An application of Lasso regression in Factor Augmented HAR models

Daniel Velasquez-Gaviria
Maastricht University

August 11, 2021

Abstract

Many questions have emerged in the econometric literature since the formulation of the heterogeneous autoregressive (HAR) model. Amongst them, what is the best way to estimate it? Furthermore, what variables we should include? Several approaches have arisen, for instance, decomposing the information available in datasets in common factors and adding them as explanatory variables. A second stream proposes numerous varieties of the Least Absolute Shrinkage and Selection Operator (Lasso) model for the automatic selection of the autoregressive structure of the unrestricted AR22. For such models, the selection of the sparsity parameter has a crucial role in the level of fitting and forecasting. In this paper, we aim to enter the discussion in two ways. The first one is by including the factors within the structure of the Lasso regression in the unrestricted AR22 and jointly selecting the variables to include in the autoregressive structure and the multiple common factors, leading to a parsimonious model with only the combination of variables that minimize a loss function. Second, by finding the sparsity parameter, we propose to minimize the BIC in-sample and minimize certain loss function based on the out-of-sample model's forecasts. We perform the application on 15 Realized Volatilities of World Stock Indices. Our results evidence improvement in the fitting and forecasting afforded by the inclusion of the common factors and with the Lasso approach for model selection.

Keywords: HAR models; Realized Volatility; Forecasting; Lasso, Factor Augmented.

1 Introduction

The heterogeneous autoregressive (HAR) model proposes an additive cascade structure to characterize the different time components of volatility in high-frequency data, such as daily, weekly and monthly effects, Corsi (2009). This model has proved to successfully represent the long-run persistence of Realised Volatility. The HAR model can be expressed as an AR22. Therefore, many questions have arisen in the econometric literature about the best way to estimate it and determine the optimal combination of variables for its forecasting.

To address this question, different perspectives have emerged. One of them is using principal components that account for the comovements among global stock indices. This idea stems from the initial approach of Stock and Watson (1988), on the possible influence of common factors in economic series and its subsequent formalization in Bernanke et al. (2005), with the augmented vector autoregressive (FAVAR) model. Kim and Baek (2020) exploit this idea in the context of the HAR model, adding principal components of the realized volatility of other stock indices as explanatory variables, resulting in a better fitting and a higher level of forecasting.

Another stream exploits the Lasso regression from Tibshirani (1996) to parsimoniously recover the autoregressive structure of the unconstrained HAR model, e.g. Audrino and Knaus (2016); Audrino et al. (2018); Wilms et al. (2016), among others. The theoretical argument underlying the inclusion of auto-selection of which variables to include in the model arises from the notion that not all the time horizons of the HAR model contribute to the fitting or forecasting the realized volatility.

In Lasso regression, the sparsity parameter plays a critical role in the model performance in-sample and out-of-sample. When the sparsity parameter is equal to zero, results are equivalent to OLS, whereas as the sparsity parameter increases, the less relevant parameters are shrunk to zero. For the estimation of this parameter, information criteria or cross-sectional validation processes are usually used. Cross-sectional validation cannot be done using the traditional K-Folds method proposed by Friedman et al. (2010), considering that we are working with time autoregressive data. Francesco et al. (2019), proposes a method of K-folds that does not overlap and thus selects the parameter that minimizes a loss function. We propose to minimization the BIC for the in-sample estimation and the minimization of a loss function in the out-of-sample forecasts. This loss function can be the mean absolute error (MAE) or the mean square error (MSE), among others.

Lastly, we propose to use the Lasso regression to parsimoniously select the regression structure of the unrestricted AR22 model, jointly with the number of common factors. By this, we expect the fitting and forecasting of realized volatilities improves since neither the number of factors or lags in the model are arbitrarily chosen, instead Lasso selects the ones that minimize a loss function. In the paper, we compare the in-sample and out-of-sample fitting and forecasting of the HAR, AR22, Factor Augmented HAR (FHAR), Factor Augmented AR22 (FAR22), Lasso AR22 (LAR22) and Lasso Factor Augmented (LFAR22), for the lasso we perform two ways of selecting the sparsity parameter, the traditional cross-sectional validation method (for the sake of comparison) and the second one with the minimization of a loss function.

