Prediction of volatility with penalized mixture distributions

Master Thesis Submitted to

Prof. Dr. Ostap Okhrin

Prof. Dr. Wolfgang K. Härdle

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E.- Centre for Applied Statistics and Economics

Humboldt-Universität zu Berlin

by

Nina Grygorenko

(553382)

in partial fulfillment of the requirements

for the degree of

Master of Science

Berlin, August 19, 2015

Acknowledgement

I would like to thank Prof. Dr. Ostap Okhrin for giving me the opportunity to do research under his guidance as well as for his support during the onerous time of writing this thesis. I would also like to express many thanks to Prof. Dr. Wolfgang Härdle for his encouragement over the entire course of my studies at Humboldt University and for the knowledge that I gained from his courses. I am indebted to all members of the chair of statistics for the knowledge I was enriched with during my master's program.

Nobody, but my parents, could provide support and love at the level I have received during the wonderful adventure of my education – for that I will remain eternally grateful to them.

Abstract

The research presented in this dissertation deals with robustness of two models describing daily realized volatility of stock market indexes, the HAR model introduced by Corsi (2009) and Lasso introduced by Tibshirani (1996). These models have been selected because of their ability to recover long memory dependence in the data. Comparison of forecast accuracy of these two models has been performed following research provided by Audrino and Knaus (2012). The results for stock market indexes obtained in this research coincide with the results reported by Corsi (2009) and Audrino and Knaus (2012).

A new type of estimator, the MLE based on penalized mixture distribution was proposed and investigated for the two models explored in this thesis. This estimator can replicate non-normal distribution of data and select appropriate mixture distribution function, which fits the analyzed data well. However, according to the empirical results described in this thesis, the new estimator doesn't improve forecast accuracy of both analyzed models for stock market indexes.

Key words: Realized volatility, long memory dependence, non-normal distribution, HAR model, Lasso, penalized mixture distribution

Contents

List of Abbreviations

ACF	autocorrelation function
AR	autoregressive
ARCH	autoregressive conditional eteroskedasticity
ARFIMA	autoregressive fractionally integrated moving average
GARCH	generalized autoregressive conditional heteroskedasticity
HAR	heterogeneous autoregressive
Lasso	least absolute shrinkage and selection operator
MLE	maximum likelihood estimator
MSE	mean squared error
OLS	ordinal least square
PDF	probability density function
TTSE	two time scale estimator

List of Figures

List of Tables

1 Introduction

Stock market and prediction of stock markets movements is a topic investigated by many scientists during the last several decades. A wide variety of models and approaches designed to estimate, analyze and forecast stock prices have been developed. However, it was also clear that stock prices are stochastic in nature and, consequently, they are very difficult to predict. Stock price returns are less stochastic and, thus, they are easier to forecast. As a result the models concentrate currently on estimation, analyses and prediction of stock returns. One of the most important components of stock returns models is volatility of the returns. Various approaches and models were introduced in order to estimate and forecast the volatility of stock prices returns.

The first approach to the volatility estimation was based on the assumption that conditional volatility is a latent observation, and therefore it couldn't be observed and estimated in a direct way. With that realization in mind a wide variety of conditional autoregressive models, such as ARCH proposed by (Engle, 1982), GARCH proposed by Bollerslev (1986), Exponential GARCH (EGARCH) proposed by Nelson (1991), Threshold GARCH (TGARCH) introduced by Zakoian (1994), Fractionally Integrated GARCH (FIGARCH) introduced by Baillie et al. (1996) and others, have been developed. Other approaches to estimation of conditional volatility are stochastic volatility models, first introduced by Taylor (1986). Also, in order to estimate the latent volatility Exponentially Weighted Moving Average (EWMA) model was introduced; it was advocated by Riskmetrics methodology proposed by Morgan (1996). The common problem of the existing models, designed to estimate latent volatility was that all of them failed to replicate some stylized facts, which is observed in financial time series data. Moreover, for some of these models (for example stochastic volatility model) the estimation procedure is complicated.

The search for appropriate model capable to estimate volatility leads to analysis of high frequency data. Analysis of high frequency data provides an opportunity to estimate volatility as an observed variable. The realized volatility was conceived as volatility measure and was first calculated by Andersen and Bollerslev (1998). This measure developed rapidly and a wide variety of approaches to calculation of realized volatility have been introduced.

With the advancements in modeling a necessity to create conditional volatility model based on realized volatility, which would be able to replicate all empirical properties and at the same time would be easier in estimation and interpretation, was increasingly clearly perceived. The

HAR model proposed by Corsi (2009) constituted a significant progress towards satisfaction of the criteria described above. The main empirical property of realized volatility data covered by the HAR model is long-memory dependence. In this model the daily realized volatility today, aggregated weekly realized volatility, and monthly realized volatility were selected as explanatory variables of the daily realized volatility tomorrow. Estimation of the HAR model could be provided by the simple OLS estimator. The HAR model achieved high popularity due to its good statistical properties and the parsimony of the model.

One of the most effective and widely applied approaches to solve the problem of model selection is Lasso. Nardi and Rinaldo (2011) applied Lasso to autoregressive process in order to solve the problem of appropriate model order selection in time series data. Lasso could also provide selection of lags, which should be included in the model as explanatory variables. Therefore, Lasso could be used to check correctness of explanatory variable selection in the HAR.

Evidence that Lasso should replicate HAR model asymptotically, in the case when HAR model is a true data generation process, was provided by Audrino and Knaus (2012). However, based on the empirical estimations in research reported also by Audrino and Knaus (2012), it was shown that Lasso does not recover HAR model. Therefore, this provides evidence against the assumption, that the HAR model represents a true data generation process.

There is one more problem present in the existing realized volatility estimations. Empirical research clearly shows that the realized volatility follows a non-normal distribution. Features such as high kurtosis and fat tails could be observed in the realized volatility data and they are not accurately captured by the existing estimations. Consequently, appropriate selection of distribution could increase estimation accuracy of MLE estimator and, hence, could lead to an increased quality of forecast. Therefore, an appropriate selection of distribution remains a very actual problem.

Motivated by the perceived shortcomings in the existing results, which were exposed by the application of Lasso in model selection, an idea to apply this operator to probability weights in mixture distribution will be introduced in this research. In such an approach different distributions could be included in the mixture distribution probability density function, and the weights associated with the distributions not relevant to the analyzed data will be shrunk to zero by Lasso. As a result selection of an appropriate distributions to analyzed data will be provided.

The theoretical part of this research will be divided into four sections. In the first section a short description of different approaches to calculation of realized volatility based on intraday high frequency data is provided. In the second section a general description of HAR model and its extensions is given. The third part contains general information about Lasso and and its application to autoregressive process. In the fourth section an estimator based on penalized mixture distributions is introduced.

The empirical part of this research is divided into three sections. In the first section robustness of HAR model and Lasso is checked using daily realized volatility of stock exchange indexes. In that section the evidence that Lasso doesn't recover HAR model provided by Audrino and Knaus (2012) is also examined on the daily realized volatility of stock exchange indexes. In the second section the results of estimation HAR model and Lasso using MLE based on penalized mixture distribution is presented. In the third section, comparison of the forecast accuracy of HAR model and Lasso estimated using two different estimators, standard and MLE based on penalized mixture distributions, is provided.

2 Realized volatility measures

Volatility is a statistical measure, which represents standard deviation from the mean value of returns for assets, stock market indexes and other financial instruments. This measure began its rapid development and its dominance among other statistical measures with development of financial markets. The main topic of the continuing research in financial time series is oriented towards making their forecast more accurate. Based on the fact, that daily volatility of financial instruments, especially of stocks, is almost unpredictable, scientists concentrated their effort on the forecast volatility of returns of financial instruments, which is easier to predict.

This research is focused on a volatility measure called realized volatility. The realized volatility measure is a consistent nonparametric estimation of price movements over certain time interval.

The realized volatility measure plays an important role in practical estimation and forecasting of volatility. It is possible to forecast realized volatility employing the classical time series techniques using daily realized volatility observations calculated based on intraday high frequency data. Also, the realized volatility measure could be used as measure of realization of latent volatility.

The intuition of realized volatility could be presented using stochastic volatility (diffusion) process:

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) \tag{1}$$

where $p(t)$ is logarithm of instantaneous price, $\mu(t)dt$ is continuous finite variation process, $dW(t)$ is a standard Brownian motion and $\sigma(t)$ is a stochastic process independent of $dW(t)$.

In the process of rapid development of the realized volatility measure many different estimators for this measure were created. Firstly, simple time discrete model with no microstructure noise was introduced to estimate realized volatility. In this model the realized variance defined as sum of observed intraday high frequency squared returns was introduced as a realized volatility measure.

Then a continuous time models also with no microstructure noise was introduced. This model was based on different sampling schemes, such as calendar time sampling, transaction time sampling, business time sampling and tick time sampling.

After these models had been introduced, the question about presence of microstructure noise and bias, which is caused by microstructure noise in estimation of the realized volatility, were widely discussed. It was concluded that all estimators created without taking into account microstructure noise are biased. An example of such estimator is the realized variance using different sampling.

Three main assumptions about possible structure of microstructure noise were described first by Bandi and Russell (2005) and summary was provided by McAleer and Medeiros (2008). The first assumption is about noise structure. This means that the microstructure noise has mean, which is equal to zero and is a covariance stationary stochastic process. The second assumption is that microstructure noise has independent and identically distributed noise structure. According to this assumption the microstructure noise has mean, which is equal to zero and is random variable which is an independent and identically distributed. Under this assumption noise is process which is independent of the price. The third assumption is that the microstructure noise has dependent noise structure. The microstructure noise has mean which is equal to zero as well as stationary and strong mixing stochastic process, with the mixing coefficients decaying exponentially.

There are many estimators introduced by different scientists which are unbiased, however not all of those estimators are consistent. There are two main estimators of the realized volatility measures presented in the literature which are robust with respect to noise and consistent, and those are: the two time scale estimator (TTSE) introduced by Zhang et al. (2005) and by Ait-Sahalia et al. (2006), and the realized kernel by Barndorff-Nielsen et al. (2006).

TTSE presented Zhang et al. (2005) is described using the following formula:

$$RV_t^{(ZMA)} = \frac{1}{K}\sum_{k=1}^{K} RV_t^{(K)} + \frac{\bar{n}_t}{n_t} RV_t^{(all)}, \text{ where} \tag{2}$$

$$RV_t^{(all)} = \sum_{i=1}^{n_t} (r_{t,i}^*)^2 + \sum_{i=1}^{n_t} r_{t,i}^* \nu_{t,i} + \sum_{i=1}^{n_t} \nu_{t,i};$$

$$RV_t^{(k)} = \sum_{i=1}^{n_t^k} r_{t,i}^2;$$

$$\bar{n}_t = \frac{1}{K}\sum_{k=1}^{K} n_t^{(k)} = \frac{n_t - K + 1}{K};$$

where $n_t^{(k)}$ is the number of observations in each subgrid, n_t is the number of observations in the full grid, K is the number of subgrids, $r_{(t,i)}^* = p_{(t,i)}^* - p_{(t,i-1)}^*, p_{(t,i-1)}^*$ is the latent efficient price process, $r_{(t,i)} = r^*_{(t,i)} + \epsilon_{(t,i)} - \epsilon_{(t,i-1)} = r^*_{(t,i)} + v_{(t,i)}$. The price model in this estimator is a diffusion process, microstructure noise is independent and identically distributed. The estimator provided by Ait-Sahalia et al. (2006) includes dependent microstructure noise process. Microstructure noise and efficient price are independent.

The kernel estimator proposed by Barndorff-Nielsen et al. (2006) is described by

$$RV_t^{(BHLS)} = RV_t^{(all)} + \sum_{h=1}^{H} k\frac{h-1}{H}(\widehat{\gamma}_h + \widehat{\gamma}_{-h}), \text{ where} \tag{3}$$

$$\widehat{\gamma}_h = \frac{n_t}{n_t - h}\sum_{j=1}^{n_t} r_{t,j} r_{j+h}$$

where $k(x)$ is the kernel weight function. In Barndorff-Nielsen et al. (2006) three different kernel functions were compared: Barlett kernel, 2nd order kernel and Epanechnikov kernel. H in the above formula is the bandwidth, which is selected according to Barndorff-Nielsen et al. (2006) as $H = cn_t^{(2/3)}$, where n_t is the number of observations, and c is a constant, which can be optimally chosen as a function of the kernel $k(x)$.

The realized kernel estimator is sensitive to choice of bandwidth. The price model of this estimator is also diffusion process. No assumptions about microstructure noise was made in this realized variance estimator.

Improvement of this estimator was provided by Barndorff-Nielsen et al. (2009). It was proposed to use Parzen kernel as kernel weight function. This kernel weight function satisfied smoothness condition $k^{'}(0) = k^{'}(1) = 0$ and, at the same time, always produced a non-negative estimate, which is important because the realized volatility could be only positive. The Parzen kernel function is described by

$$k(x) = \begin{cases} 1 - 6x^2 - 6x^3 & 0 \leq x \leq \frac{1}{2}, \\ 2(1-x)^3 & \frac{1}{2} < x \leq 1 \\ 0 & x > 1. \end{cases} \tag{4}$$

Preferred bandwidth, according to Barndorff-Nielsen et al. (2009), is

$$H^* = c^* \xi^{(4/5)} n^{(3/5)} \tag{5}$$

with $c^* = \{\frac{k"(0)^2}{k^{0.0}_{\cdot}}\}^{\frac{1}{5}}$ which is equal to $c^* = 3.5134$ for Parzen kernel, and $\xi^2 = \frac{\omega^2}{\sqrt{T \int_0^T \sigma_u^4 du}}$ with unknown quantiles ω^2 and $\int_0^T \sigma_u^4 du$,which is called integrated quarticity.

3 Methodology

It is first noted that in the realized volatility data features such as persistence of long memory dependence and non-normal distribution are present. In this context selection of an appropriate model becomes one of the most important components in the forecast quality. In this section two models, which could fit realized volatility well, are described. The first model is Heterogeneous Autoregressive (HAR) model introduced by Corsi (2009), which tries to replicate long memory dependence of the realized volatility. The second model is the least absolute shrinkage and selection operator (Lasso) introduced by Tibshirani (1996), which is one of the most popular and efficient estimator applied in order to solve model/explanatory variables selection problems. In this section an estimator for HAR model and for Lasso is also described. The estimator is based on penalization of mixture distributions using Lasso and therefore designed to replicate non-normal distribution of the daily realized volatility.

3.1 HAR model

3.1.1 General description of HAR model

HAR model was introduced by Corsi (2009). The main idea of this model was to find a model, which will cover persistence of long memory dependence, which is clearly observed in the realized volatility data but at the same time a model, which will be easy in estimation and interpretation.

