Politická ekonomie

Volume 75
Forthcoming articles
https://doi.org/10.18267/j.polek.1526
Open Access

VŠE / PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Enhancing the HAR model: PCA and scaled PCA methods on heterogeneous realized volatilities

Huachen Zhang, Huifang Liu, Han Yan, Shenglin Ma

Huachen Zhang, School of Finance, Capital University of Economics and Business, Beijing, China

Huifang Liu, School of Economics and Management, Shandong Youth University of Political Science, Jinan, Shandong, China

Han Yan, School of Business, Nankai University, Tianjin, China

Shenglin Ma (*corresponding author*, email: sz202209002@st.nuc.edu.cn), School of Economics and Management, North University of China, Taiyuan, China

Abstract

This study extends the heterogeneous autoregressive (HAR) model by employing the principal component analysis (PCA) and scaled PCA (sPCA) methods on lagged heterogeneous realized volatilities to construct the HAR-PCA and HAR-sPCA models using high-frequency data from 20 stock indices. The in-sample fitting and out-of-sample forecasting performances consistently show that the HAR-PCA and HAR-sPCA models have superior performance. The findings suggest that the heterogeneous principal components of lagged volatilities alleviate the subjectivity associated with the selection of the heterogeneous lag terms. This emphasizes the importance of optimizing the long-period lags under the HAR framework to better capture the heterogeneity of investors.

Keywords: Volatility, heterogenous autoregressive model, principal component analysis, scaled principal component analysis, heterogeneous market hypothesis

JEL Classification: C52, C53, C58

1. Introduction

Volatility modelling and forecasting stand as vibrant arenas within financial econometrics. Notable contributions include the works of Engle (1982), Bollerslev (1986), Hull and White (1987), Baillie *et al*. (1996), Granger and Ding (1996), Bollerslev *et al*. (2009), Corsi (2009), Buncic and Gisler (2016) among others. Volatility modelling and forecasting stand as vibrant arenas within financial econometrics. Due to its concise model form and easy parameter estimation, the heterogeneous autoregressive (HAR) model proposed by Corsi (2009) has become one of the most widely used volatility models.

Although the HAR model has gained popularity in applications, some issues need to be discussed. First, some studies challenge the lag structure of the HAR model. Audrino and Knaus (2016) and Ding *et al*. (2021) pointed out that the lag structure of the HAR model cannot always be recovered, and lags exceeding order 22 contain useful information for predicting future RV. Second, existing studies on the rationality of the lag structure of the HAR model often struggle to provide a comprehensive economic explanation. For example, Audrino *et al*. (2018) selected different lag structures for different forecast intervals. While their study showed improvements in volatility forecasting accuracy, they did not elaborate on the economic implications of these lagged volatility components. Third, many previous studies have added exogenous variables to improve the prediction effect of volatility. Though exogenous variables contain relevant factors influencing RV, the lagged volatility plays a major role in volatility modelling (Christensen *et al*., 2022). Given the long-memory property of volatility, it is necessary to incorporate numerous lagged terms in volatility modelling, and this suggests a method involving selection and dimension reduction on lagged volatilities. For the stock market index, it is economically important to monitor systematic risk by forecasting the volatility of market index, but as an average price index, it lacks the corresponding fundamental characteristics and thus, it is more convenient to utilize information from lagged volatilities than to search for limited exogenous factors when forecasting the volatility of index.

Regarding the problems existing in the above-mentioned research, we consider the heterogeneous lag terms that surpass the order of (1, 5, 22) and tend to apply dimension reduction methods to them to extend the framework of HAR. In addition, we hope to strengthen the economic interpretation compared with previous literature. We aim to obtain a flexible lag structure, which avoids subjectivity in lag structure selection. Applying dimension reduction methods to the heterogeneous lag terms offers distinct advantages over applying them to the ordinary lag terms. For one thing, incorporating the heterogeneous lag terms can effectively measure the volatility across different periods, allowing it to capture the long-mem-