We use 15 world stock indices 10 minutes realized volatilities. The data was taken fom the Oxford-Man Institute of Quantitative Finance website (https://realized.oxford-man.ox.ac.uk). The sample period is from January 3, 2000, to December 12, 2017. Although we have suggested that the sample may be extended to 2020, we will take this into account for a later version.

In the next section we present the methodology, then the data, next the estimation results and finally the discussion.

2 Methodology

The intraday realized volatility is calculated by

$$RV_t^{(d)} = \sum_{j=1}^{N} r_{t,j}^2 \tag{1}$$

where $r_{t,j}$ stands for the intraday 10 minutes log returns for 1 day. $t = 1, 2, ...T$.

2.1 HAR model

For the HAR model we compute the weekly and monthly effects, following Corsi (2009)

$$logRV_t^{(w)} = \frac{1}{5}\left(\sum_{p=1}^{5} logRV_{t-p}^{(d)}\right); \tag{2}$$

$$logRV_t^{(m)} = \frac{1}{22}\left(\sum_{p=1}^{22} logRV_{t-p}^{(d)}\right) \tag{3}$$

The general HAR model is given by

$$logRV_t^{(d)} = \alpha + \beta_d logRV_{t-1}^{(d)} + \beta_w logRV_{t-1}^{(w)} + \beta_m logRV_{t-1}^{(m)} + \varepsilon_t \tag{4}$$

where $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. This equation can be estimated by OLS.

2.2 AR22 model

We can express the general HAR modelas an unrestricted AR22 model, and estimate the following equation by OLS

$$logRV_t^{(d)} = \beta_0 + \sum_{p=1}^{22} \beta_p logRV_{t-p}^{(d)} + \epsilon_t \tag{5}$$

where $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. p represents the lag order. Notice that this model has 23 parameters and the general HAR only 4 parameters.

2.3 FHAR model

The idea of this model is to include information from other stock indices captured in common factors and incorporate them as explanatory variables within the AR22 estimation Kim and Baek (2020). The motivation for using this model is to control for comovements between markets, which increases the level of fitting and forecasting given the high correlation between world economies. A disadvantage of this is that it does not allow us to identify the origin of the shocks.

Matrix X_t stands for the set of information of q foreing logRVs.

$$X_t = \begin{bmatrix} logRV_{1,1} & ... & logRV_{1,q} \\ logRV_{2,1} & ... & logRV_{2,q} \\ ... & ... & ... \\ logRV_{T,1} & ... & logRV_{T,q} \end{bmatrix}$$

The Principal Components corresponding to X_t are given by,

$$F_j = \sum_{k=1}^{q} e_{jk} X_k$$

where e_j are q eigenvectors of variance covariance funcion of X_t $\Sigma = cov(X_t)$. We arbitrary use the first fourth factors as an explanatory variable in HAR. Nonetheless, we know that there are optimization criteria for selecting the number of factors, e.g., Bai and Ng (2008). Each factor F_j is a linear combination of the realized volatility from other stock indices, they are independent each other. The first factor account for as much of the variability in the data, then the second accounts for the remaining information, and so on.The FHAR model can be expressed as

$$logRV_t^{(d)} = \alpha + \beta_d logRV_{t-1}^{(d)} + \beta_w logRV_{t-1}^{(w)} + \beta_m logRV_{t-1}^{(m)} + \sum_{m=1}^{M} \omega_m F_{t-1,m} + \gamma_t \qquad (6)$$

where $\gamma_t \overset{iid}{\sim} N(0, \sigma^2)$. We estimate the FHAR by OLS.

2.4 FAR22 model

For this model we incorporate the factors in the estimation of the AR22. The FAR22 model is expressed as,

$$logRV_t^{(d)} = \zeta_0 + \sum_{p=1}^{22} \zeta_p logRV_{t-p}^{(d)} + \sum_{m=1}^{M} \omega_m F_{t-1,m} + \chi_t \qquad (7)$$

where $\chi_t \overset{iid}{\sim} N(0, \sigma^2)$. We estimate the FAR22 by OLS.