The idea of HAR model was based on Heterogeneous Market Hypothesis introduced by Mueller et al. (1993). Heterogeneous Market Hypothesis could be explained using the empirical findings presented by Mueller et al. (1993). The first finding is that different participants of the heterogeneous market include different time horizons and different dealing frequencies. The second finding is that increase of market participation leads to growth of market volatility, while in homogeneous market hypotheses increase of market participants leads to improvement of convergence. The third finding is that different geographical location of market participants produce heterogeneity in the market. Summarizing the findings described above, it could be concluded that heterogeneity of the traders and positive correlation between the number of market participants and volatility was recognized by Heterogeneous Market Hypothesis. This hypothesis was reflected in the Heterogeneous Autoregression Conditional Heteroscedasticity (HARCH) model created by Mueller et al. (1997). It was shown by Mueller et al. (1997)

that HARCH model produced better results in comparison with Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model.

Also important for HAR model creation was the evidence about asymmetric propagation of volatility provided by Mueller et al. (1997) and by Arneodo et al. (1998). Asymmetric propagation of volatility, which is based on heterogeneous market hypothesis means that volatility over longer time interval has stronger influence on realized volatility tomorrow than volatility over shorter time interval. Inclusion of the asymmetric propagation in the model could replicate such stylized facts of time series data as long memory dependence and fat tail distribution.

HAR model uses partial volatilities as explanatory variables of the daily realized volatility, which are generated by the action of certain market component. According to this, HAR model could be characterized as additive model, which consists of different partial volatilities with hierarchical structure. There are three partial volatility components, which are included in this model. These partial volatilities are realized volatility aggregated over different time horizons: time horizon of one day (short-term), time horizon of one week (medium-term, aggregated five daily realized volatility values) and time horizon of one month (long-term, aggregate 22 daily realized volatility values). Partial volatilities create a volatility cascade of heterogeneous volatility components.

Daily, weekly and monthly partial realized volatilities could be described by the following formulas:

$$\widetilde{\sigma}^{(d)}_{t+1d} = c^{(d)} + \phi^{(d)} RV^{(d)}_t + \gamma^{(d)} E_t[\widetilde{\sigma}^{(w)}_{t+1w}] + \widetilde{\omega}^{(d)}_{t+1d}, \tag{6}$$

$$\widetilde{\sigma}^{(w)}_{t+1w} = c^{(w)} + \phi^{(d)} RV^{(w)}_t + \gamma^{(w)} E_t[\widetilde{\sigma}^{(m)}_{t+1m}] + \widetilde{\omega}^{(w)}_{t+1w}, \tag{7}$$

$$\widetilde{\sigma}^{(m)}_{t+1m} = c^{(m)} + \phi^{(m)} RV^{(m)}_t + \widetilde{\omega}^{(m)}_{t+1m} \tag{8}$$

where $RV^{(d)}_t$, $RV^{(w)}_t$, and $RV^{(m)}_t$ describe daily realized volatility, weekly realized volatility and monthly realized volatility respectively, $\phi^{(d)}$, $\phi^{(w)}$, $\phi^{(m)}$, $\gamma^{(w)}$, $\gamma^{(m)}$ are coefficients near explanatory variables, $c^{(d)}$, $c^{(w)}$, $c^{(m)}$ are the intercepts of the model, and where $\widetilde{\omega}^{(d)}_{t+1d}$, $\widetilde{\omega}^{(w)}_{t+1w}$, $\widetilde{\omega}^{(w)}_{t+1w}$ represent daily volatility innovations, weekly volatility innovations, and monthly volatility innovations respectively. Volatility innovations are contemporaneously and serially independent variations with mean equal to zero and truncated left tail in order to ensure positive values of partial volatilities according to Corsi (2009). It could be clearly observed in the formulas presented above, that partial volatility consists of two parts. One part is the autoregressive process of order one (AR(1)) and the other part is the expectation of partial volatility for the next longer time horizon.

After straightforward recursive substitution of partial volatility equation is performed, the following equation is obtained:

$$\sigma_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \widetilde{\omega}_{t+1d}^{(d)} \tag{9}$$

This equation could be characterized as stochastic volatility model, which include three components. These components are the past realized volatilities observed over different time intervals. From this equation the time series model of realized volatility was derived, which was provided by Corsi (2009).

The left-hand side of the above equation, $\sigma_{t+1d}^{(d)}$ could also be represented by the following formula:

$$\sigma_{t+1d}^{(d)} = RV_{t+1d}^{(d)} + \omega_{t+1d}^{(d)} \tag{10}$$

where $\omega_{t+1d}^{(d)}$ describes daily realized volatility, which is latent and at the same time $\omega_{t+1d}^{(d)}$ describes the estimation errors.

From the last two equations a simple cascade time series model for daily realized volatility was derived. This model could be described by the formula:

$$RV_{t+1d}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \omega_{t+1d} \tag{11}$$

where $\omega_{t+1d} = \widetilde{\omega}_{t+1d}^{(d)} - \omega_{t+1d}^{(d)}$

The HAR model with daily realized volatility, weekly realized volatility, and monthly realized volatility could be also represented as simple AR (22) process. This could be described by the following formula:

$$RV_{t+1d}^{(d)} = \alpha + \sum_{i=1}^{22} \theta_i RV_{t-(i-1)d}^{(d)} + \omega_{t+1d} \tag{12}$$

with the constraints

$$\theta_i = \begin{cases} \beta^{(d)} + \frac{1}{5}\beta^{(w)} + \frac{1}{22}\beta^{(m)} & \text{for } i = 1 \\ \frac{1}{5}\beta^{(w)} + \frac{1}{22}\beta^{(m)} & \text{for } i = 2,3,4,5 \\ \frac{1}{22}\beta^{(m)} & \text{for } i = 6,...,22 \end{cases} \tag{13}$$

Empirical and simulation studies provided by Corsi (2009), clearly show that the HAR model could cover a wide variety of stylized facts related to the realized volatility data, which is the financial time series data. In particular, it is observed that the HAR model could replicate fat tails. For realized volatility it is clearly observed that its kurtosis is much higher than the kurtosis observed for normal distribution. Therefore, the probability density function of the realized volatility is leptokurtic. The HAR model also covers well such stylized fact as the long memory dependence, which remains strong through very long time interval and could be clearly observed in autocorrelation functions. There are other stylized facts, which could be covered by the HAR model, such as tails cross-over, self-similarity, multifractality, and volatility cascade. At the same time the HAR model is still very easy in economic interpretation. Moreover, this model could be estimated using simple OLS estimator.

Empirical studies, which were done by Corsi (2009), shows that HAR(3) model in comparison with Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model, Risk Metrics model, AR(1) model, AR(3) model, and Autoregressive Fractionally Integrated Moving Average (ARFIMA) model provide better (lower) measures of forecast accuracy such as root mean squared error, mean absolute error, mean absolute percentage error and Theil inequality coefficient. This result remains valid for in-sample performance and for out-of-sample performance for one day time period, one week time period and two weeks time period. The Minzer-Zarnowitz regression test also provides evidence about better forecast accuracy of HAR(3) model in comparison with GARCH model, Risk Metrics model, AR(1) model, AR(3) model, and ARFIMA model.

Simple estimation and interpretation, good forecast accuracy of HAR model lead to different extensions and applications of this model.

3.1.2 Extensions of HAR model

Several extensions and generalizations of the HAR model have been proposed. The Leveraged HAR (LHAR) model was introduced by Corsi and Reno (2009). This model extended the heterogeneous structure of volatility to account for the leverage effect. Other extension of the HAR model is the Leverage HAR with Continuous volatility and jumps model (LHAR-CJ), also proposed by Corsi and Reno (2009). This model, additionally, takes into account jump-diffusion, which could cause large finite sample bias. As a result model developed by Corsi and Reno (2009) could cover three heterogeneous components: continuous volatility, leverage and jumps. Empirical evidence of better forecast accuracy of LHAR-CJ in comparison with HAR model was provided by Corsi and Reno (2009).

Tree-structured heterogeneous autoregressive (tree-HAR) process was proposed by Audrino and Corsi (2008) to modeling and estimation of the tick-by-tick realized covariance measure, which was obtained as quotient of the realized covariance and the realized volatility. This model according to Audrino and Corsi (2008) could replicate two stylized facts of realized correlation: long-memory strong temporal dependence and existence of structural breaks. The empirical results, which were obtained by Audrino and Corsi (2008), show that tree-HAR model provide better forecast accuracy of realized correlations than AR(1) model, ARMA(1,1) model, ARIMA(1,1,1) model and HAR according to mean absolute error, mean squared error measure, and R-squared coefficient.

The Heterogeneous Autoregressive Gamma with Leverage (HAGL) model was proposed by Corsi et al. (2011) in order to estimate option prices. This is a discrete-time stochastic volatility option pricing model, which is based on the historical high-frequency data. Unobservable returns volatility in this model is included as realized volatility. Realized volatility in HAGL model presented and estimated as autoregressive gamma process of order p. Calculation of weekly (medium-term volatility) and monthly (long-term volatility) realized volatilities in this model differ from that in HAR model and could be described through daily realized volatility (short-term volatility) using following formulas:

$$RV_t^{(w)} = \frac{1}{4} \sum_{i=1}^{4} RV_{t-i}^{(d)}$$

$$RV_t^{(m)} = \frac{1}{17} \sum_{i=5}^{21} RV_{t-i}^{(d)}$$

According to those formulas, realized volatilities aggregated over different horizons are calculated in such a way that there is no overlapping between volatilities related to different terms (short-term, medium-term and long-term).

It was shown by Corsi et al. (2011) that HAGL model could replicate more accurately Q-dynamics of option prices returns. Therefore HAGL model provides better estimation results for option prices returns than GARCH type models.

In summary, the above short review of existing contributions documents the fact that the HAR model has been successfully extended in different directions and applied for different data. This only confirms good statistical and economical features of the HAR model.

3.2 Lasso

3.2.1 General description of Lasso

Lasso was introduced by Tibshirani (1996). The main idea behind Lasso was to find a model, which will introduce regression shrinkage and, as a result, provide selection of explanatory variables for the regression. Therefore, the goal of Lasso was to solve such problem as model selection. This is an actual and important problem in modern econometrics because of its huge impact on forecast accuracy.

The well known and widely applied approaches for solution of explanatory variables selection problem are such approaches as ridge regression or subset selection. However, both of these two approaches have strong drawbacks. The problem of ridge regression is interpretation of results, because this model continuously reduces coefficients near explanatory variables. However, none of the model coefficients is equal to zero. The second approach, subset selection approach, provides results that could be interpreted easily. A drawback of this approach is that the model selection problem doesn't provide one clear solution. This shortcoming has to do with the fact that the model selection is based on inclusion or exclusion of different explanatory variables in the model. Therefore, a clear answer to the question of which model should be selected isn't arrived at by this subset selection approach. Also, estimation of a wide variety of different models in order to select the best of these models could be excessively time consuming.

Lasso proposed by Tibshirani (1996) is described by an operator, which reduces some coefficients of explanatory variables and sets others to zero. Therefore Lasso provides easily interpretable result and, at the same time, results could be estimated easily. Consequently, Lasso could solve the drawbacks of ridge regression and subset selection while, at the same time, retains the positive features of both of these approaches. One of the biggest advantages of Lasso is that two actions are accomplished by a single operator: the first action is model selection and the second action is estimation of the model. One more advantage of Lasso (among several others) is that it also provides an effective estimation, whereby the number of explanatory variables grows with the growth of the sample size.

Lasso could be described using the following formula, which was introduced by Tibshirani

(1996):

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha,\beta} \left\{ \frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha - \sum_{j=1}^{K}\beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^{K}|\beta_j| \leq t, \tag{14}$$

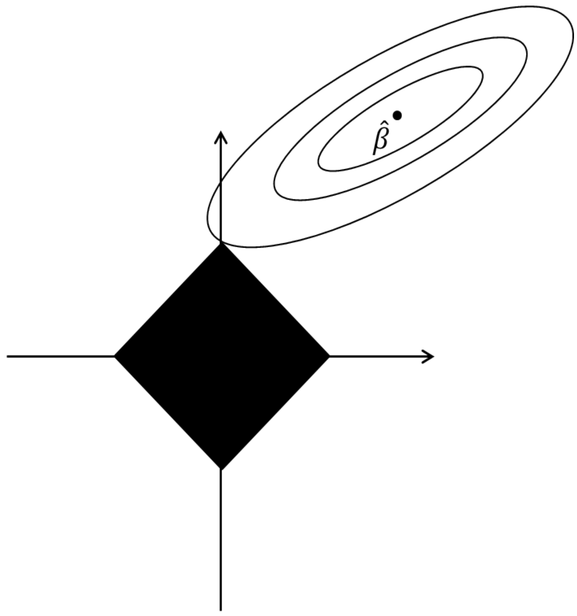
where y_i is the dependent variable, x_{ij} are the explanatory variables, α is the intercept, β_j are the coefficients near explanatory variables and t is a tuning parameter. The tuning parameter should be bigger or equal to zero.

Lasso also could be written in the Lagrangian form as:

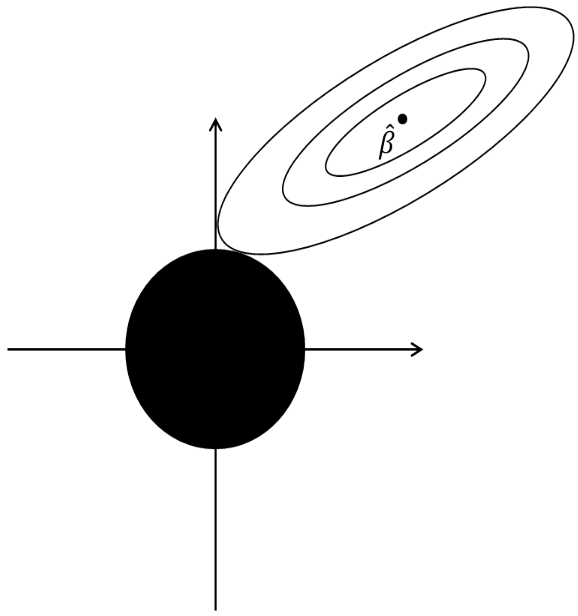
$$\beta^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{K}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{K}|\beta_j| \right\} \tag{15}$$

where λ is a regularization parameter. There is one to one correspondence between λ and t , which can be concluded based on the last two equations.

Geometrical representation of Lasso estimation and ridge regression could be observed on Figure 1 .