ory properties of volatility. For another, after applying dimension reduction techniques such as PCA or scaled PCA (sPCA) to the heterogeneous lag terms, the resulting new variables are a recombination of different time scales, thereby avoiding the conscious selection of the (1, 5, 22) lag structure. Given the critical importance of economic explainability in our research, the linear dimension reduction methods are more suitable due to their superior explanatory power. A new research question is whether the nonlinear models, with their complex functional compositions, will offer superior predictions of volatility compared to linear models. Some scholars argued that there is no evidence that nonlinear ML models can statistically outperform linear models in general (Branco *et al*., 2024, Dudek *et al*., 2024). PCA method (Pearson, 1901), one of the linear models, obtains new variables from the linear combination of all features, so it is an interesting research perspective to analyze the weights of these linear combinations to investigate the market behavior. Huang *et al*. (2022) proposed a new supervised learning technique to improve PCA, namely sPCA, which introduces the target information based on PCA. We construct HAR-PCA and HAR-sPCA models and investigate the economic implications of the principal components extracted from these two models. Moreover, Lasso, Ridge, and Elastic Net (ENet) are widely employed linear dimension reduction techniques, from which we construct the HAR-Lasso, HAR-Ridge, and HAR-ENet models. However, their economic explanatory power is marginally weaker. We will compare the performance of all the models mentioned in the out-of-sample analysis section. For a broad assessment of model performance, we use 20 stock market indices.

The in-sample results show that the four heterogeneous principal components extracted by the PCA and sPCA methods have similar coefficient fluctuation patterns across the 20 indices. They are related to market power and heterogeneous investors, with good economic interpretability. The market power represented by the first principal component plays a persistent and moderate role in future investment activity. The subsequent three principal components represent distinct dimensions of market power associated with heterogeneous investment horizons, thereby enabling the differentiation of investor behavior. The goodness-of-fit for both the HAR-PCA and HAR-sPCA models surpasses that of the HAR model across nearly all indices.

The out-of-sample analysis results indicate superior performance of the HAR-PCA and HAR-sPCA models in comparison to other competing models such as HAR, HAR-LASSO, HAR-Enet, and HAR-Ridge models. The robustness analysis encompasses variations in the rolling window size and the maximum lag. Further analysis shows that both the HAR-PCA and HAR-sPCA models exhibit robust forecasting efficacy for horizons of one week and one month ahead. Overall, our findings suggest that, by optimizing the lag structure of volatil-

ity with PCA and sPCA, we managed to better approximate the heterogeneity of investors in the market and get an improvement in the forecasting power in many scenarios.

The contributions of this paper can be summarized as follows. First, our proposed HAR-PCA and HAR-sPCA models, which integrate PCA and sPCA methods with lagged heterogeneous volatilities for the first time, present a novel way to overcome the limitations of the traditional HAR model's restrictive lag structure. The heterogeneous principal components derived from these models provide valuable insights into capital market structure. Second, compared to the subjective determination of fixed lag structure stipulated by the HAR model, we confirm the proposition of heterogeneity of market traders from a more scientific perspective by analyzing the coefficients of four types of heterogeneous principal components extracted by PCA and sPCA methods. This is attributed to the strong economic interpretability of the models we constructed by integrating various linear dimensionality reduction methods. Besides, the selection and dimension reduction in volatility lags can improve the forecasting performance of HAR models. Third, although incorporating numerous exogenous variables may enhance volatility prediction accuracy, dimension reduction based on such variables is not a fundamental solution. This paper instead utilizes information derived directly from volatility itself. In this regard, the HAR-PCA and HAR-sPCA models improve upon the HAR model in terms of both economic relevance and predictive performance, while also providing a solid foundation for the potential future integration of exogenous variables into the framework.

2. Literature review

Over the last decade or so, the use of intraday high-frequency data for research on volatility has become a primary research direction, due to the increasing use of electronic transactions and the development of data storage technologies. Andersen and Bollerslev (1998) used high-frequency data to derive realized volatility (RV). Subsequently, RV has emerged as the most widely accepted and utilized volatility proxy in literature.

The idea proposed by Corsi (2009) has become the dominant approach to modeling and forecasting the volatility of financial asset returns. Informed by the heterogeneous market hypothesis (Müller *et al*., 1993) and the existence of asymmetric interactions across different volatility time frames (Müller *et al*., 1997), Corsi (2009) proposed the HAR model, which was subsequently applied to high-frequency data. It delineates a stratification akin to a "volatility cascade phenomenon", which triggers stratification from low to high frequencies. The model amalgamates nonparametric realized variance at daily, weekly, and monthly intervals with a parametric autoregressive model, capturing short, medium, and long-term

perspectives of market participants while reflecting the long memory of volatility. Due to its straightforward parameter estimation, concise structure, and superior out-of-sample forecasting performance, the HAR model has been widely used in the realized volatility modeling (Chen and Ghysels, 2011, Audrino and Knaus, 2016).