2.5 LAR22 model

Lasso performs the selection of variables from the euclidean space of each realized volatility to include in the model. The model possesses the ability to select variables and estimate in one step as follows

$$\hat{\beta}_{AR22} = \arg\max_{\beta} \left\{ \sum_{t=p+1}^{T} \left(logRV_t - \sum_{j=1}^{p} \beta_j logRV_{t-j} \right)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right\} \qquad (8)$$

The first term of the equation is a loss function between the data and the model estimate $\left(logRV_t - \sum_{j=1}^{p} \beta_j logRV_{t-j} \right)^2$. This function is usually the quadratic error. The second term of the equation is the penalty function $\lambda \sum_{j=1}^{p} \mid \beta_j \mid$ as originally proposes by Tibshirani (1996). When the parameter lambda associated with the penalty function is equal to zero, the Lasso results are equivalent to OLS. When lambda increases, the least important coefficients go to zero; hence fewer variables are included in the regression. Accordingly, we expect an improvement in the model fitting and forecasting performance.

2.6 LFAR22 model

Determining the number of common factors is not a trivial task. Some papers suggest choosing it by using information criteria, e.g., Bai and Ng (2008). Other authors recommend using the Lasso model for auto-selection, e.g., Kneip et al. (2011). In this paper, we use the second criterion. Hence, we include in the estimation the auto-selection of the factors jointly with the 22 lags autoregressive structure. Then, the LFAR22 model can expressed as

$$\hat{\beta}_{LFAR} = \arg\max_{\beta} \left\{ \sum_{t=p+1}^{T} \left(logRV_t - \sum_{j=1}^{p} \beta_j logRV_{t-j} - \sum_{m=1}^{M} \omega_m F_{t-1,m} \right)^2 + \lambda \sum_{j=1}^{p} \sum_{m=1}^{M} \mid \omega_m + \beta_j \mid \right\} \tag{9}$$

Notice that the penalty function includes jointly the coefficients of the autoregressive structure and the common factors.

2.7 Ordered Lasso

2.8 Selection of sparsity parameter

In Lasso regression, a major concern is the choice of the sparsity parameter (λ). This parameter determines the number of coefficients taken to zero. When the sparsity parameter is equal to zero, all the variables are included, thereby increasing the variance of the prediction but technically decreasing its bias. Conversely, when the sparsity parameter increases, the estimate's variance decreases as fewer parameters are included, but the bias increases, resulting in less accurate forecasts. The choice of this parameter is traditionally made using information criteria or by the cross-validation method.

2.8.1 Cross-validation

Cross-validation is usually performed through the traditional K-Folds method, Tibshirani (1996); Friedman et al. (2010). To reproduce this method, the sample is divided into K subsamples (G_k) or "folds". Audrino et al. (2018), suggests not to overlap the folds to preserve the dependent structure of some types of data, most of the papers set K=5,10. The model is estimated on K-1 folds, and the error is predicted by taking the remaining group as a basis. The process is repeated K times, where all the folds serve as the base group. At each iteration, the mean quadratic error is computed. From this, we choose the lambda that minimizes the following function,

$$\arg\min_{\lambda}\left\{\frac{1}{T}\sum_{k=1}^{K}\sum_{G_K}(logRV_t-\widehat{logRV}_t^k)^2\right\} \tag{10}$$

where $\widehat{logRV}_t^k$ is the prediction of the K^{th}-fold.

2.8.2 In-Sample

In-sample, we select the Audrino and Knaus (2016) approach, minimizing the BIC.

2.8.3 Out of sample

In the out-of-sample analysis, we propose minimizing a loss function based on the forecast; this loss function is the mean absolute error or the mean square error (or any other loss function).

For the one day forecasting, we take $S = T*(0.95)$ [1] as the training data. The forecasting is performed from $S+1$ to T in an expanding window framework. We collect the errors $e_t = \{e_{S+1}, e_{S+2}, ..., e_T\}$. We show the process in the second part of Figure 1. For the errors we perform the MAE, the MSE and the Diebold and Mariano (2002) test.

$$MAE = \frac{1}{T-S+1}\sum_{t=S+1}^{T} \mid e_t \mid ; MSE = \frac{1}{T-S+1}\sum_{t=S+1}^{T} e_t^2 \tag{11}$$

The key question is how to select the $\lambda_{minMAE,S}$? (see Figure 1)

For this purpose we take $EW = (S * 0.95)$. From $EW+1$ to S we calculate the errors of the forecasts in an expanding window framework $e_t^{EW} = \{e_{EW+1}, e_{EW+2}, ..., e_{EW+S}\}$ as we show in the first part of Figure 1. Our λ_{minMAE} is made up of $S - WE + 1$ errors, each one from a one day growing expanding window. With these errors we construct the MAE_S^{EW} as

$$MAE_S^{EW} = \frac{1}{S-WE+1}\sum_{t=WE+1}^{S} \mid e_t^{EW} \mid \tag{12}$$

Then, we minimize the MAE_S^{EW}

[1] We only forecast the 5% of the data due to our time constraints to hand in the paper, but must be increased to at least the 25%.