(a) Lasso estimation



(b) Ridge regression

Figure 1: Geometrical representation

This figure shows the main difference in the estimation approach of these two models. The estimation of Lasso provides a solution when elliptical contour first touches the square. This occurs at a corner, which leads to a zero coefficient. In the ridge regression contour the ellipse has to touch a circle, which does not have corners, and therefore zero coefficients could be observed very rarely.

Main difference between lasso and ridge regression is that in the Lasso L_1-norm was applied to penalized coefficients β ($\lambda \sum_{j=1}^{p} |\beta_j|$) while in the ridge regression L_2-norm was applied to penalized coefficients β ($\lambda \sum_{j=1}^{p} \beta_j^2$). As a result in the ridge regression those coefficients are shrunk by a constant factor and in the Lasso they are shrunk by a constant factor but with truncating at 0.

For Lasso the problem of optimal lambda selection is crucial. There are several known approaches to solve this problem. The cross-validation approach, generalized cross-validation approach, and the information criteria approach, discussed by Efron and Tibshirani (1993) and by Tibshirani (1996), are most often applied to the solution of the problem of optimal lambda selection for Lasso.

There is known a wide variety of generalizations and extensions of Lasso. Compressive sensing was designed by Donoho (2004) and by Candes (2006); Fused Lasso was proposed by Tibshirani et al. (2005); Elastic net was introduced by Zou and Hastie (2005); Adaptive Lasso was conceived by Zou (2006); Grouped Lasso was proposed by Yuan and Lin (2006); Graphical Lasso was introduced by Yuan and Lin. (2007) and by Friedman et al. (2007); Dantzig selector was introduced by Candes and Tao (2007); Matrix completion was suggested by Candes and Tao (2007) and by Mazumder et al. (2010); Near isotonic regularization was proposed by Tibshirani et al. (2011).

3.2.2 Application of Lasso to autoregressive process

Lasso became a very popular operator for solving model, explanatory variables and other selection problems. It has been applied to various linear and generalized linear models. Among others, Lasso was applied to autoregressive process to solve the problem of selection of explanatory variables. Application of Lasso to autoregressive process (AR) modeling was first introduced by Nardi and Rinaldo (2011).

In the classical time series approaches it is assumed that the time series data fits a mix of two processes. The first process is the autoregressive process and the second one is the moving average process. In other words the time series data fits the Autoregressive Moving Average (ARMA) model. However, for ARMA model it was assumed that the order of the two processes, i.e. the autoregressive process and the moving average process, are known in advance. But this assumption is not fully realistic and caused problems in estimation of ARMA model. As mentioned above, the Lasso provides effective results when the number of parameters increases with increasing number of observations. Therefore Lasso could provide effective results for ARMA model, which was described by Nardi and Rinaldo (2011).

$$x_t = c + \varphi_1 x_{t-1} + ... + \varphi_p x_{t-p} + z_t, \quad t = 1...n$$

where $x_1, \ldots, x_n$ are n observations from AR(p) process, Z_t is the error in the AR process, and represents sequence of independent Gaussian variable with $EZ_t = 0$, $E|Z_t|^2 = \sigma^2$ and $cov(Z_t, X_s) = 0$ for all $s < t$ and c is the intercept.

The autoregressive process of order p penalized using Lasso can be written as:

$$\varphi^{lasso} = \arg\min_{\varphi} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \varphi_0 - \sum_{j=1}^{p} \varphi_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\varphi_j| \right\} \tag{16}$$

Applying Lasso to the autoregressive process provides a solution for the problem of process order selection. Lasso reduces coefficients near irrelevant explanatory variables in the autoregressive process to zero. This provides opportunity to choose explanatory variables that fit the autoregressive process of order p in the best way among all relevant orders of this process. Moreover, as mentioned before, model selection and model estimation will be provided at the same time, when Lasso is applied.

Asymptotic properties of the Lasso estimator applied to AR process such as model selection consistency, estimation and prediction consistency, asymptotic distribution were derived and proved by Nardi and Rinaldo (2011).

Also, an approach to lambda selection for Lasso applied to the autoregressive process was proposed by Nardi and Rinaldo (2011). According to their suggestion, lambda could be found using the following formula:

$$\lambda_n = \sqrt{\frac{\log n \log p}{n}}$$

As further extension of Lasso in application to time series data the Lag weighted Lasso for time series was proposed by Park and Sakaori (2013). The main idea of their model is that the Lasso introduces penalties on coefficients, which are differently weighted. These weights give an opportunity to introduce into the model such feature of time series data as lag effect. In this way the forecast accuracy could be improved in comparison with Lasso and adaptive Lasso when applied to time series data, as shown by Park and Sakaori (2013).

In other research related to the application of Lasso operator to time series data comparison of the Lasso applied to the autoregressive process with the HAR model was performed. This research was presented by Audrino and Knaus (2012). Their hypothesis was that Lasso applied to the autoregressive process should replicate the HAR model, if the HAR model is a true data generation process. This hypothesis is reasonable because the HAR model could be rewritten as AR(22) process, which was shown in this work earlier.

The main goal of Audrino and Knaus (2012) research wasn't improvement of forecast accuracy, rather it was the analysis of realized volatility dynamics. The HAR model assumes that daily realized volatility depends on three partial volatilities aggregated over different time intervals: daily, weekly and monthly. Audrino and Knaus (2012) investigated how these frequencies could really replicate real dynamics of the daily realized volatility. Lasso was applied to the autoregressive process of order 100 (AR(100)) in order to answer this question. It was found that Lasso replicate the HAR model when only the first 22 lags (daily, weekly and monthly realized volatility), from the 100 included in the model, will be non-zero by Lasso estimation.

It was shown by Audrino and Knaus (2012) that the first lag related to daily realized volatility was chosen 1000 times from 1000 replications (100%) for all nine stocks which were analyzed. Inspection of the first five lags related to the weekly realized volatility shows that lags from the second to the fifth were selected 100% times from all replications for six stocks; for other stocks some lags within that range were chosen from 96% to 99% times. Analysis of the first 22 lags related to monthly realized volatility shows that the lags form the sixth to the twenty second weren't selected 100% time for any of the nine analyzed stocks. The percentage of time when a specific lag from within that range was selected fluctuates between 0% to 61%.

Also, it is observed that there are lags beyond the twenty second with percentage of selection higher than zero. This results could lead to the conclusion that Lasso doesn't replicate the HAR model. Therefore, the assumption that the HAR model is a true data generation processes couldn't be confirmed by the results of this analysis. Moreover, it was mentioned by Audrino and Knaus (2012)) that the results of their research provides some evidence that realized variance could be described more accurately using model which includes variables only from short time horizons.

Also, the research of Audrino and Knaus (2012) documented that there was almost no difference in the mean squared error, and therefore also in the forecast accuracy, between Lasso and the HAR models. This was supported by the results of Diebold-Mariano test.

3.3 Estimator based on penalized mixture distribution

In two previous sections of this thesis two different approaches to estimation of realized volatility, which could replicate persistence of long-memory dependence in the data, was described. In this section new variant of MLE estimator, which could be applied to these two approaches, will be proposed. The main aim of this new estimator is the solution of appropriate distribution selection problem.

The HAR model and Lasso applied to autoregressive process was estimated using simple Ordinal Least Square (OLS) estimator. Also, both of these models could be estimated using maximum likelihood estimator (MLE) based on one of the already known distributions.

The normal distribution is selected most often as an appropriate distribution to fit the analyzed data well. However, the probability density function and tests used to check whether or not the data is distributed normally, such as Jarque-bera test and Shapiro-Wilk test, clearly reject the hypothesis that daily realized volatility data has normal distribution. Therefore, the issue of which distribution could fit realized volatility data well remains open. Moreover, sometimes the analyzed data couldn't be described well using only one distribution. Description of the data using mixture distributions is an appropriate solution in such case. Mixture distributions could fit well the data which in addition to a global maximum has also some local maxima as well as data which has heavy tails and high kurtosis.

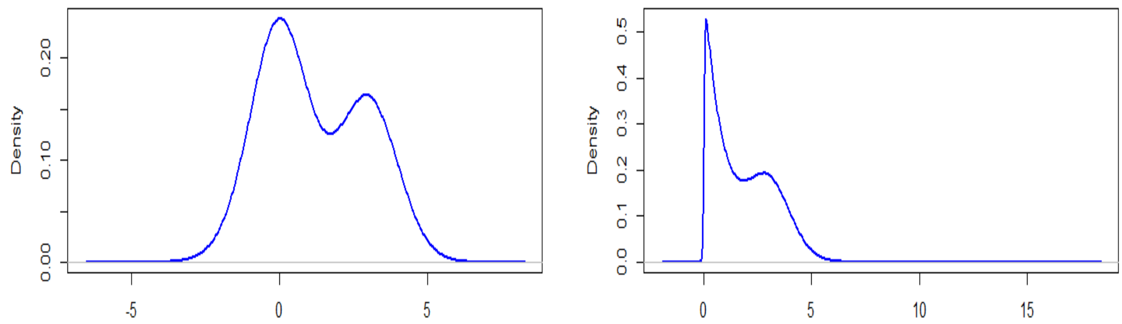
One of the biggest problems and disadvantages in estimation using MLE is that in order to apply this estimator distribution of the analyzed data should be known. Application of MLE to wrong distribution could lead to big estimation error and, therefore, inaccurate forecast could be provided. However, the problem of unknown distribution of the data is common in modern econometrics and is the source of serious difficulties. Search for appropriate mixture of some distributions to fit data well increases complexity of distribution selection problem. Therefore, a question, which often arises, is how to efficiently select a distribution, or a mixture of distributions, which will fit the estimated data well.

Probability density function of mixture distributions could be described using the following formula:

$$f(x) = \sum_{i=1}^{K} \omega_i f_i(x) \tag{17}$$

where K is number of distributions included in probability density function of mixture distributions , $f_i(x)$ is corresponding probability density function of each distribution, and ω_i is corresponding weights of each distribution, such that each weight should satisfied following restrictions: $\omega_i \geq 0$ and $\sum_{i=1}^{N} \omega_i = 1$.

Examples of probability density function of mixture distribution could be observed on Figure 2.



(a) mixture of normal (mean=3,sd=1) and Student's t (df=30) distributions

(b) mixture of normal (mean=3,sd=1) and Weibull (shape=1,scale=1) distributions

Figure 2: PDF of mixture distributions

In view of the above description the log likelihood function based on mixture distribution could be expressed by the following formula:

$$\log L(\theta, \omega; x_1 ... x_N) = \frac{1}{N} \sum_{j=1}^{K} \omega_j \sum_{i=1}^{N} \log f_j(x_i|\theta) \tag{18}$$

where $\log L(\theta, \omega; x_1 ... x_N)$ is likelihood function which consists of two inputs. The first input is the vector of parameters θ, which includes contains parameters of distributions included in the mixture and parameters of the model. The second input is the vector of observations $x_1 ... x_N$ with N being the number of observations.

The distribution selection problem for MLE could be more specifically stated as determination of which existing distributions should be included in mixture distribution probability density function. This problem is similar to the problem of appropriate explanatory variable selection. As described earlier, Lasso is a very efficient estimator applicable to the solution of the model selection problem. Lasso was also successfully applied to autoregressive process in order to solve problem of order selection. Therefore, in this research it is proposed to apply Lasso to solve the problem of appropriate distribution selection.

To solve the problem of appropriate distribution selection in the model Lasso could be applied as a penalization to weighting parameter of mixture distributions ω_i. The maximum likelihood function with Lasso applied to weighting parameter of mixture distributions could be presented using the following formula:

$$L(\theta, \omega; x_1, \ldots, x_n) = \sum_{j=1}^{K} \omega_j \sum_{i=1}^{N} \log\ f_j(x_i|\theta) + \lambda \sum_{j=1}^{K-1} |\omega_j| \tag{19}$$

where λ is the penalization parameter and $f_j(x_i|\theta)$ is the probability density function with parameters θ.

According to the above expression, the maximum likelihood estimator (MLE) could be described using the following formula:

$$\begin{aligned}(\theta, \omega_{lasso}) =& \arg\min_{\theta} \{-L(\theta, \omega; x_1, \ldots, x_n)\} = \\ =& \arg\min_{\theta} \left\{ -\frac{1}{N} \sum_{j=1}^{K} \omega_j \sum_{i=1}^{N} \log\ f_j(x_i|\theta) + \lambda \sum_{j=1}^{K-1} |\omega_j| \right\}\end{aligned} \tag{20}$$

Application of Lasso to weights of mixture distributions will shrink weights of some distributions to zero and, therefore, a selection of distributions appropriate for the analyzed data will be provided. Estimation of this model could be achieved by using some of the already known optimization algorithm with constraints applied to weighting parameters of mixture distributions. Penalizing parameter lambda could be selected using cross-validation, generalized cross-validation, and information criteria approaches. The result obtained using MLE based on penalized mixture distributions will be easy to interpret.

In order to check accuracy of this estimator several simulations were performed. Data related to different distributions was used in those simulations. Then, the simulated data and different distributions, including true and false distributions, were estimated using MLE based on the penalized mixture distributions. The obtained simulation results showed that this estimator choose correct weights for the considered distributions, with the average estimation accuracy of approximately 84.3%. This result shows that MLE based on penalized distribution provides a choice of distributions which is appropriate for the analyzed data.

The HAR model and Lasso applied to autoregressive process could be estimated using MLE based on penalized mixture distributions. First, an assumption that the residuals in

HAR model and Lasso have mixture distribution should be made. Second, distributions that could potentially fit well the residuals in the model should be selected. Then, MLE based on penalized mixture distribution could be applied in order to estimate HAR model and Lasso applied to autoregressive process.

Empirical results of this estimator applied to the HAR model and Lasso employed to autoregressive process in order to estimate and forecast daily realized volatility will be provided in section four.

4 Empirical Results

4.1 Data description

The daily realized volatility data used in this research was taken from the Realized Library of Oxford-Man Institute of Quantitative Finance, which is provided on its website `http://realized.oxford-man.ox.ac.uk`. This data was calculated based on the raw high frequency data, which Oxford-Man Institute of Quantitative Finance received from Reuters DataScope Tick history database.

The data processed in such manner, was used to calculate realized volatility measures and was related only to the time interval within which the stock exchange was open. This approach helps to eliminate seasonal fluctuations. The realized volatility measures calculated by the Oxford-Man Institute of Quantitative Finance, don't take into account the overnight volatility as well as the volatility at the beginning of the trading day (first few minutes), because then a big error could be introduced.

The main topic of the research presented in this thesis is volatility, with particular emphasis on its estimation employing realized kernel with Parzen weight function, as described by Barndorff-Nielsen et al. (2009). This estimator is more robust with respect to some market frictions, which make it more efficient than others realized variance measures.

The data set provided by Realized Library of Oxford-Man Institute of Quantitative Finance and used in this research covers the period of around 15.5 years (from the 3rd of January 2000 till the 20th of May 2015, and includes almost 4000 observations).

