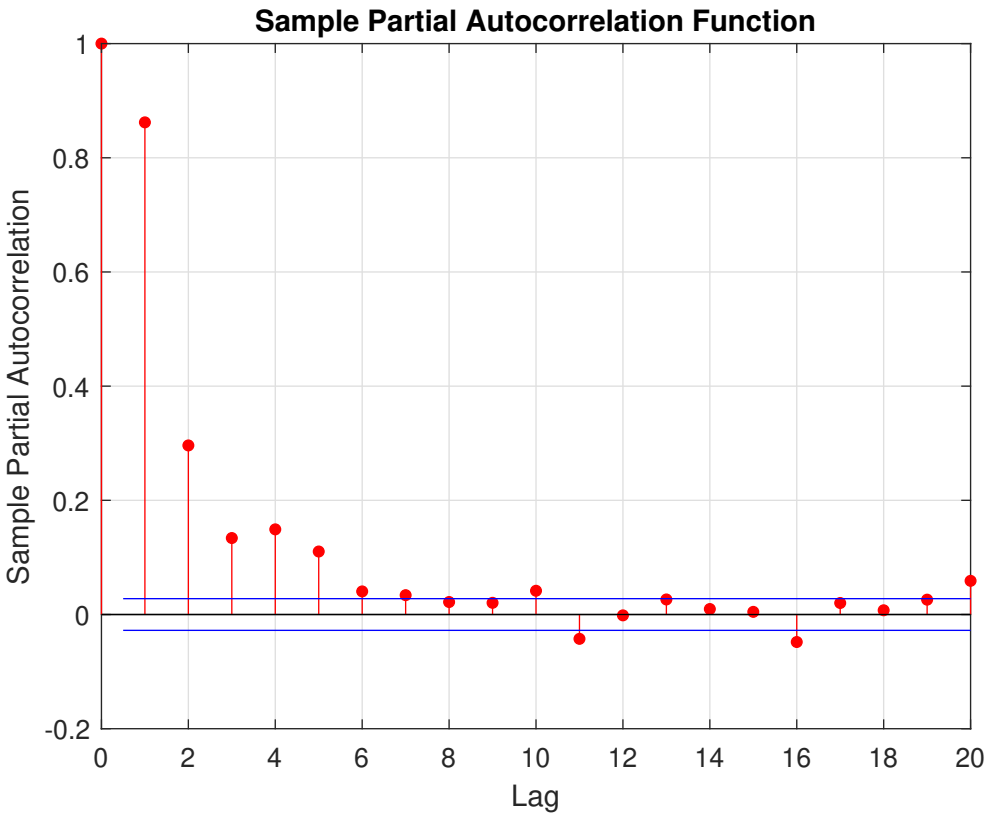

Figure 10. Partial autocorrelation AEX

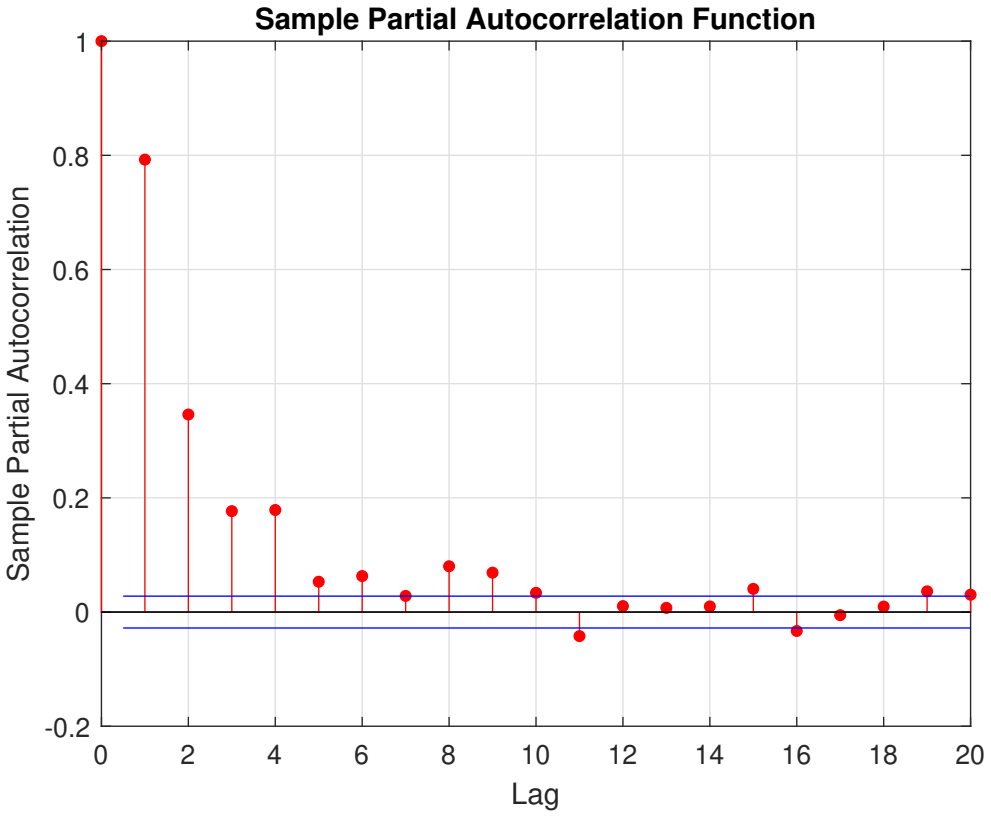