$$\arg\min_{\lambda}\left\{MAE_S^{EW}\right\} \tag{13}$$

Resulting in the $\lambda_{minMAE,S}$. We use this value, for the out-of-sample forecast of $\widehat{y}_{S+1}$ and the error e_{S+1}. Then we repeat the process, expanding the window until we reach T-1. The advantage of using λ_{minMAE} is that it results from the minimizing of an out-of-sample loss function.

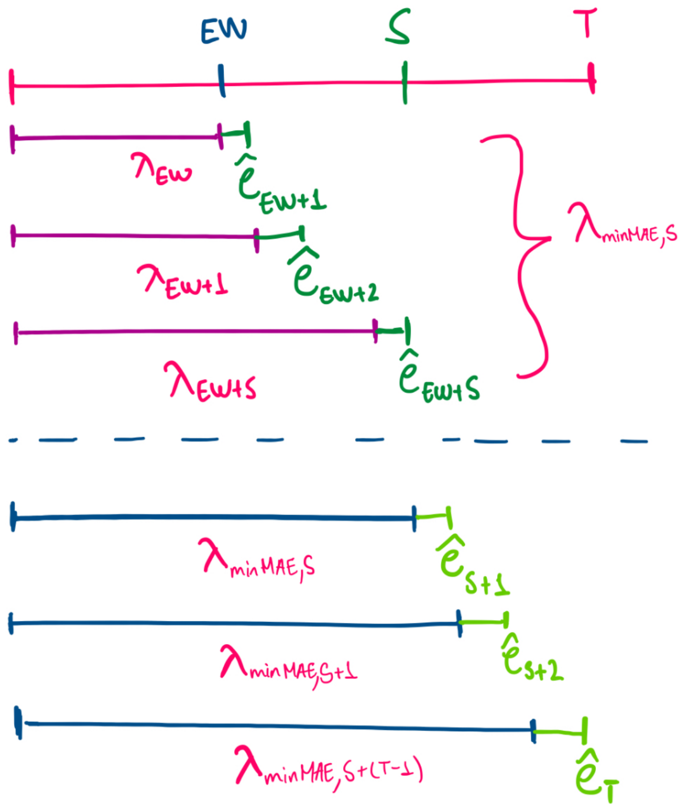

Figure 1: selection of λ

For the optimization, we use the modified limited-memory quasi-Newton method, using function values and gradients to build up a picture of the surface to be optimized, but allowing for box constraints Byrd et al. (1995); Zhu et al. (1995). We set up the constraint that lambda must be positive.

Table 1: Descriptive statistics for $LogRV_t$.

	Mean	SD	Skewness	Kurtosis	ADF
S&P500	9.818	1.168	0.286	3.262	-8.124
Dow Jones	9.851	1.148	0.332	3.397	-8.376
Nasdaq	9.602	1.131	0.414	3.041	-7.795
Russel	9.541	0.987	0.430	3.613	-8.297
DAX	9.977	1.042	0.481	3.121	-10.140
FTSE	9.290	1.062	0.292	3.179	-9.131
CAC	9.391	0.978	0.328	3.205	-10.491
AEX	9.618	1.037	0.462	3.217	-8.792
Swiss	9.940	0.939	0.840	3.835	-10.563
IBEX	9.280	0.982	0.032	2.951	-11.162
FTSE MIB	9.461	1.000	0.308	3.004	-11.567
KOSPI	9.593	1.088	0.244	2.794	-9.588
Nikkei	9.596	0.974	0.304	3.605	-9.092
IPC Mex	9.776	0.989	0.615	3.430	-9.421
Bovespa	8.704	0.854	0.408	4.195	-8.566

3 Data

We use 15 world stock indices 10 minutes realized volatilities. We use the data from the Oxford-Man Institute of Quantitative Finance website (https://realized.oxford-man.ox.ac.uk) Heber et al. (2009). The sample period is from January 3, 2000, to December 12, 2017. Although we have suggested that the sample may be extended to 2020, we will take this into account for a later version.