For this research five stock market indexes described in this paragraph were chosen. Standard&Poor's 500 (S&P 500) is the stock market index in the United States, which includes 500 large companies listed on the NYSE (New York Stock Exchange) and on the NASDAQ (National Association of Securities Dealers Automated Quotations, Stock Exchange)). The Financial Times Stock Exchange 100 Index (FTSE 100) is the stock market index in the United Kingdom, which includes 100 companies listed on London Stock Exchange. Nikkei Heikin Kabuka (Nikkei 225) is the stock market index in Japan for Tokyo Stock Exchange. Deutscher Aktienindex (DAX) is the stock market index in Germany, which includes 30 major German companies listed on the Frankfurt Stock Exchange. Hang Seng is the stock market index in Hong Kong (China), which includes largest companies listed on the Hong Kong Stock Exchange). These indexes represent biggest stocks exchanges in the world, located in its different regions (America, Europe and Asia).

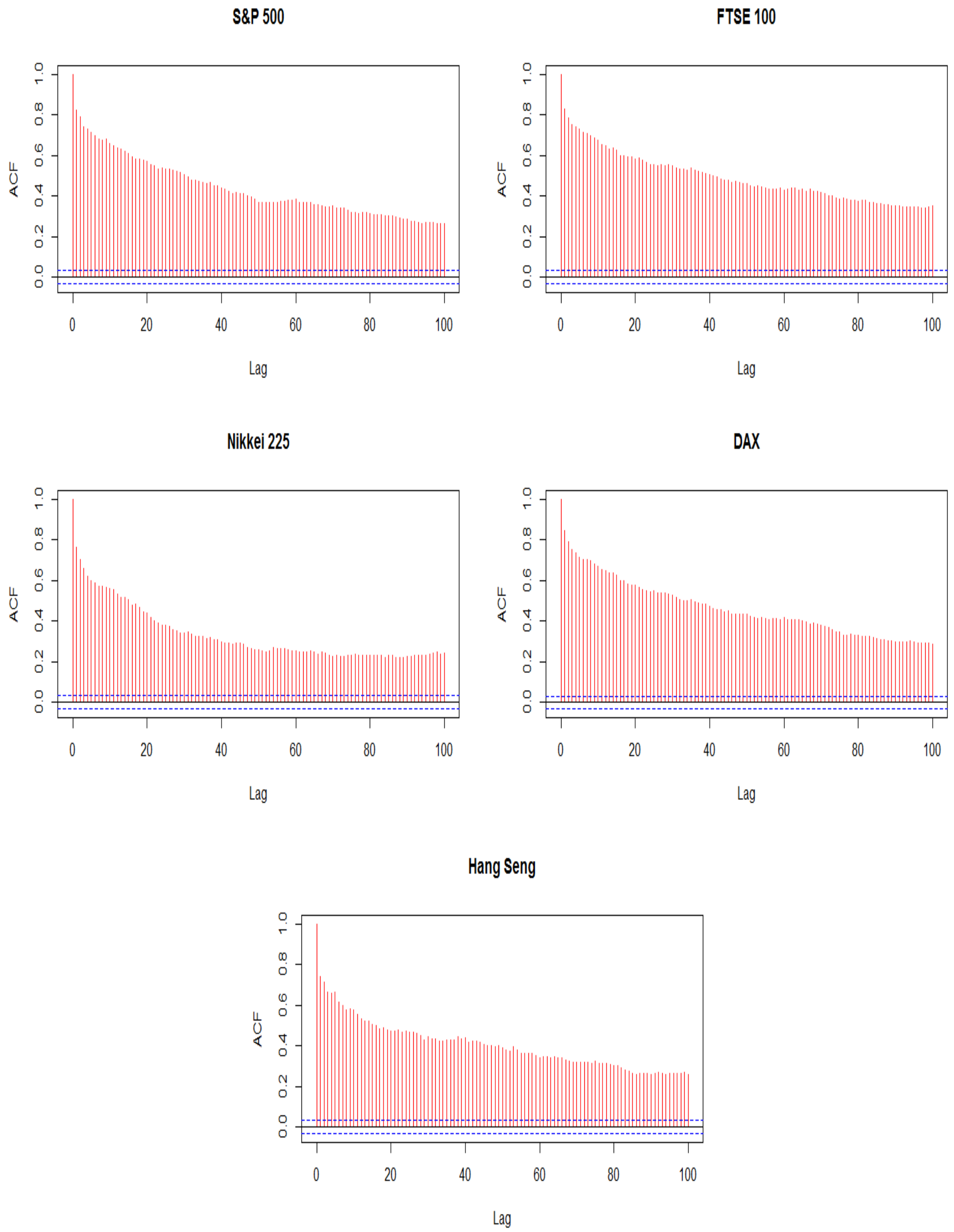


Figure 3: Autocorrelation function of realized volatility of stock market indexes

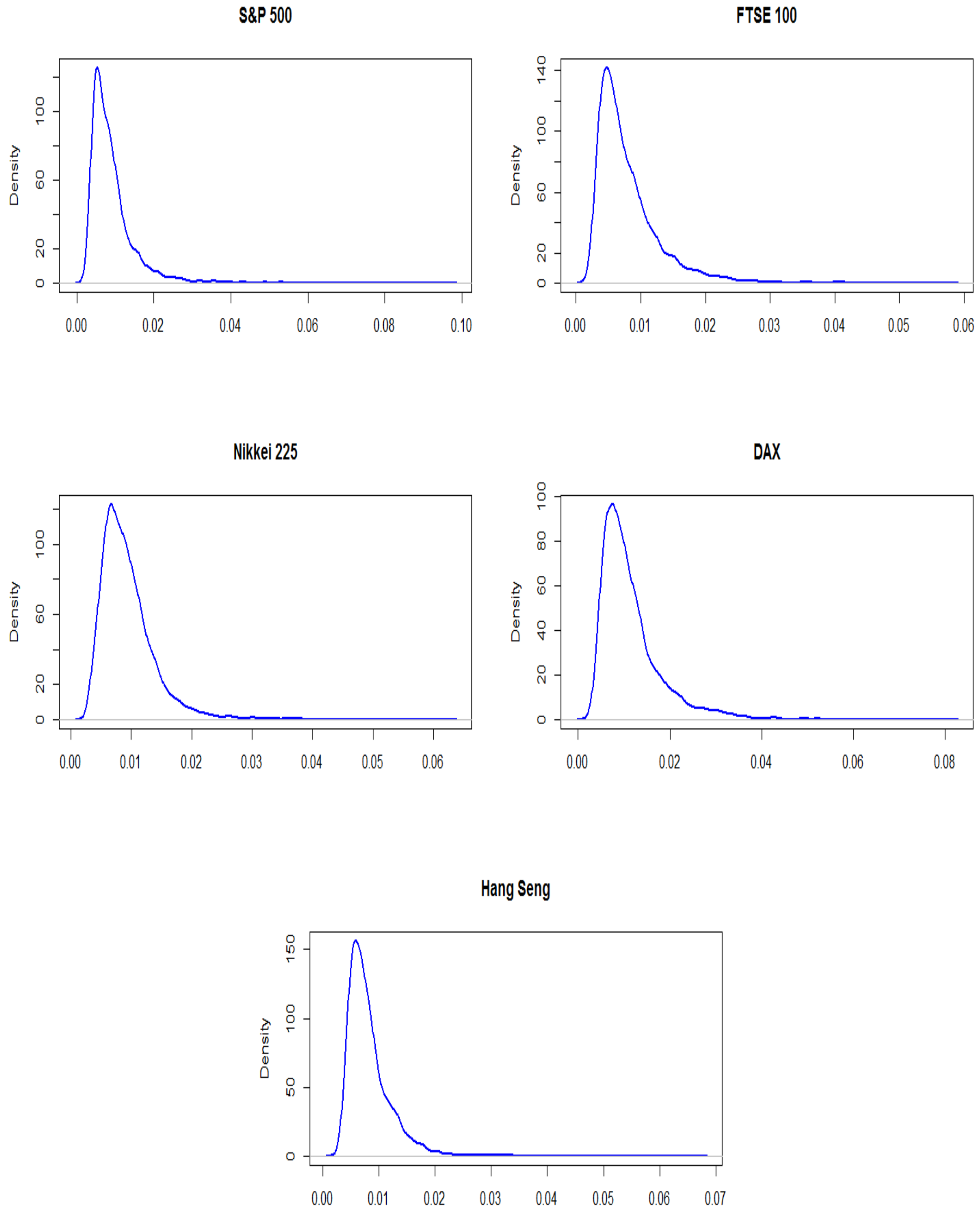


Figure 4: Probability density function of realized volatility of stock market indexes

Slowly decaying autocorrelation function is observed for all 5 stock market indexes analyzed in this research in Figure 3. It is clearly shown in that figure that even after 100 lags (approx. 5 months) significant autocorrelation exists. This means that long memory persistence exists. Therefore, the important characteristics of the proposed model, must relate to long memory dependence.

The probability density plots of daily realized volatility of stock market indexes analyzed in this research are provided on Figure 4. It can be clearly observed that the distribution of the daily realized volatility of all five stock market indexes is leptokurtic and, at the same time, right-skewed.

Descriptive statistics of realized volatility of different stock market indexes is shown in Table 1.

	Mean	Std Dev	Kurtosis	Jarque-Bera	Prob
S&P 500	0.009	0.006	21.989	84594.867	0.000
FTSE 100	0.008	0.005	11.136	24100.382	0.000
Nikkei 225	0.010	0.005	17.750	54728.155	0.000
DAX	0.012	0.007	12.414	29534.387	0.000
Hang Seng	0.008	0.005	28.922	131023.157	0.000

Table 1: Descriptive statistics of realized volatility of stock market indexes

It could be observed, that kurtosis of realized volatility for all five indexes is much higher than in normal distribution. This confirmed what was already observed on the probability density plot, namely that the realized volatility of stock market indexes is leptokurtic. The standard deviation is very small for all five indexes, which show very small variation from the mean value in daily realized volatility, as clearly seen also in the probability density plot. Therefore, it is logical to conclude that Jargue-Bera test reduced H_0 hypothesis about normality of data for all indexes.

According to the results provided by autocorrelation functions, probability density plot and descriptive statistics of the data, which clear show presence of long memory dependence and non-normal distribution, it is sensible to apply the HAR model and Lasso to this data.

4.2 Estimation of HAR model and Lasso for realized volatility

4.2.1 Estimation of HAR model using OLS and MLE based on penalized mixture distribution

HAR model introduced by Corsi (2009) as was mentioned in section 3 can be described by the following formula:

$$RV^{(d)}_{t+1d} = c + \beta^{(d)} RV^{(d)}_t + \beta^{(w)} RV^{(w)}_t + \beta^{(m)} RV^{(m)}_t + \omega_{t+1d}$$

To estimate HAR model, daily, weekly and monthly realized volatility is needed. The daily realized volatility is available, however weekly and monthly volatilities need to be calculated in order to estimate HAR model.

The daily realized volatility was aggregated in weekly and monthly realized volatility using the following formulas:

$$RV^{(w)}_t = \frac{1}{5}\left(RV^{(d)}_t + RV^{(d)}_{t-1d} + RV^{(d)}_{t-2d} + RV^{(d)}_{t-3d} + RV^{(d)}_{t-4d}\right) \tag{21}$$

$$RV^{(m)}_t = \frac{1}{22}\left(RV^{(d)}_t + RV^{(d)}_{t-1d} + ... + RV^{(d)}_{t-21d}\right) \tag{22}$$

Kurtosis of realized volatility for all aggregation levels (daily, weekly and monthly) for all five stock market indexes analyzed in this research is shown in Table 2.

	Daily RV	Weekly RV	Monthly RV
S&P 500	22.01	13.07	9.94
FTSE 100	11.18	5.72	3.39
Nikkei 225	17.74	15.16	12.75
DAX	12.60	7.33	3.39
Hang Seng	30.44	17.44	6.41

Table 2: Kurtosis of realized volatility of stock market indexes

It is clearly observed that for all five stock market indexes kurtosis decreases, when aggregation period increases. Therefore, aggregation makes data more smooth, and tails become thinner. As a result the distribution of the data after aggregation is less leptokurtic than it was without aggregation.

After aggregation was done, all variables, which are needed to estimate the HAR model, are known. Therefore, parameters of the HAR model could be estimated. As shown before, mean of all variables are near zero. Therefore, the HAR model without intercept was chosen for estimation.

In Table 3 coefficients of the explanatory variables included in the model and their p-values for all five stock exchange indexes (S&P 500, FTSE 100, Nikkei 225, DAX, and Hang Seng) are shown. These results have been obtained according to estimation parameters of the HAR model using OLS. The model was estimated in Program R and lm function was used for the estimation of the model.

		RV daily	RV weekly	RV monthly
S&P 500	Coefficient	0.405	0.379	0.194
S&P 500	p-value	0.000	0.000	0.000
FTSE 100	Coefficient	0.390	0.398	0.201
FTSE 100	p-value	0.000	0.000	0.000
Nikkei 225	Coefficient	0.429	0.310	0.250
Nikkei 225	p-value	0.000	0.000	0.000
DAX	Coefficient	0.459	0.370	0.234
DAX	p-value	0.000	0.000	0.000
Hang Seng	Coefficient	0.261	0.519	0.209
Hang Seng	p-value	0.000	0.000	0.000

Table 3: Estimation coefficients of the HAR model using OLS

RV daily, RV weekly and RV monthly coefficients are significant - at 99% level for S&P 500, FTSE 100, Nikkei 225, DAX and Hang Seng stock market indexes. The sign of all coefficients for all stock market indexes is positive. Therefore, the daily realized volatility is sum of partial volatilities with different weights. The results received in my research coincide with the result which was received by Corsi (2009).

The second estimator, which is proposed for HAR model is MLE based on penalized mixture distributions. Two normal distributions, lognormal distribution, Weibull distribution and Student's t-distribution were selected for this research. Distributions were selected taking into account different distribution parameters (mean and standard deviation for normal and lognormal distributions, shape and scale parameters for Weibull distribution, degrees of freedom and non-centrality parameter for Student's t-distribution), and their ability to fit, as well as possible, the residuals of the estimated model.

Lasso penalty was applied to weighting parameters of distributions, which was included in mixture distribution function. This penalty should help to select the mixture of distributions that fits the model in the best way.

Maximum likelihood function with penalized mixture distribution was maximized in Program R using function constrOptim with Nelder-Mead algorithm. Nelder-Mead method is a numerical optimization method, which uses only function values and does not use its derivatives.

The obtained results are shown in Table 4.