Currently, a prevalent approach for extending HAR involves the introduction of exogenous variables and subsequent variable selection and dimensionality reduction. For example, Chen and Ghysels (2011) added asymmetric news impact to HAR model, Christensen *et al*. (2022) introduced some firm characteristics and macroeconomic indicators, and Niu *et al*. (2022) and Guo *et al*. (2022) used dimension reduction methods to select uncertain variables and then introduced them into the HAR model, among others. In addition, some related studies have also been carried out under the AR framework (Ma *et al*., 2022). Another line of research extending the HAR model focuses on incorporating more sophisticated measures of realized volatility. Andersen *et al*. (2007) enhanced the model's predictive performance by employing realized bi-power variation to separate jump components from the smooth continuous moves. Patton and Sheppard (2015) further improved forecast accuracy by incorporating realized semi-variances of positive and negative returns. Bollerslev *et al*. (2016) introduced realized quarticity into the HARQ model, allowing for time-varying parameter optimization. It is worth noting that this study does not directly advance this line of research; however, given the foundational importance of improved realized measures for volatility forecasting, we provide a review of this strand of HAR model extensions in this context.

On the other hand, some scholars also studied the rationality of the HAR model structure. The reason is that the HAR model divides the investors into three categories according to the investment period of daily, weekly, and monthly, introducing a degree of subjectivity in the process of discerning heterogeneous investors. For example, researchers such as Audrino and Knaus (2016), Ding *et al*. (2021), and other scholars have employed a LASSO-type method within the AR or constrained AR framework to select lag structure. Their studies show that the lag structure of the HAR model cannot always be recovered, and lags beyond 22 periods also contain factors that affect future volatility, making it reasonable to consider longer lags and more flexible lag structures. Moreover, with the improvement of computer performance, some scholars have recently used machine learning methods in this field. For example, Christensen *et al*. (2022) studied volatility modelling based on machine learning methods. They concluded that the random forest and neural network models they tested had a better overall prediction effect than a broad suite of HAR models. Guo *et al*. (2022) employed an AR framework to forecast volatility in crude oil prices, complementing it with

machine learning techniques. Their research suggests that the sPCA method combined with the AR model performs best in predicting volatility.

Reviewing past literature, it is not difficult to find the following problems. First, in the current discussion on the rationality of the lag structure of HAR models, it is generally argued that more flexible lag structures and longer-term lag terms need to be considered. Second, the existing research examining the rationality of the lag structure within the HAR model often tends to devote a substantial portion of its content to tedious discussions about lag term selection and encounters difficulties in providing comprehensive economic interpretations. For example, Audrino *et al*. (2018) and Ding *et al*. (2021) found that the lag structure changes as the economic situation changes. Although their study improves the volatility forecasting effect in some cases, they did not explain the economic meaning of these lagged volatility components. Thirdly, numerous works have incorporated exogenous variables to improve forecasting performance. Existing research mainly focuses on the selection and dimension reduction of exogenous variables. Though exogenous variables contain relevant factors influencing RV, lagged volatility is also an important variable for forecasting volatility. In fact, lagged volatility plays a major role in volatility modelling. Extracting predictive signals for future volatility directly from the intrinsic information embedded in realized volatility constitutes a more fundamental and direct approach compared to incorporating external information sources. HAR is also a model built upon lagged information. Accordingly, selecting and dimensionally reducing the lagged terms of volatility itself constitutes a more fundamental approach for model improvement.

In response to the challenges identified in the preceding research, we consider a flexible choice of heterogeneous lag terms that surpass the standard order of (1, 5, 22) and apply PCA and sPCA methods to them, extending the framework of the HAR model.

3. Methodology

3.1 HAR model

Let the logarithmic price process X_t follow a continuous-time stochastic dynamics defined as:

$$dX_t = \mu_t dt + \sigma_t dW_t \tag{1}$$

where w_t is a standard Brownian motion, μ_t is the trend component, a non-random càdlàg process of finite variation, and σ_t is the time-varying volatility, also a càdlàg process, which is independent of W_t.

The integrated volatility (*IV*) over a 1-day interval [$t - 1d$, t] quantifies latent volatility and is expressed as:

$$IV_{\mathrm{t}}^{(d)} = \int_{t\text{-}1d}^{t} \sigma_s^2 ds \tag{2}$$

since *IV* is unobservable, it is empirically approximated using realized volatility (*RV*). *RV* aggregates squared intraday log-returns sampled at high frequencies within a trading day:

$$RV_t = \sum_{i=1}^{n} (X_{t-1+\frac{i}{n}} - X_{t-1+\frac{i-1}{n}})^2 \tag{3}$$

here, n denotes the number of equally spaced intraday observations. Under regularity conditions, RV_t converges in probability to $IV_t^{(d)}$ as $n \to \infty$ (Barndorff-Nielsen and Shephard, 2002).