In table 1, the first column stands for the absolute value of the mean $logRV_t$. The second for the standard deviation—the third for the skewness, and the fourth for the Kurtosis. ADF stands for the calculated Augmented Dickey and Fuller's statistic[2]. The mean values are between 8.704 for Bovespa and 9.940 for the Swiss index. The series with the lowest standard deviation is the Bovespa index. In contrast, SP500 has the highest. All the series have positive skewness. Except for the KOSPI and IBEX indices, all series exhibit excess kurtosis.

In Figure 2 we present the $LogRV_t$ for the S&P500. Realized volatility has extreme values in the 2008 and 2012 financial crises. Nonetheless, it also presents high persistence

[2] For ADF test we select the max-lag as $\sqrt{T}$ and select the best model minimizing the BIC

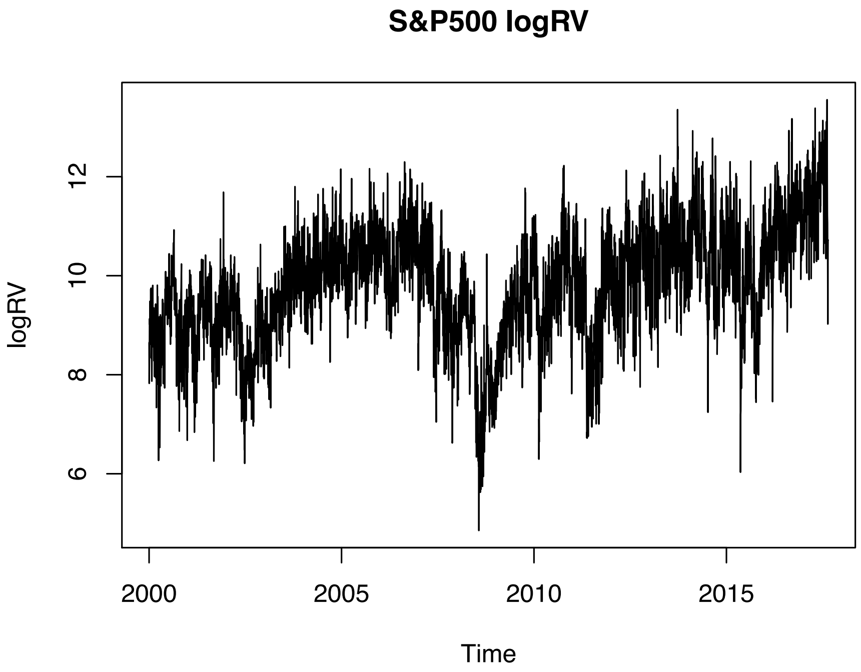

Figure 2: Caption

at the end of 2015 due to the change in oil prices.

4 Application

We first present the in-sample estimation, and second the out-of-sample.

4.1 In Sample

We present the results of the estimation in Table 2.