	Normal distrib1	Normal distrib2	Weibull distrib	Lognorm distrib	Student's t-distrib	RV day	RV week	RV month
S&P 500	0.0%	32.2%	0.3%	0.2%	67.3%	0.387	0.366	0.245
FTSE 100	69.1%	0.1%	3.3%	3.8%	23.7%	0.293	0.491	0.212
Nikkei 225	0.0%	33.1%	0.0%	0.0%	66.9%	0.318	0.325	0.360
DAX	29.9%	0.0%	0.0%	0.0%	70.1%	0.279	0.664	0.196
Hang Seng	25.6%	0.0%	0.0%	0.0%	74.4%	0.169	0.672	0.127

Table 4: Estimation coefficients of the HAR model using MLE based on penalized mixture distribution

It could be observed that for different stock market indexes different distributions were chosen using HAR model estimated by MLE based on penalized mixture distributions. For S&P 500 stock market index normal distribution, lognormal distribution, Weibull distribution and Student's t-distributions were selected; for FTSE 100 stock market index all five distributions included in mixture distribution function were chosen; for Nikkei, DAX and Hang Seng stock market indexes normal distribution and Student's t-distributions were selected. However, when the percentage of each distribution is analyzed, it is clearly observed that two distributions with the biggest percentage for each of indexes are normal distribution and Student's t-distribution. Moreover, for three of the five stock market indexes only these two distributions were chosen. Therefore, it could be concluded that a mixture of these two distributions describe realized volatility of stock market indexes in the best manner.

When the resulting coefficients of explanatory variables of the HAR model estimated in two different ways are compared, it could be noted that the sign is the same (positive) for all coefficients and for all five indexes. However, when their values are compared, some difference could be noted between coefficients near the same explanatory variable of each index. Nevertheless, the conclusion that daily realized volatility tomorrow is a weighted sum of daily, weekly and monthly realized volatilities today still remains the same.

4.2.2 Estimation of Lasso model using MLE based on normal distribution and MLE based on penalized mixture distribution

The second model which was estimated in this research was Lasso. The Lasso is an operator widely applied to solve the problem of model selection. Therefore, Lasso is suitable to analyze the data used in this research, as the autocorrelation function of stock indexes show existence of long memory dependence. This is associated with the problem of lag selection, since when the long memory dependence is present it is not clear how many lags from the past and exactly which lags should be included in the regression model. The Lasso could be employed to address the problem of selecting lags, so as to have the highest explanatory power, and therefore, to predict the future daily realized volatility in the most accurate way.

The Lasso estimation will also make it possible to perform a comparison of the Lasso lag selection with the HAR model lag selection, which selects 22 lags (one month period). It will also be possible to observe if Lasso could or could not recover the HAR model for stock indexes analyzed in this research.

The Lasso can be described by the following formula:

$$\beta_{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ -\frac{\sum_{i=1}^{N} L(y_i, x_i; \beta)}{N} + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{23}$$

In the estimation process of Lasso adopted in this research, Lasso and Elastic-Net Regularized Generalized Linear Models (glmnet) package in R were used. This package was written by Friedman et al. (2015) with the aim of fitting Lasso. The elastic-net regularization was applied for different regression models such as Gaussian, logistic and multinomial regression models, as well as other models. In this research the Lasso model for daily realized volatility of stock exchange indexes was estimated using penalized MLE based on normal distribution. As the optimization method this package uses combination of two optimization methods: Newton optimization for outer loop and weighted least-squares optimization for inner loop. The penalization parameter lambda for lasso was selected using cross-validation, with the mean squared error chosen as a measure.

The coefficients obtained from the Lasso estimated using penalized MLE based on normal distribution, are provided in Table 5. The first lag (the lag related to daily realized volatility) was selected for all five stock market indexes, which were analyzed in this research. All first five lags (the lags related to weekly realized volatility) are non-zero for stock market indexes considered herein (FTSE 100, Nikkei 225, DAX and Hang Seng). For S&P 500 stock market index four of the first five lags are non-zero (coefficient of the third lag is equal to zero). There are no stock market indexes analyzed in this research, where all first 22 lags (the lags related to monthly realized volatility) were selected. For S&P 500 stock market index 11 lags of the first 22 were selected, for FTSE 100 stock market index there were 13 lags selected, for Nikkei 225 stock market index also 13 lags were selected, 10 lags were selected for DAX stock market index, and 8 lags for Hang Seng stock market index. For all five stock market indexes there are also non-zero lags beyond the first 22 lags. For S&P 500 stock market index there were 7 lags after 22 lag selected, for FTSE 100 stock market index there were 10 such lags, for Nikkei 225 stock market index there were 4 lags, for DAX stock market index 6 lags beyond 22 were selected, and for Hang-Seng stock market index there were 6 lags. In general, for S&P 500 stock market index 18 lags were selected; for FTSE 100 stock market index 23 lags were selected; for Nikkei 225 stock market index 17 lags were selected, for DAX stock market index 16 lags were selected and for Hang Seng stock market index 14 lags were selected.

It is also interesting to know that for FTSE 100 stock market index 100 lag was selected, and for Nikkei 225 stock market index 98 was selected. This result means that even after 100 lags long memory dependence still could exist, which was also clearly observed in autocorrelation function, which was presented earlier. Also, it could be observed that for different stock market indexes, analyzed in this research and estimated by Lasso using penalized MLE based on normal distribution, different lags were selected.

Analysis of lag selection using Lasso estimated using penalized MLE based on normal distribution clearly demonstrates that there are lags with high explanatory power beyond the first 22 lags for all five stock market indexes, and also that not all lags for all five stock market

Lag	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng	Lag	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng
Int	0.0005	0.0000	0.0007	0.0006	0.0005	X_{t-51}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-1}	0.4445	0.4438	0.4488	0.5092	0.3457	X_{t-52}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-2}	0.2056	0.1628	0.1472	0.1425	0.1998	X_{t-53}	0.0000	0.0000	0.0000	0.0000	0.0200
X_{t-3}	0.0000	0.0282	0.0794	0.0148	0.0352	X_{t-54}	0.0000	0.0000	0.0130	0.0000	0.0000
X_{t-4}	0.0887	0.0856	0.0111	0.0675	0.0583	X_{t-55}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-5}	0.0405	0.0443	0.0208	0.0136	0.1274	X_{t-56}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-6}	0.0015	0.0149	0.0245	0.0000	0.0000	X_{t-57}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-7}	0.0007	0.0351	0.0000	0.0501	0.0000	X_{t-58}	0.0120	0.0000	0.0000	0.0000	0.0000
X_{t-8}	0.0285	0.0354	0.0436	0.0662	0.0000	X_{t-59}	0.0000	0.0010	0.0000	0.0000	0.0000
X_{t-9}	0.0720	0.0121	0.0322	0.0000	0.0388	X_{t-60}	0.0310	0.0000	0.0000	0.0160	0.0000
X_{t-10}	0.0000	0.0066	0.0175	0.0050	0.0224	X_{t-61}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-11}	0.0000	0.0000	0.0336	0.0000	0.0000	X_{t-62}	0.0000	0.0130	0.0000	0.0000	0.0000
X_{t-12}	0.0000	0.0000	0.0000	0.0082	0.0000	X_{t-63}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-13}	0.0156	0.0000	0.0000	0.0000	0.0000	X_{t-64}	0.0000	0.0000	0.0060	0.0030	0.0000
X_{t-14}	0.0000	0.0351	0.0162	0.0288	0.0000	X_{t-65}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-15}	0.0001	0.0000	0.0018	0.0000	0.0000	X_{t-66}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-16}	0.0000	-0.0230	0.0000	0.0000	0.0000	X_{t-67}	0.0000	0.0030	0.0000	0.0000	0.0000
X_{t-17}	0.0000	0.0000	0.0093	0.0000	0.0000	X_{t-68}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-18}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-69}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-19}	0.0103	0.0000	0.0000	0.0000	0.0000	X_{t-70}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-20}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-71}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-21}	0.0000	0.0185	0.0000	0.0000	0.0000	X_{t-72}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-22}	0.0000	0.0000	0.0000	0.0000	0.0268	X_{t-73}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-23}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-74}	0.0000	-0.0020	0.0000	0.0000	0.0000
X_{t-24}	0.0038	0.0000	0.0000	0.0000	0.0167	X_{t-75}	0.0000	-0.0020	0.0000	0.0000	0.0000
X_{t-25}	0.0000	0.0000	0.0000	0.0089	0.0032	X_{t-76}	0.0000	0.0000	0.0000	-0.0070	0.0000
X_{t-26}	0.0180	0.0000	0.0000	0.0000	0.0007	X_{t-77}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-27}	0.0006	0.0011	0.0000	0.0024	0.0000	X_{t-78}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-28}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-79}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-29}	0.0000	0.0105	0.0000	0.0000	0.0000	X_{t-80}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-30}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-81}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-31}	0.0000	0.0000	0.0006	0.0000	0.0000	X_{t-82}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-32}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-83}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-33}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-84}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-34}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-85}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-35}	0.0000	0.0372	0.0000	0.0240	0.0000	X_{t-86}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-36}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-87}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-37}	0.0019	0.0000	0.0000	0.0000	0.0000	X_{t-88}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-38}	0.0000	0.0000	0.0000	0.0000	0.0431	X_{t-89}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-39}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-90}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-40}	0.0000	0.0000	0.0000	0.0000	0.0057	X_{t-91}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-41}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-92}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-42}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-93}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-43}	0.0000	-0.0117	0.0000	0.0000	0.0000	X_{t-94}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-44}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-95}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-45}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-96}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-46}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-97}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-47}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-98}	0.0000	0.0000	0.0220	0.0000	0.0000
X_{t-48}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-99}	0.0000	0.0000	0.0000	0.0000	0.0000
X_{t-49}	0.0000	0.0000	0.0000	0.0000	0.0000	X_{t-100}	0.0000	0.0130	0.0000	0.0000	0.0000
X_{t-50}	-0.0247	0.0000	0.0000	0.0000	0.0000						

Table 5: Estimation coefficients of Lasso using MLE based on normal distribution

indexes within the first 22 lags are selected via lasso as lags with high explanatory power. Also, it is clearly observed that for different stock market indexes analyzed in this research, different lags where selected.

Comparison of lag selection via lasso estimated using penalized MLE based on normal distribution with lag selection in HAR model leads to the conclusion that Lasso selection doesn't fully recover the HAR model for the five stock market indexes analyzed in this research. This result coincides with the result obtained by Audrino and Knaus (2012), which also clearly show that Lasso applied to autoregressive process doesn't replicate HAR model.

The second estimator, which was proposed for Lasso, involves two steps. In the first step, Lasso was estimated using penalized MLE based on normal distribution, as described above. After this step is completed, lags with non-zero coefficients were identified. In the second step, after lags with highest explanatory power were chosen, coefficients related to those lags where reestimated using MLE based on penalized mixture distribution. As a result penalty was applied two times: the first time for the explanatory variables selection and the second time for distributions in order to choose distributions, which fit the residuals most precisely.

Distributions, used in estimation of this model were selected to be the same as for HAR model: two normal distributions, lognormal distribution, Weibull distribution and Student's t-distribution.

The optimization function and algorithm applied to lasso estimated using MLE based on penalized mixture distributions was function contrOptim in Program R with Nelder-Mead algorithm, which is similar to that applied to HAR model estimated using MLE with penalized mixture distribution.

The parameters for Lasso estimated using MLE based on penalized mixture distributions are provided in Table 6. For S&P 500 stock market index all five distributions were chosen, however two normal distributions have probability equal to 0.2% and 0.1% respectively. For FTSE 100 stock market index 4 of 5 distributions were selected: normal distribution, Weibull distribution, lognormal distribution and Student's t-distribution. Moreover, the normal distribution and Weibull distribution have probability 0.1%. For Nikkei 225 stock market index also 4 of 5 distributions were selected: two normal distributions (one of the normal distributions has probability 0.01%), lognormal distribution and Student's t-distribution. For DAX stock market index all five distributions were chosen, but Weibull distribution has probability

	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng
Normal distr 1	0.2%	0.0%	5.8%	1.9%	10.1%
Normal distr 2	0.1%	0.1%	0.1%	3.8%	4.4%
Weibull distr	1.5%	0.1%	0.0%	0.1%	0.0%
Lognormal distr	4.5%	2.4%	2.1%	2.5%	9.7%
Student's t-distr	93.7%	97.4%	92.0%	91.7%	75.8%
Intercept	0.000	-0.001	0.001	0.000	0.003
X_{t-1}	0.349	0.503	0.544	0.37	0.244
X_{t-2}	0.294	0.141	0.029	0.15	0.310
X_{t-3}	-	0.167	0.076	0.044	0.047
X_{t-4}	0.052	0.153	0.146	0.029	0.094
X_{t-5}	0.040	-0.013	-0.02	0.131	-0.007
X_{t-6}	0.002	0.009	0.048	-	-
X_{t-7}	0.073	-0.023	-	0.005	-
X_{t-8}	-0.097	0.003	-0.162	-0.014	-
X_{t-9}	0.040	0.020	0.056	-	-0.033
X_{t-10}	-	0.006	0.112	0.047	0.006
X_{t-11}	-	-	-0.022	-	-
X_{t-12}	-	-	-	-0.017	-
X_{t-13}	0.002	-	-	-	-
X_{t-14}	-	0.084	0.078	0.078	-
X_{t-15}	0.039	-	-0.065	-	-
X_{t-16}	-	-0.040	-	-	-
X_{t-17}	-	-	0.107	-	-
X_{t-19}	-0.013	-	-	-	-
X_{t-21}	-	-0.004	-	-	-
X_{t-22}	-	-	-	-	0.008
X_{t-24}	0.02	-	-	-	0.047
X_{t-25}	-	-	-	0.04	0.028
X_{t-26}	-0.035	-	-	-	-0.037
X_{t-27}	0.029	0.051	-	-0.038	-
X_{t-29}	-	-0.087	-	-	-
X_{t-31}	-	-	0.032	-	-
X_{t-35}	-	0.039	-	0.029	-
X_{t-37}	0.025	-	-	-	-
X_{t-38}	-	-	-	-	0.034
X_{t-40}	-	-	-	-	0.027
X_{t-43}	-	0.002	-	-	-
X_{t-50}	-0.022	-	-	-	-
X_{t-53}	-	-	-	-	-0.011
X_{t-54}	-	-	0.028	-	-
X_{t-58}	0.010	-	-	-	-
X_{t-59}	-	-0.047	-	-	-
X_{t-60}	-0.013	-	-	0.005	-
X_{t-62}	-	0.035	-	-	-
X_{t-64}	-	-	-0.018	0.019	-
X_{t-67}	-	0.044	-	-	-
X_{t-74}	-	-0.036	-	-	-
X_{t-75}	-	-0.040	-	-	-
X_{t-76}	-	-	-	-0.058	-
X_{t-98}	-	-	0.036	-	-
X_{t-100}	-	0.110	-	-	-

Table 6: Estimation coefficients of Lasso using MLE based on penalized mixture distribution

0.1%. For Hang Seng stock exchange index 4 of 5 distributions were chosen as distributions with non-zero probability: two normal distributions, lognormal distribution, and Student's t distribution.