Figure 11. Partial autocorrelation FTSE

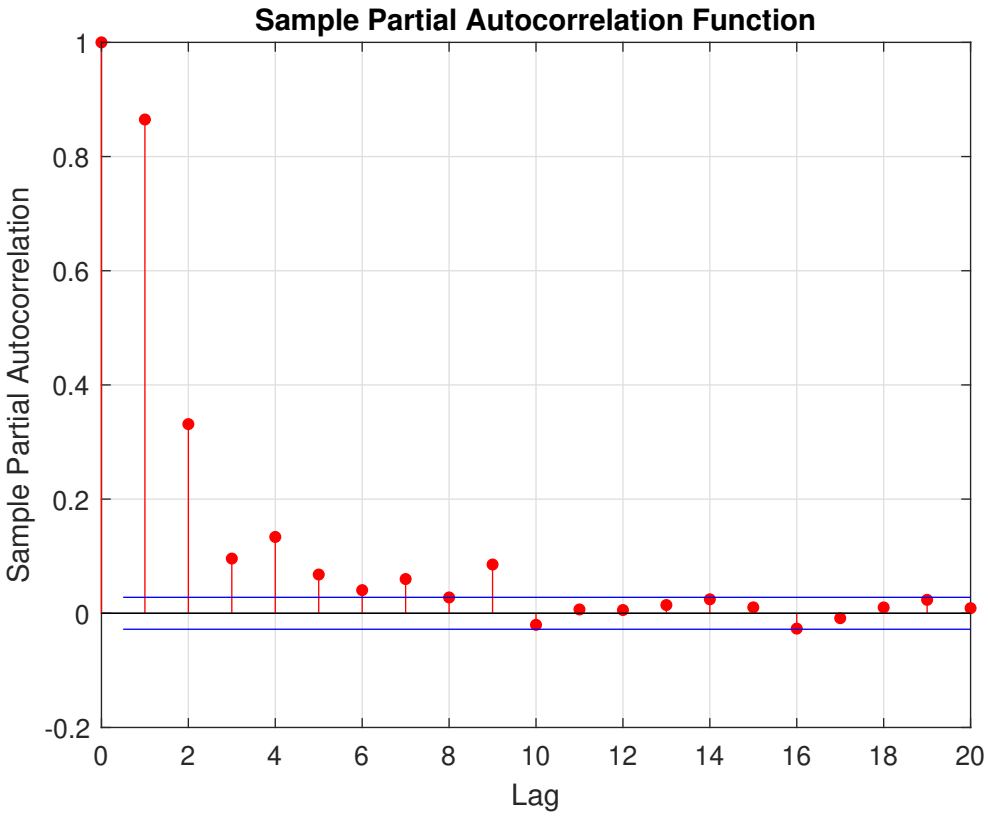

Figure 12. Partial autocorrelation SP500