Table 2: In-sample estimations

	HAR			FHAR			AR(22)			FAR(22)			$LAR(22) - CV kfold$			$LAR(22) - CV_{minBIC}$			$LFAR(22) - CV_{minBIC}$		
	R^2	R^2_{adj}	BIC	R^2	R^2_{adj}	BIC	R^2	R^2_{adj}	BIC	R^2	R^2_{adj}	BIC	R^2	R^2_{adj}	BIC	R^2	R^2_{adj}	BIC	R^2	R^2_{adj}	BIC
S&P500	0.6996	0.6994	8654.29	0.7099	0.7094	8533.95	0.7042	0.7027	8745.71	0.7133	0.7116	8641.61	0.7039	0.7029	8683.65	0.7039	0.7029	8683.47	0.7122	0.7111	8574.28
Dow Jones	0.6676	0.6673	8949.61	0.6830	0.6825	8773.41	0.6725	0.6709	9042.92	0.6863	0.6845	8885.38	0.6721	0.6708	9014.20	0.6722	0.6708	9014.00	0.6853	0.6841	8824.11
Nasdaq	0.7567	0.7566	7405.04	0.7582	0.7578	7412.21	0.7592	0.7580	7519.60	0.7603	0.7589	7532.20	0.7591	0.7582	7463.06	0.7591	0.7582	7462.92	0.7597	0.7588	7467.22
Russel	0.5926	0.5923	8512.49	0.5976	0.5970	8491.39	0.5984	0.5964	8608.36	0.6027	0.6003	8595.11	0.5983	0.5967	8576.39	0.5980	0.5966	8554.41	0.5965	0.5954	8546.16
DAX	0.7523	0.7521	6777.70	0.7625	0.7621	6625.49	0.7550	0.7537	6888.94	0.7648	0.7634	6742.02	0.7545	0.7538	6805.75	0.7545	0.7539	6805.45	0.7641	0.7633	6654.30
FTSE	0.7422	0.7421	7111.58	0.7522	0.7518	6971.52	0.7447	0.7434	7228.71	0.7541	0.7527	7095.97	0.7445	0.7437	7164.27	0.7445	0.7437	7164.00	0.7540	0.7530	7022.81
CAC	0.7329	0.7328	6557.84	0.7464	0.7460	6362.24	0.7351	0.7338	6681.90	0.7481	0.7466	6492.17	0.7347	0.7340	6603.48	0.7347	0.7340	6603.39	0.7456	0.7448	6435.38
AEX	0.7548	0.7546	6699.61	0.7648	0.7644	6548.02	0.7578	0.7566	6804.14	0.7673	0.7660	6659.50	0.7574	0.7568	6718.11	0.7574	0.7568	6718.05	0.7655	0.7647	6602.78
Swiss	0.7497	0.7495	5924.98	0.7564	0.7560	5838.58	0.7528	0.7516	6029.10	0.7593	0.7578	5945.94	0.7524	0.7518	5944.58	0.7524	0.7518	5944.48	0.7574	0.7568	5853.20
IBEX	0.7362	0.7360	6542.87	0.7412	0.7408	6491.58	0.7388	0.7375	6657.58	0.7437	0.7422	6607.96	0.7384	0.7375	6606.00	0.7382	0.7374	6600.23	0.7431	0.7421	6542.16
FTSE MIB	0.7218	0.7216	6935.62	0.7280	0.7275	6869.71	0.7237	0.7223	7064.87	0.7297	0.7281	7001.55	0.7234	0.7225	7002.05	0.7234	0.7225	7001.77	0.7293	0.7282	6940.86
KOSPI	0.7704	0.7703	6771.88	0.7746	0.7742	6724.56	0.7733	0.7721	6876.74	0.7771	0.7758	6834.30	0.7732	0.7723	6835.49	0.7732	0.7723	6836.09	0.7769	0.7759	6788.03
Nikkei	0.6285	0.6283	7992.89	0.6454	0.6448	7820.70	0.6323	0.6305	8106.74	0.6487	0.6466	7939.30	0.6320	0.6307	8059.89	0.6321	0.6307	8059.60	0.6483	0.6468	7885.51
IPC Mex	0.5446	0.5443	9007.85	0.5500	0.5493	8989.19	0.5487	0.5464	9127.67	0.5538	0.5512	9110.46	0.5486	0.5467	9103.23	0.5486	0.5467	9102.98	0.5520	0.5502	9061.67
Bovespa	0.5875	0.5872	7270.99	0.5952	0.5945	7221.47	0.5911	0.5891	7391.01	0.5986	0.5963	7342.91	0.5910	0.5893	7359.51	0.5900	0.5886	7344.37	0.5973	0.5959	7273.29

Table 3: Out of sample estimations, MAE and MSE.