After analyses of all selected distributions were performed, it could be easily observed that the main probability for all five stock market indexes analyzed in this research has the Student's t-distribution, with probability more than 90% for S&P 100, FTSE 100, DAX and Nikkei 225 stock market indexes. Hang Seng stock market index has probability for Student's t-distribution equal to 75.8%. This means, that Student's t-distribution could fit residuals of Lasso better than other distributions.

Comparison of the distributions, selected for Lasso with the distributions selected for HAR model leads to the following conclusions. Two distributions with the highest probability for HAR model were the normal and Student's t distributions, however for Lasso the distribution with the highest probability is clearly the Student's t-distribution. Distributions selected as appropriate for stock market indexes is different for Lasso and for HAR model for all five indexes.

If one compares the estimated coefficients near the selected lags in Lasso estimated using penalized MLE based on normal distribution with those in Lasso estimated using MLE based on penalized mixture distributions an additional conclusion can be drawn. The values of the coefficients are different for all five indexes, and even signs for some coefficients for all five stock market indexes are also different, when Lasso estimated using two different estimators is employed.

It also could be observed that the coefficients have smaller difference in HAR model estimated using two different estimators then in Lasso estimated using two different estimators. This could be explained by that fact that in HAR model the estimation using MLE based on penalized mixture distribution was fully performed in one step, whereas estimation of Lasso using MLE based on penalized mixture distribution was divided in two separate steps.

4.3 Forecasting realized volatility using HAR model and Lasso

In this section forecast accuracy measure for HAR model estimated using OLS and MLE based on penalized mixture distribution and for Lasso estimated using penalized MLE based on normal distribution and MLE based on penalized mixture distribution, will be examined. The main forecast accuracy measures selected in this research are the mean squared error and

Minzer-Zarnowitz regression test proposed by Mincer and Zarnowitz (1969).

The mean squared error could be described using the following formula:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{RV}_{t+1}^{(d)} - RV_{t+1}^{(d)}\right)^2 \tag{24}$$

Whereas the Minzer-Zarnowitz regression test could be described using the following formula:

$$RV_{t+1}^{(d)} = \beta_0 + \beta_1\widehat{RV}_{t+1}^{(d)} + error \tag{25}$$

This section provides also the values of the R-squared coefficient, adjusted R-squared coefficient, and plot, which will compare real daily realized volatility with the forecasting values of daily realized volatility using different models.

4.3.1 Forecast accuracy of HAR model for realized volatility

For HAR model estimating, using OLS for quality analysis of the model and for forecast accuracy, the following parameters were estimated: R-squared coefficient, adjusted R-squared coefficient, mean squared error, Akaike information criterion, and Schwarz (Bayesian) information criterion. The results are presented in Table 7.

	R-squared	Adjusted squared	Mean squared error	Akaike criterion	Schwarz criterion
S&P 500	0.92	0.92	$1.0 * 10^{-5}$	-33120.54	-33095.55
FTSE 100	0.93	0.93	$0.7 * 10^{-5}$	-34836.68	-34811.67
Nikkei 225	0.92	0.92	$0.9 * 10^{-5}$	-32377.46	-32352.60
DAX	0.93	0.93	$1.3 * 10^{-5}$	-32600.69	-32575.64
Hang Seng	0.92	0.92	$0.7 * 10^{-5}$	-31615.67	-31591.03

Table 7: Forecast accuracy statistic for HAR model estimated using OLS

It is observed that R-squared coefficient and adjusted R-squared coefficient for all five stock market indexes are higher than 0.92. This result means that the data (daily realized volatility of S&P 500, FTSE 100, Nikkei 225, DAX and Hang Seng stock market indexes) fit statistical model well.

The mean squared error for all five stock market indexes is smaller than $1.3 * 10^{-5}$. The smallest mean squared error is observed in FTSE 100 and Hang Seng stock exchange indexes and it is equal to $0.7 * 10^{-5}$. This result provides strong evidence of the high forecast accuracy of the HAR model, which was estimated using OLS.

High forecast accuracy could also be observed in Figure 5, where daily realized volatility is compared with daily realized volatility forecasted by HAR model, which was estimated using OLS.

As the additional measure of forecast accuracy for HAR model estimated using OLS, the Minzer-Zarnowitz regression test was performed. The results of this regression test could be seen in Table 8.

For all five stock market indexes in the Minzer-Zarnowitz regression test intercept is almost zero, whereas the coefficient near forecast daily realized volatility values is almost one with p-values of t-test which showing high significance. The R-squared coefficient in this test is higher than 0.63 for all five stock market indexes, which were analyzed in this research. These results are sufficient for confirmation of unbiased forecast of HAR model estimating using OLS for S&P 500, FTSE 100, Nikkei 225, DAX and Hang Seng stock market indexes. This coincides with the results about high forecast accuracy which were obtained earlier.

	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng
Intercept coefficient	0.000	0.000	0.001	0.000	0.001
Intercept p-value	0.000	0.000	0.000	0.000	0.000
RV forecast coefficient	0.966	0.967	0.946	0.967	0.950
RV forecast p-value	0.000	0.000	0.000	0.000	0.000
R-squared	0.730	0.739	0.635	0.748	0.638
F-statistic p-value	0.000	0.000	0.000	0.000	0.000

Table 8: Minzer-Zarnowitz regression test for HAR model estimated using OLS

Summarizing all obtained results for all five stock market indexes analyzed in this research the following conclusion could be provided. HAR model estimated using OLS is a model which data fits well. It is also a model, which forecasts with high accuracy the daily realized volatility of S&P 500, FTSE 100, Nikkei 225, DAX and Hang Seng stock market indexes. This also confirms the results about appropriateness of HAR model estimated using OLS for daily realized volatility presented by Corsi (2009).

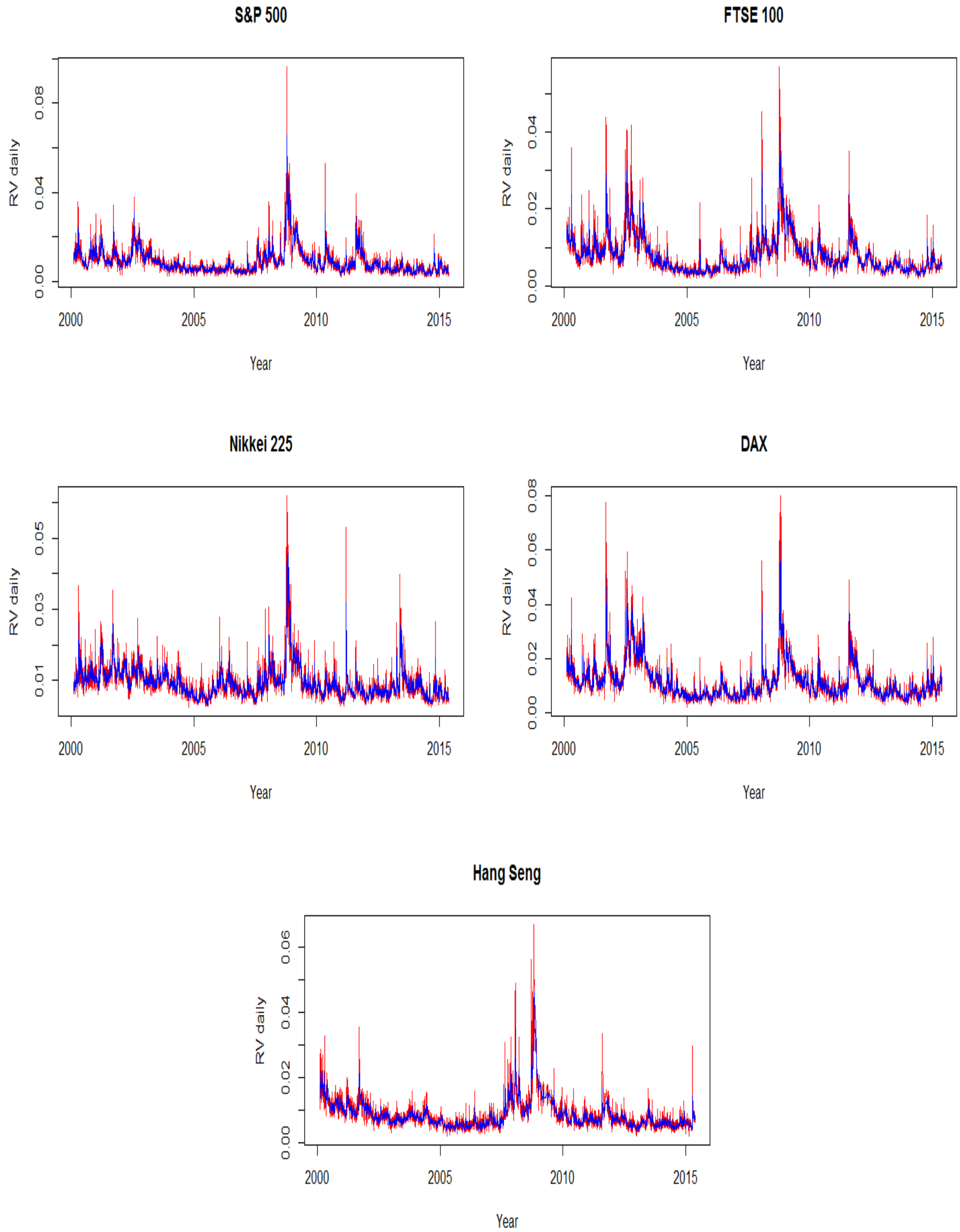


Figure 5: Comparison of daily realized volatility with value forecasted by HAR model using OLS

For HAR model estimating, using MLE based on penalized mixture distribution for analysis quality of the model and for forecast accuracy, the following indicators were estimated: R-squared coefficient, adjusted R-squared coefficient, mean squared error. The results are presented in Table 9.

	Mean squared error	R-squared	Adjusted R-squared
S&P 500	$1.0 * 10^{-5}$	0.73	0.73
FTSE 100	$0.7 * 10^{-5}$	0.74	0.74
Nikkei 225	$0.9 * 10^{-5}$	0.63	0.63
DAX	$1.3 * 10^{-5}$	0.74	0.74
Hang Seng	$0.7 * 10^{-5}$	0.63	0.63

Table 9: Forecast accuracy statistic for HAR model estimated using MLE based on penalized mixture distribution

The R-squared coefficient and adjusted R-squared coefficient are higher than 0.7 for S&P 500, FTSE 100, and DAX stock market indexes. The R-squared coefficient and adjusted R-squared coefficient for Nikkei 225 and Hang Seng stock market indexes is approximately 0.63. This mean that HAR model estimated using MLE based on penalized mixture distributions fits data quite well. However HAR model estimated using MLE based on penalized mixture distribution in comparison with HAR models estimated using OLS, show lower R-squared coefficient and adjusted R-squared coefficient. This means that the data analyzed in this research fit better the HAR model, which was estimated using OLS than HAR model, which was estimated using MLE based on penalized mixture distributions.

The mean squared error for all five stock market indexes analyzed in this work is less than $1.3 * 10^{-5}$. Therefore, high forecast accuracy could be concluded for HAR model estimated using MLE based on penalized mixture distributions. Moreover it could also be observed, that there is almost no difference in the mean squared error for all stock market indexes, which were analyzed in this research by HAR model estimated using OLS in comparison with HAR model estimated using MLE based on penalized mixture distributions. This means that both estimators provided in this research show approximately the same high forecast accuracy for all five stock market indexes.

High forecast accuracy could also be observed in Figure 6, where daily realized volatility is

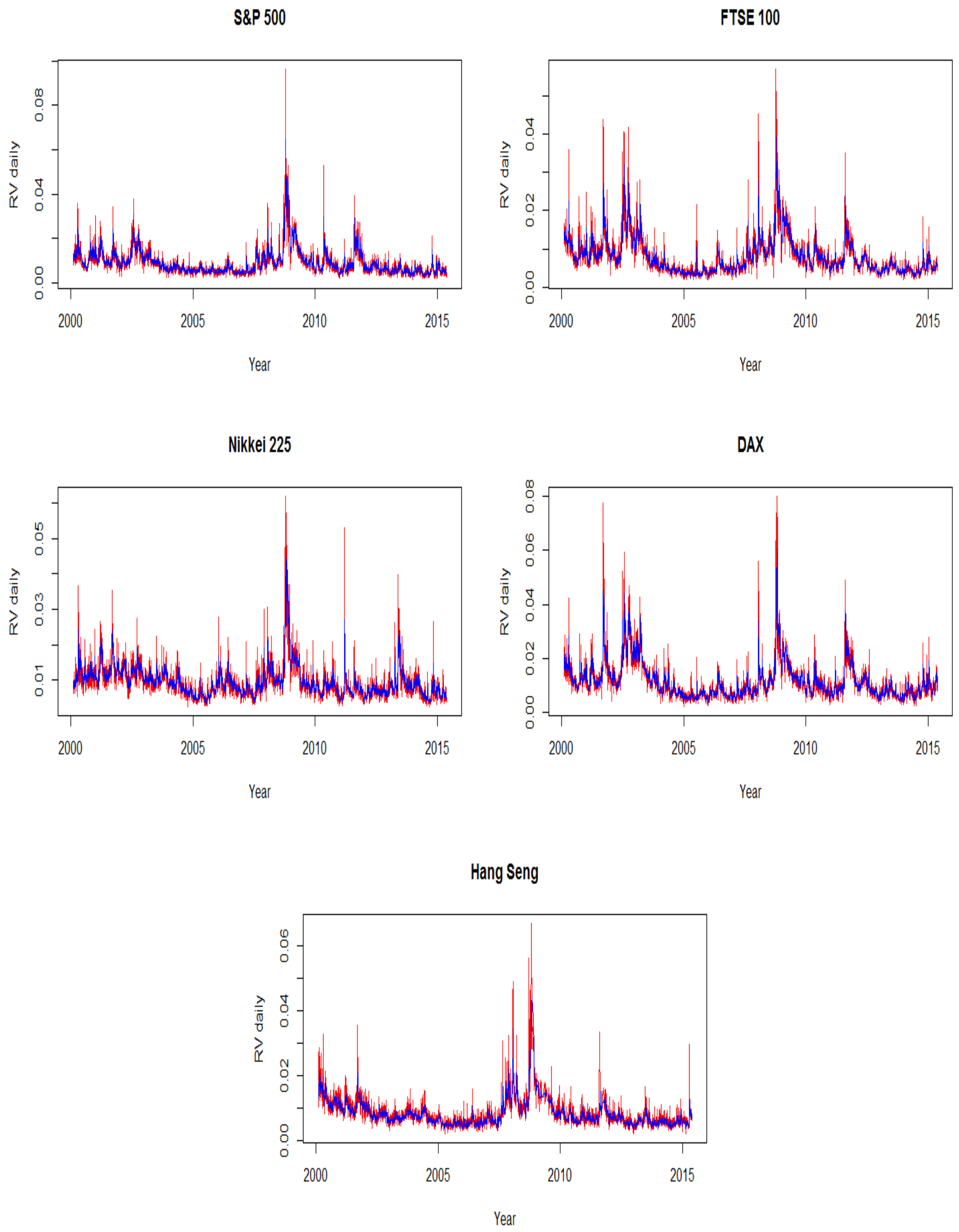


Figure 6: Comparison of daily realized volatility with value forecasted by HAR model using MLE based on penalized mixture distribution

compared with daily realized volatility forecasted by HAR model, which was estimated using MLE based on penalized mixture distribution.