For longer time horizons, such as weekly or monthly periods (with 5 and 22 trading days, respectively), *RV* is computed as the average of daily *RV* values over the given period:

$$\begin{aligned} RV_t^{(w)} &= \frac{1}{5}\sum_{i=0}^{4} RV_{t-id}^{(d)} \\ RV_t^{(m)} &= \frac{1}{22}\sum_{i=0}^{21} RV_{t-id}^{(d)} \end{aligned} \tag{4}$$

The standard HAR model can be derived as follows:

$$RV_{t+1d}^{(d)} = \beta_0 + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \varepsilon_{t+1d} \tag{5}$$

where $RV_t^{(d)}$, $RV_t^{(w)}$, and $RV_t^{(m)}$ represent the daily, weekly, and monthly aggregated *RV* values, respectively. ε_{t+1d} denotes a zero-mean innovation term. Estimates of coefficients, β_0, $\beta^{(d)}$, $\beta^{(w)}$ and $\beta^{(m)}$, can be consistently estimated via ordinary least squares (OLS) regression.

The HAR model simplifies the form of volatility modelling as well as the difficulty of model parameter estimation by introducing volatility on daily, weekly, and monthly time scales into the AR model. However, the shortcomings of the HAR model are that, according to the previous literature, the longer-term lag terms also contain useful information for future RV. In addition, the selection of the three volatility components is subjective, and each type of volatility component is only the average value within a certain time range in the past. Ignoring the case of unequal weights may cause the volatility component to lack sufficient explanation in the model, which could lead to inadequate consideration of additional

market volatility patterns and investors market's investment operations. Therefore, on the basis of the HAR model, this paper considers the combination of the PCA method and the recently improved sPCA method (Huang *et al*., 2022) to explain the investment behavior of heterogeneous investors as well as the corresponding heterogeneous volatility from a distinct perspective. It is expected to explain better prediction performance.

3.2 HAR-PCA model

Müller *et al*. (1997) proposed that the effect of coarse volatility in predicting fine volatility is better than that of the reverse prediction. Coarse volatility refers to volatility observed over longer time intervals in the past or a volatility level with coarser statistical granularity, representing long-term traders in the market structure. The fine volatility refers to the volatility of a shorter time interval or finer statistical granularity in the past, representing the short-term traders in the market structure.

Inspired by Müller *et al*. (1997), referring to the concept of coarse volatility they introduced, we similarly construct the heterogeneous lag term of RV that encompasses information on market trading across various time horizons. The multiperiod volatilities are normalized by time horizons, enabling a direct comparison of volatilities defined over different time horizons. We refer to this as heterogeneous lagged volatility, which captures the overall volatility level over the interval from $t - d$ $(k - 1)$ to t, and it is formally defined in equation (6). When $k = 5$ (22), Equation (6) becomes $RV_t^{(w)}$ ($RV_t^{(m)}$) in the HAR model.

$$RV_t^{(k)} = \frac{\sum_{i=0}^{k-1} RV_{t-id}}{k} \tag{6}$$

When modeling volatility using many heterogeneous lagged volatilities with different maturities, these variables need to be reduced in dimension due to multicollinearity among the independent variables for the linear regression model. The principal component analysis (PCA) (Pearson, 1901) is a dimensionality reduction method that can extract important information from high-dimensional data sets and is well suited to the research of this paper. Specifically, we applied the PCA method to the set of heterogeneous lagged volatility variables to extract principal components and replace the volatility components in the HAR model, thereby constructing the HAR-PCA model.

In order to implement the HAR-PCA model, the number of principal components, m, is a parameter that needs to be predefined before employing the PCA method to obtain new volatility components. In addition, the variable set used to extract the principal components is composed of heterogeneous volatilities with different lag lengths, which is expressed as

$\Omega = \left\{ RV_t^{(1)}, RV_t^{(2)}, \ldots, RV_t^{(k-1)}, RV_t^{(k)} \right\}$, so the maximum lag period k of Ω is also a parameter that needs to be determined in advance. Elaboration on the values of parameters m and k is provided in Section 4.