	HAR		FHAR		AR22		FAR22		LAR22$_{cv}$		LAR22$_{m}inBIC$		LFAR22	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
S&P500	0.6315	0.6123	0.6278	0.5966	0.6251	0.5918	**0.6181**	**0.5753**	0.6281	0.5969	0.6254	0.5918	0.6202	0.5792
Dow Jones	**0.5785**	0.5682	0.5805	0.5476	0.5788	0.5542	0.5880	**0.5408**	0.5801	0.5587	0.5792	0.5550	0.5917	0.5432
Nasdaq	0.6839	0.6902	0.6773	0.6837	0.6663	0.6486	0.6615	0.6456	0.6678	0.6497	0.6673	0.6498	**0.6610**	**0.6442**
Russel	0.6332	0.6294	0.6267	0.6203	0.6130	0.6128	0.6111	0.6042	0.6152	0.6105	0.6159	0.6133	**0.6085**	**0.5990**
DAX	0.3700	**0.2137**	0.3487	0.2147	0.3651	0.2144	**0.3483**	0.2146	0.3701	0.2157	0.3682	0.2152	0.3727	0.2254
FTSE	**0.5038**	**0.3957**	0.5484	0.4514	0.5133	0.4158	0.5564	0.4704	0.5112	0.4104	0.5114	0.4123	0.5476	0.4398
CAC	0.4056	**0.2629**	0.4235	0.2857	0.4058	0.2654	0.4252	0.2909	0.4036	0.2637	**0.4024**	0.2619	0.4152	0.2833
AEX	0.4292	0.2597	0.4365	0.2712	**0.4271**	0.2610	0.4320	0.2719	0.4288	0.2605	0.4294	0.2613	0.4376	0.2709
Swiss	0.3422	0.1928	0.3357	0.1999	0.3444	0.1930	**0.3352**	0.1999	0.3450	**0.1924**	0.3457	0.1935	0.3367	0.1985
IBEX	0.3911	0.2496	0.3934	0.2680	**0.3774**	**0.2376**	0.3808	0.2545	0.3806	0.2398	0.3796	0.2385	0.3810	0.2433
FTSE MIB	0.4462	0.3177	0.4550	0.3203	0.4391	0.3080	0.4489	0.3120	0.4380	0.3070	**0.4369**	**0.3068**	0.4413	0.3023
KOSPI	**0.3517**	0.2129	0.3763	0.2202	0.3522	0.2062	0.3705	0.2146	0.3524	**0.2057**	0.3570	0.2082	0.3593	0.2101
Nikkei	0.5454	0.4850	0.5382	0.5074	0.5426	0.4849	**0.5297**	0.5027	0.5421	**0.4809**	0.5417	0.4833	0.5321	0.5034
IPC Mex	0.4142	**0.2855**	0.4374	0.3045	0.4153	0.2908	0.4383	0.3084	0.4136	0.2895	**0.4137**	0.2899	0.4339	0.3042
Bovespa	**0.3001**	0.1588	0.3056	**0.1533**	0.3055	0.1644	0.3107	0.1588	0.3025	0.1620	0.3030	0.1625	0.3063	0.1586

Table 2 reveals that the FAR22 model has the highest R2 for all the indices. The FHAR has the highest adjusted R2 in all cases, except for FTSE, FTSE MIB, KOSPI, and Nikkei. The FHAR model has the lowest BIC for all indices, except for the Nasdaq, for which the HAR model holds the lowest BIC. The above results suggest that adding Factors significantly contributes to the in-sample fitting. However, adding a larger number of parameters penalizes the BIC and, consequently, favors the FHAR model. Furthermore, the LFHAR22 model has lower BIC compared with the FAR22. Evidencing that not all the variables of the unconstrained model contribute to the fitting, and some parameters are shrunk to zero. Such a result is proof of the usefulness of the Lasso model for auto-selecting explanatory variables.

4.2 Out of Sample

To evaluate the performance of out-of-sample predictions, we calculated the MAE and the MSE. We also performed the Diebold and Mariano (2002) test.

In the out-of-sample results in Table 3, we observe that for the SP500, the best predictive model is the FAR22. The best MAE is with the HAR for the Dow Jones, but the best MSE is with the FAR22. For the Nasdaq, the best model is the LFAR22 as well as for the Russell index. For the DAX, the best MAE is for FAR22, but the best MSE is for HAR. For the

Table 4: Two sided Diebold-Mariano Test

	S&P500					
	FHAR	AR22	FAR22	LAR22_cv	LAR22_minBIC	LFAR
HAR	0.5905	0.8406	0.7948	0.5100	0.6081	0.5839
FHAR		0.9001	0.3663	0.6989	0.6039	0.7942
AR22			0.9742	0.3388	0.9112	0.4347
FAR22				0.5429	0.3618	0.6317
LAR22_cv					0.6846	0.4996
LAR22_minBIC						0.7794
	Dow Jones					
	FHAR	AR22	FAR22	LAR22_cv	LAR22_minBIC	LFAR
HAR	0.9833	0.9146	0.8704	0.6547	0.9500	0.5377
FHAR		0.9327	0.6536	0.6138	0.5988	0.4976
AR22			0.9913	0.4471	0.9498	0.2117
FAR22				0.6630	0.7006	0.5283
LAR22_cv					0.6308	0.3510
LAR22_minBIC						0.5102

FTSE, the best model is HAR. For the CAC, the best MAE is for LAR22minBIC, but the best MSE is for HAR. For the AEX, the best MAE is for the AR22, but the best MSE is for the HAR. For the Swiss index, the best MAE is for FAR22, but the best MSE is for LAR22cv. For the IBEX, the best model is the AR22. For the FTSE MIB, the best model is the LAR22cv. For the Kospi, the best MAE is the HAR, but the best MSE is the LAR22cv. For the Nikkei, the best MAE is the FAR22, but the best MSE is the LAR22cv. For the IPC mex, the best MAE is held by the LAR22cv, and the HAR holds the best MSE. Finally, for the Bovespa index, the best MAE goes to the HAR and the best MSE to the FHAR.