The Minzer-Zarnowitz regression test was also provided as forecast accuracy measure for HAR model estimated using MLE based on penalized mixture distribution. The results are presented in Table 10.

	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng
Intercept coefficient	0.000	0.000	0.000	0.000	0.001
Intercept p-value	0.001	0.001	0.003	0.000	0.000
RV forecast coefficient	0.953	0.966	0.956	0.952	0.96
RV forecast p-value	0.000	0.000	0.000	0.000	0.000
R-squared	0.729	0.737	0.628	0.742	0.636
F-statistic p-value	0.000	0.000	0.000	0.000	0.000

Table 10: Minzer-Zarnowitz regression test for HAR model estimated using MLE based on penalized mixture distribution

The results of the Minzer-Zarnowitz test show, that for all stock market indexes analyzed in this research intercept is zero, and coefficient near forecast values of daily realized volatility is near one with p-value of t-test, which show high significance. The R-squared coefficient in this test is higher than 0.62 for all stock market indexes, which were analyzed in this research. The results of this test constitute a sufficient basis to say that the HAR model provide unbiased forecast for S&P 500, FTSE 100, Nikkei 225, DAX and Hang Seng stock market indexes. The coefficients of Minzer-Zarnowitz regression test for HAR model estimated using MLE based on penalized mixture distribution are almost the same as results of the same test for HAR model estimated using OLS. This also confirmed that there is no difference in forecast accuracy between HAR model estimated using two different approaches for stock indexes analyzed in this research.

After analysis of all coefficients, tests, measures and plots one can conclude that both estimators OLS and MLE based on penalized mixture distribution are appropriate for estimation of HAR model. However, conclusion that MLE based on penalized mixture distribution for HAR model is better than OLS couldn't be made.

4.3.2 Forecast accuracy of Lasso for realized volatility

The same measures, tests, coefficients and plots, as those applied for the HAR model, were also applied for Lasso to show model quality and forecast accuracy. These are: R-squared coefficient, adjusted R-squared coefficient, mean squared error, Minzer-Zarnowitz regression test, and plots, which compare real values of daily realized volatility with daily realized volatility values, which were forecasted using this Lasso. Applying the same measures, test, coefficients and plot will make comparison of results more precise and easier in interpretations.

The mean squared error, R-squared coefficient, and adjusted R-squared coefficient for Lasso, which was estimated using penalized MLE based on normal distribution, are provided in Table 11.

	Mean squared error	R-squared	Adjusted R-squared
S&P 500	$1.0 * 10^{-5}$	0.74	0.73
FTSE 100	$0.7 * 10^{-5}$	0.75	0.74
Nikkei 225	$0.9 * 10^{-5}$	0.64	0.63
DAX	$1.3 * 10^{-5}$	0.76	0.75
Hang Seng	$0.7 * 10^{-5}$	0.65	0.64

Table 11: Forecast accuracy statistic for Lasso model estimated using MLE based on normal distribution

The R-squared coefficient and the adjusted R-squared coefficient is between 0.73 and 0.76 for S&P 500, FTSE100 and DAX stock market indexes. For Nikkei 225 stock market index R-squared coefficient is almost 0.64, and for Hang Seng stock market index, R-squared coefficient is around 0.65. The R-squared coefficient and the adjusted R-squared coefficient for Lasso estimated using penalized MLE based on normal distribution in comparison with the HAR model estimated using two different estimators is lower. Therefore, it can be clearly concluded that data fit Lasso worse than the HAR model estimated using OLS and using MLE based on penalized mixture distribution.

The mean squared error is less or equal the $1.0 * 10^{-5}$ for four of the five stock market indexes analyzed in this research (S&P 500, FTSE100, Nikkei 225 and Hang Seng). For DAX stock market index mean squared error is slightly higher than $1.0 * 10^{-5}$ and it is equal to

$1.3*10^{-5}$. The mean squared error of Lasso estimated using penalized MLE based on normal distribution is approximately equal to mean squared error of HAR model estimated using OLS and MLE based on penalized mixture distributions for all five stock market indexes, which were analyzed in this research. This result could lead to the conclusion that forecast accuracy is high in Lasso estimated using penalized MLE and coincided with forecast accuracy in HAR model estimated using two different estimators. This result coincide with result reported by Audrino and Knaus (2012).

The forecast accuracy of this model is also shown in Figure 7, where the daily realized volatility is compared with the realized volatility forecasted by Lasso, estimated using penalized MLE based on normal distribution.

Also Minzer-Zarnowitz regression test was applied for Lasso estimated using penalized MLE based on normal distribution to check forecast accuracy. The results are , which were received, presented in Table 12.

	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng
Intercept coefficient	0.000	0.000	0.000	0.000	0.000
Intercept p-value	0.253	0.310	0.091	0.190	0.048
RV forecast coefficient	1.013	1.011	1.023	1.014	1.027
RV forecast p-value	0.000	0.000	0.000	0.000	0.000
R-squared	0.737	0.746	0.640	0.756	0.654
F-statistic p-value	0.000	0.000	0.000	0.000	0.000

Table 12: Minzer-Zarnowitz regression test for Lasso using MLE based on normal distribution

The intercept in the Minzer-Zarnowitz regression test is almost zero for S&P 500 stock market index, FTSE 100 stock market index, Nikkei 225 stock market index, DAX stock market index and Hang Seng stock market index. Coefficient near forecast values of daily realized volatility is slightly higher than one, and highly significant according to p-value of t-test for all five stock market indexes, which were analyzed in this research. The results of this test provide evidence that the Lasso estimated using penalized MLE based on normal distribution perform unbiased forecast for S&P 500, FTSE 100, Nikkei 225, DAX and Hang Seng stock market indexes. Comparison of the results of Minzer-Zarnowitz regression test

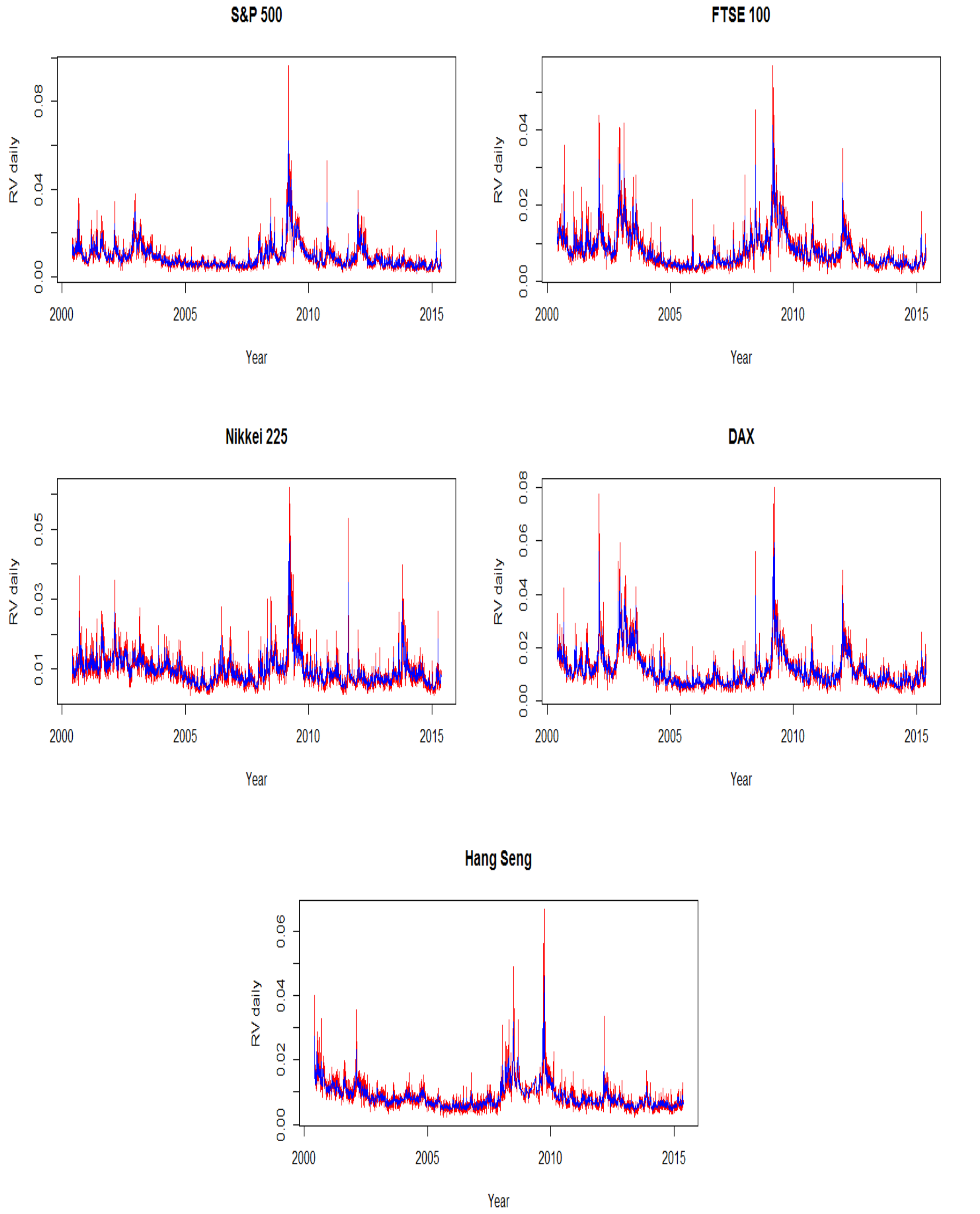


Figure 7: Comparison of daily realized volatility with value forecasted by Lasso using MLE based on normal distribution

for Lasso estimated using penalized MLE based on normal distribution and for HAR model, which was estimated using OLS and using MLE based on penalized mixture distribution, provide strong confirmation, that HAR model estimated using two different estimators provide approximately similar forecast accuracy.

R-squared coefficient, adjusted R-squared coefficient and mean squared error were calculated for Lasso estimated using MLE based on penalized mixture distribution. The results are provided in Table 13.

	Mean squared error	R-squared	Adjusted R-squared
S&P 500	$1.5 * 10^{-5}$	0.60	0.60
FTSE 100	$0.8 * 10^{-5}$	0.71	0.71
Nikkei 225	$1.2 * 10^{-5}$	0.52	0.52
DAX	$1.8 * 10^{-5}$	0.65	0.65
Hang Seng	$0.8 * 10^{-5}$	0.65	0.63

Table 13: Forecast accuracy statistic for Lasso model estimated using MLE based on penalized mixture distribution

The R-squared coefficient is 0.7 for FTSE 100 stock market index, 0.65 for DAX stock market index, 0.6 for S&P 500 and Hang Seng stock market indexes, 0.52 for Nikkei 225 stock market index. R-squared coefficient and adjusted R-squared coefficient for all stock market indexes analyzed in this research is the lower in comparison with these coefficients for HAR model estimated using OLS and using MLE based on penalized mixture distributions, and for Lasso estimated using penalized MLE based on normal distribution. This mean that daily realized volatility of all five stock market indexes fit Lasso estimated using MLE based on penalized mixture distributions worse than this data fit all other models, which were analyzed in this research.

The mean squared error in the lasso estimated using MLE based on penalized mixture distributions is higher than $1.5*10^{-5}$ for S&P 500, Nikkei 225 and DAX stock market indexes. The mean squared error for FTSE 100 stock market index and Hang Seng stock market index is approximately equal to $0.8 * 10^{-5}$. This result is higher for all stock market indexes analyzed in this research than in the HAR model estimated using OLS and MLE based on penalized mixture distributions and Lasso estimated using penalized MLE based on normal distribution. Therefore, forecast accuracy of Lasso estimated using MLE based on penalized mixture distributions is definitely lower.

Forecast accuracy could be observed in Figure 8, where daily realized volatility is compared with realized volatility forecasted by Lasso, estimated using MLE based on penalized mixture distribution.

Minzer-Zarnowitz regression test was provided as forecast accuracy measure for Lasso estimated using MLE based on penalized mixture distribution. The results could be observed in Table 14.

	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng
Intercept coefficient	0.001	0.001	-0.001	0.001	-0.002
Intercept p-value	0	0	0	0	0
RV forecast coefficient	1.157	0.854	0.924	1.146	1.203
RV forecast p-value	0	0	0	0	0
R-squared	0.723	0.734	0.611	0.746	0.63
F-statistic p-value	0	0	0	0	0

Table 14: Minzer-Zarnowitz regression test for Lasso using MLE based on penalized mixture distribution

The intercept in Minzer-Zarnowitz regression test is equal to 0.001 for S&P 500, FTSE 100, and DAX stock market index and is equal to -0.001 and -0.002 for Nikkei 255 and Hang Seng stock market indexes respectively. Coefficient near forecast values of daily realized volatility is around one for all five stock market indexes, which were analyzed in this research. For unbiased forecast intercept should be zero and forecast values of daily realized volatility should be approximately equal to one. The result of Minzer-Zarnowitz regression test for Lasso estimated using MLE based on penalized mixture distributions is worse than the results of this test for all model, which were described earlier.

It could be observed that Lasso estimated using penalized mixture provided less accurately forecast than Lasso estimated using MLE based on penalized mixture distributions. This result could be explained by the complexity of estimation which was performed in two separate steps.

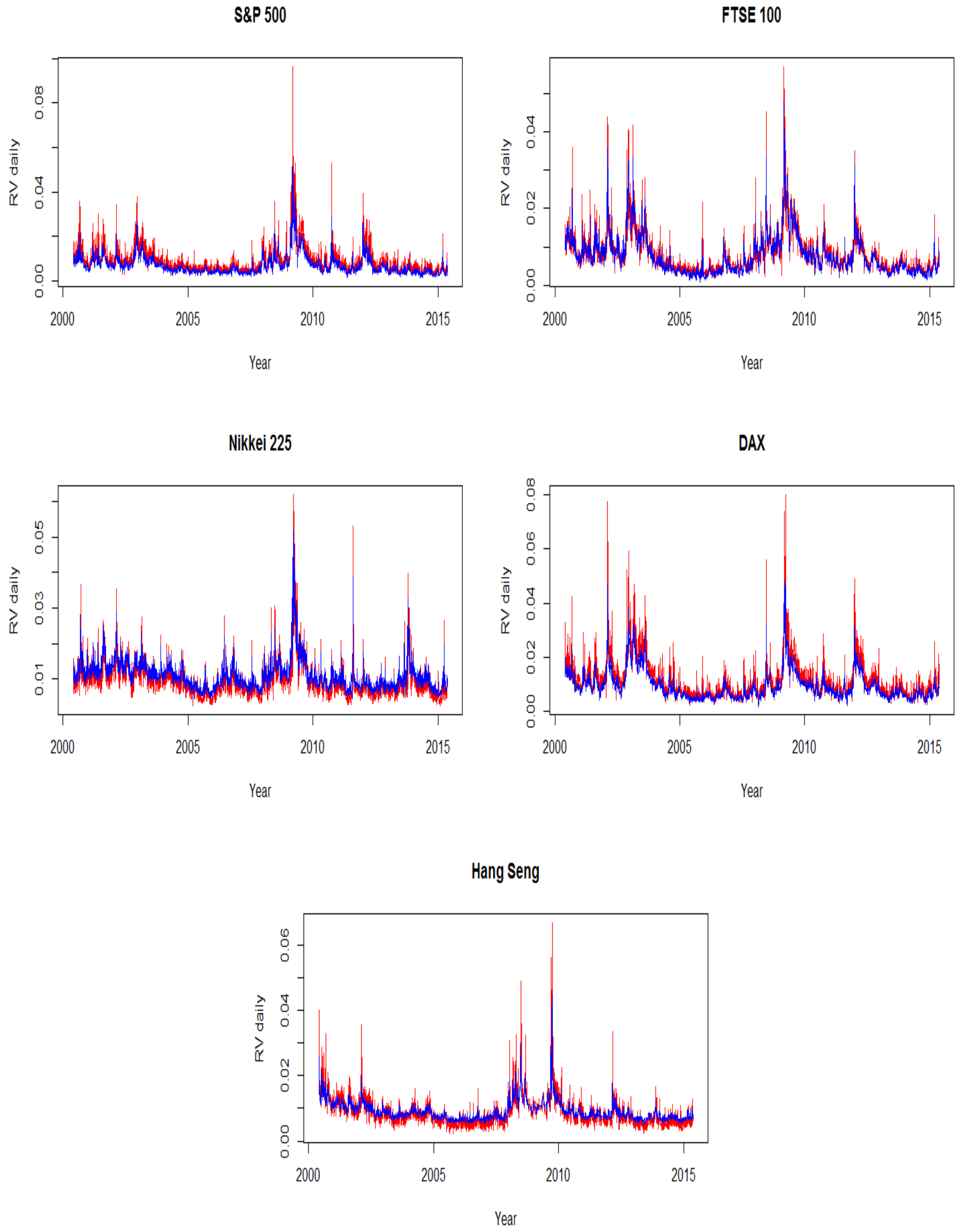


Figure 8: Comparison of daily realized volatility with value forecasted by Lasso using MLE based on penalized mixture distribution

4.3.3 Models comparison

As a summary of the results presented in this subsection, comparison of the models analyzed earlier will be performed. This final overview of the HAR model and Lasso estimated using different estimators will allow to have a synthesis of their relative merits. As a result, it will provide a better understanding of which model is better to use in estimating the daily realized volatility of stock market indexes.

In order to compare the forecast power of different models, Diebold-Mariano test proposed by Diebold and Mariano (1995) was chosen. This test performs a comparison of forecast errors of different models.

Hypothesis of this test could be described using the following formulas:

$$H_o : d_t = 0 \text{ vs. } H_1 : d_t \neq 0$$

where $d_t = g(e_{1i}) - g(e_{2i})$, with e_{1i} and e_{2i} being the forecast errors of the two models compared, g is the loss function which, usually, is the squared error loss or the absolute error loss.

The results of Diebold-Mariano test resulting from the comparison of the HAR model estimated using OLS (HAR ols), HAR model estimated using MLE based on penalized mixture distribution(HAR pen mix), Lasso estimated using penalized MLE based on normal distribution(Lasso norm) and Lasso estimated using MLE based on penalized mixture distribution(Lasso pen mix), are provided in Table 15.

	S&P 500	FTSE 100	Nikkei 225	DAX	Hang Seng
HAR ols vs. HAR pen mix	0.112	0.243	0.079	0.128	0.108
HAR ols vs. Lasso norm	0.745	0.510	0.350	0.767	0.793
HAR ols vs. Lasso pen mix	0.002	0.189	0.000	0.002	0.403
HAR pen mix vs. Lasso norm	0.706	0.471	0.542	0.568	0.726
HAR pen mix vs. Lasso pen mix	0.003	0.241	0.000	0.004	0.460
Lasso norm vs. Lasso pen mix	0.000	0.000	0.000	0.000	0.000

Table 15: Diebold-Mariano test results

It could be observed that, according to the results of Diebold-Mariano test, there is no difference between the HAR model estimated using OLS, the HAR model estimated using

MLE based on penalized mixture distribution and Lasso estimated using penalized MLE based on normal distribution for S&P 500, FTSE 100, Nikkei 225, DAX and Hang Seng stock market indexes. There is a difference between Lasso estimated using penalized MLE based on normal distribution and Lasso estimated using MLE based on penalized mixture distribution for all five indexes analyzed in this research. There is also a difference between Lasso estimated using MLE based on penalized mixture distribution and HAR model estimated using OLS, and between Lasso estimated using MLE based on penalized mixture distribution HAR model estimated using MLE based on penalized mixture distributions for S&P 500, Nikkei 225 and DAX stock market indexes. However, there is no difference between these models for FTSE 100 and Hang Seng stock market indexes.

The results of this test clearly confirm approximately the same forecast accuracy of the HAR model estimated using OLS, HAR model estimated using MLE based on penalized mixture distribution, and Lasso estimated using penalized normal distribution for all five indexes analyzed in this research. Forecast accuracy of Lasso estimated using penalized mixture distribution is different and, according to mean squared error, lower than forecast accuracy of three other models which were analyzed in this research.

It could be observed that estimation using MLE based on penalized mixture distribution doesn't improve forecast accuracy of the model. Moreover, for Lasso, the MLE based on penalized mixture distribution performs even worse than other models.

5 Conclusions

The research described in this thesis relates to the estimation and forecast of the daily realized volatility calculated employing realized kernel with Parzen weight function. Two models, used in estimation and forecast of the daily realized volatility: the HAR model and Lasso applied to the autoregressive process, were investigated. Both of these models are able to cover such important feature of the daily realized volatility as long memory dependence, which was the reason for selecting them to be the topic of this research. Classical estimators have been applied to these models: the OLS was applied to HAR model and MLE based on normal distribution was applied to Lasso. One more estimator, MLE based on the penalized mixture distribution, was also applied to HAR model and Lasso.

In the first part of the thesis, the results received by Corsi (2009) and by Audrino and Knaus (2012) were compared with the results obtained in this research. The HAR model is an additive model which consists of three partial volatilities aggregated over daily, weekly and monthly time interval. This model, estimated using OLS, yields good forecast accuracy. However, Lasso applied to autoregressive process doesn't replicate the HAR model which, therefore provides evidence that HAR is not a true data generating process for daily realized volatility. The obtained results also demonstrated that there is no difference between forecast accuracy of the HAR model estimated using OLS and Lasso estimated using penalized mixture distribution. All these results coincide with the results performed by Corsi (2009) and by Audrino and Knaus (2012).

Such an unanticipated result for Lasso could possibly be explained by the complexity of the two-step estimation procedure and accumulation of errors incurred in each of the steps. In case of the MLE based on the penalized mixture distribution one of the main challenges is the optimization algorithm, which should be used in order to maximize the maximum likelihood function with penalized parameters. The issue with optimization algorithm is that it becomes more sensitive with increasing complexity of the function that should be maximized or minimized. This is precisely the case with the MLE based on penalized mixture distribution – the maximum likelihood function contains many parameters describing different component distributions as well as parameters defining their contributions to the mixture, which significantly increases complexity of that function in comparison with OLS and MLE based on normal distribution. Therefore, it is possible that changing the optimization algorithm to a more appropriate for this problem could improve estimation results. However, this is beyond the scope of the presented research.

In summary, a general conclusion pertaining to the results of this research (and other academic pursuits) can be formulated as follows: A more complicated estimator which covers more features of the analyzed data and theoretically, should provide better estimation and more accurate forecast may fail unless a corresponding increase in the sophistication in the optimization algorithm is introduced. Without that any improvement of model and estimation may be negated by the increase of biasness and sensitivity errors in optimization. Therefore, it is always an open question how to achieve a balance between the model and the estimator which could cover all features of the analyzed data and the complexity of optimization function.

References

Ait-Sahalia, Y., P. A. Mykland, and L. Zhang (2006): "Ultra high frequency volatility estimation with dependent microstructure noise," *Journal of Econometrics*, 160, 160–175.

Andersen, T. G. and T. Bollerslev (1998): "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *International Economic Review*, 39(4), 885–905.

Arneodo, A., J. Muzy, and D. Sornette (1998): "Casual cascade in stock market from the "infrared" to the "ultraviolet"," *European Physical Journal B*, 2, 277–282.

Audrino, F. and F. Corsi (2008): "Modeling tick-by-tick realized correlation," SEPS Discussion paper series, University of St. Gallen.

Audrino, F. and S. Knaus (2012): "Lassoing the HAR model: A model selection perspective on realized volatility dynamics," SEPS Discussion paper series, University of St. Gallen.

Baillie, R., T. Bollerslev, and H. Mikkelsen (1996): "Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 74, 3–30.

Bandi, F. M. and J. R. Russell (2005): "Microstructure noise, realized volatility, and optimal sampling," Unpublished paper, Graduate School of Business, University of Chicago.

Barndorff-Nielsen, O. E., P. Hansen, L. A., and N. Shephard (2009): "Realised kernels in Practice: trades and quotes," *Econometrics Journal*, 12, 1–32.

Barndorff-Nielsen, O. E., P. H. Hansen, A. Lunde, and N. Shephard (2006): "Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise," Unpublished manuscript, Stanford University.

Bollerslev, T. (1986): "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 21, 307–328.

Candes, E. (2006): "Compressive sampling," *Proceedings of the International Congress of Mathematicians*, Madrid, Spain.

Candes, E. and T. Tao (2007): "Rejoinder: the Dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, 35, 2392–2404.

Corsi, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7(2), 174–196.

Corsi, F., N. Fusari, and D. La Vecchia (2011): "Realizing Smiles: Options Pricing with Realized Volatility," National Centre of Competence in Research Financial Valuation and Risk Management, Working paper No.678.

Corsi, F. and R. Reno (2009): "HAR volatility modelling with heterogeneous leverage and jumps," Working paper.

Diebold, F. and R. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

Donoho, D. (2004): "Compressed Sensing," Department of Statistics Stanford University, Manuscript.

Efron, B. and R. Tibshirani (1993): *An Introduction to the Bootstrap*, London: Chapman & Hall, 1 ed.

Engle, R. F. (1982): "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, 50(4), 987–1007.

Friedman, J., T. Hastie, N. Simon, and R. Tibshirani (2015): "Lasso and Elastic-Net Regularized Generalized Linear Models," Package, `https://cran.r-project.org/web/packages/glmnet/glmnet.pdf`.

Friedman, J., T. Hastie, and R. Tibshirani (2007): "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 1–10.

Mazumder, R., T. Hastie, and R. Tibshirani (2010): "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322.

McAleer, M. and M. Medeiros (2008): "Realized volatility: a review," *Econometric Review*, 27(1-3), 10–45.

Mincer, J. and V. Zarnowitz (1969): *The Evaluation of Economic Forecasts: Economic Forecasts and Expectations.*, New York: National Bureau of Economic Research.

MORGAN, J. P. (1996): "J.P. Morgan/Reuters Riskmetrics," Technical Document 4, New York: J.P. Morgan.

MUELLER, U., M. DACOROGNA, R. DAV, R. OLSEN, O. PICTET, AND J. WEIZSACKER (1997): "Volatilities of different time resolutions - analysing the dynamics of market components," *Journal of Empirical Finance*, 4, 213–239.

MUELLER, U., M. DACOROGNA, R. DAV, O. PICTET, R. OLSEN, AND J. WARD (1993): "Fractals and intrinsic time - a challenge to econometricians," *XXXIXth International AEA Conference on Real Time Econometrics*, Luxembourg.

NARDI, Y. AND A. RINALDO (2011): "Autoregressive Process Modeling via the Lasso Procedure," *Journal of Multivariate Analysis.*, 102(3), 528–549.

NELSON, D. B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59(2), 347–370.

PARK, H. AND F. SAKAORI (2013): "Lag weighted lasso for time series model," *Computational Statistics*, 28(2), 493–504.

TAYLOR, S. (1986): *Modelling Financial Time Series*, Chichester: John Wiley & Sons, 1 ed.

TIBSHIRANI, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, 58(1), 267–288.

TIBSHIRANI, R., H. HOEFLING, AND R. TIBSHIRANI (2011): "Nearly-isotonic regression," *Technometrics*, 53, 54–61.

TIBSHIRANI, R., M. SAUNDERS, S. ROSSET, J. ZHU, AND K. KNIGHT (2005): "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society Series B*, 91–108.

YUAN, M. AND Y. LIN (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B*, 68(1), 49–67.

YUAN, M. AND Y. LIN. (2007): "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94(1), 19–35.

ZAKOIAN, J.-M. (1994): "Threshold Heteroskedastic Models," *Journal of Economic Dynamics and Control*, 18, 931–955.

Zhang, L., P. A. Mykland, and Y. Ait-Sahalia (2005): "A tale of two time scales: determining integrated volatility with noisy high frequency data," *Journal of the American Statistical Association*, 100, 1394–1411.

Zou, H. (2006): "The Adaptive Lasso And Its Oracle Properties," *ournal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B*, 67(2), 301–320.

Declaration of Authorship

I hereby confirm that I have authored this Master's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, August 19, 2015

Nina Grygorenko