According to the MAE, the models with the best predictive capability are FAR22, followed by HAR, then LAR22cv.According to the MSE, the best model is HAR, followed by LAR22cv and LFAR22.

In the Diebol-Mariano test, we compare the prediction accuracy of the models. The null hypothesis implies that the two models have the same accuracy. Conversely, the alternative hypothesis implies that the two predictions differ in the level of prediction. Since the loss function we compare is the MAE which is linear, not quadratic, we use power 1.

Due to the large size of the results, we only show the model comparison for the SP500 and the Dow Jones. As we expected, there is no difference between the two predictions. Nevertheless, this is due to the small amount of data to calculate the out-of-sample MAE. In our future work we will extend the prediction and the window for optimization.

5 Discussion

- **Conclusions**

  We took a sample of 15 realized volatilities and estimated a series of combinations of the HAR model with augmented factors and the Lasso regression. Our in-sample results determine that factor aggregation and variable auto-selection add fitting capacity. Two reasons for this are that the factors control for comovements between world stock markets. Second, the Lasso regression penalizes the less relevant coefficients for the estimation, resulting in a more parsimonious structure.

  According to the MAE, the models with the best predictive capability are FAR22, followed by HAR, then LAR22cv.According to the MSE, the best model is HAR, followed by LAR22cv and LFAR22.

  The Diebold-Marino test reveals no differences between the predictions of the models. Although this may be since we are forecasting only 50 values, we anticipate that by extending the estimation window to at least 25% of the data, the results may be more conclusive.

- **Limitations**

  Our method needs a lot of computational power. The λ_{minMAE} estimation was truncated to 50 errors. We also made 50 forecasts for each model. Otherwise, it would have been impossible to present results.

- **Future Research**

  To include other variants for the penalty function of the Lasso, such as the hierarchical Lasso and the ordered Lasso. Both have the advantage of including the own dependency structure of the realized volatility. Studies such as Wilms et al. (2016); Francesco et al. (2019) agree that the ordered Lasso improves the fitting and the forecasting performance.

We desire to investigate more ways to estimate the sparsity parameter. We anticipate that more flexible methods would lead to tighter results, especially by including a larger number of lags in the autoregressive structure.

References

Audrino, F., C. Huang, and O. Okhrin (2018). Flexible har model for realized volatility. *Studies in Nonlinear Dynamics & Econometrics 23*(3).

Audrino, F. and S. D. Knaus (2016). Lassoing the har model: A model selection perspective on realized volatility dynamics. *Econometric Reviews 35*(8-10), 1485–1521.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics 120*(1), 387–422.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing 16*(5), 1190–1208.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7*(2), 174–196.

Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & economic statistics 20*(1), 134–144.

Francesco, A., H. Chen, and O. Ostap (2019). Flexible har model for realized volatility. *Studies in Nonlinear Dynamics & Econometrics 23*(3).

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Heber, G., A. Lunde, N. Shephard, and K. Sheppard (2009). Oxford-man institute's realized library, version 0.1.

Kim, D. and C. Baek (2020). Factor-augmented har model improves realized volatility forecasting. *Applied Economics Letters 27*(12), 1002–1009.

Kneip, A., P. Sarda, et al. (2011). Factor models and variable selection in high-dimensional regression analysis. *Annals of statistics 39*(5), 2410–2447.

Stock, J. H. and M. W. Watson (1988). Testing for common trends. *Journal of the American statistical Association 83*(404), 1097–1107.

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

Wilms, I., J. Rombouts, and C. Croux (2016). Lasso-based forecast combinations for forecasting realized variances. *Available at SSRN 2873354*.

Zhu, C., R. Byrd, P. Lu, and J. Nocedal (1995). A limited memory algorithm for bound constrained optimisation. *SIAM J. Sci. Stat. Comp 16*, 1190–1208.