After obtaining the principal components according to the above method, the final form of the HAR-PCA model is as follows:

$$RV_{t+1d}^{(d)} = \theta_0 + \sum_{i=1}^{m} \theta_i HPC_i + \varepsilon_{t+1d} \tag{7}$$

where HPC_i represents the heterogeneous principal component obtained by the PCA method for i = 1, ..., m, and each principal component is obtained by multiplying a set of weights with the elements in the set of heterogeneous volatility variables Ω.

The benefits of the HAR-PCA model are as follows. First, when the lag period is given, the subjectivity of variable selection on the right-hand side of the HAR model equation is overcome to some extent. Based on the heterogeneous market hypothesis, the heterogeneous volatility variable set is constructed, which contains complicated market information. Employing PCA on this specific set of variables yields extracted principal components that offer a more scientific depiction of the heterogeneous market structure and trader composition. This approach avoids the subjective generalization of market participants with varying horizons.

Second, it obtains principal components based on the heterogeneous volatility variable set and then replaces the volatility components in the HAR model. This approach extends the temporal horizon based on the heterogeneous volatility variable set, providing greater flexibility in assimilating a broader spectrum of pertinent historical information. Simultaneously, this approach maintains model parsimony and ease of estimation. Third, the purpose of this study is not only to discuss the predictive performance of the model but also to improve the HAR model from an economic perspective. According to (Granger, 2008), linear models have better economic interpretability than nonlinear models. Each principal component is a linear combination of heterogeneous lagged volatilities, which implies that the HAR-PCA model remains fundamentally a linear model. Therefore, the application of the PCA method also preserves the HAR model's inherent economic interpretability. Fourth, extracting principal components from the heterogeneous lagged RV variable set by the PCA method seems to be a possible method to distinguish market structure and investors. Despite certain limitations, this method is considered more scientific than the HAR model for differentiating market structure and investor classifications.

3.3 HAR-sPCA model

The HAR-PCA model is constructed based on unsupervised learning. This model lies in its omission of the target information to be predicted during the principal component acquisition process, resulting in a lower level of information richness compared to supervised learning methodologies. The scaled principal component analysis (sPCA) method proposed by Huang *et al*. (2022) is a supervised learning method that is implemented in two steps. First, each predictor variable is scaled based on its predicted slope for the target variable. Second, PCA is applied to the scaled predictor variables to extract principal components. This method helps shift the weights of principal components toward predictor variables with stronger predictive ability. The sPCA method is also highly suitable for our research topic. We introduce this supervised learning method into the HAR-PCA model, similarly, constructing the HAR-sPCA model.

We predict the target with sPCA in two steps. First, we form a panel of scaled predictors $(\gamma_1 RV_t^{(1)}, \gamma_2 RV_t^{(2)}, \ldots, \gamma_{k-1} RV_t^{(k-1)}, \gamma_k RV_t^{(k)})$ based on the lagged heterogeneous realized volatility variable set Ω. Here, the scaled coefficient γ_l is the estimated slope from regressing the target on the *k*th (standardized) predictor:

$$RV_{t+1d}^{(d)} = v_l + \gamma_l RV_t^{(l)} + u_{l,t+1d} \qquad l = 1, \cdots, k \tag{8}$$

Second, apply PCA to $(\gamma_1 RV_t^{(1)}, \gamma_2 RV_t^{(2)}, \ldots, \gamma_{k-1} RV_t^{(k-1)}, \gamma_k RV_t^{(k)})$ to extract *m* factors and use them to construct the HAR-sPCA model. The form of the HAR-sPCA model is as follows:

$$RV_{t+1d}^{(d)} = \theta_0 + \sum_{i=1}^{m} \theta_i HsPC_i + \varepsilon_{t+1d} \tag{9}$$

where $HsPC_i$ represents the heterogeneous scaled principal component obtained by the sPCA method for $i = 1, \ldots, m$. A detailed discussion on the values of k and m will be conducted in Section 4.

The advantages of the HAR-sPCA model are as follows. First, like the HAR-PCA model, it not only avoids the subjectivity in selecting lagged terms within the HAR framework but also provides greater flexibility to extend the time range by incorporating historical information. Second, the core idea of the HAR-sPCA model is to construct a model based on supervised learning. This approach entails considering the target variable for prediction during the principal component extraction process. The goal is to adjust the weights assigned to the original variables based on the relative importance of lagged heterogeneous volatility concerning the forecasting target. This subtle adjustment generates more rational principal components. Third, as evident from the procedural steps of the sPCA method, the principal components

extracted by sPCA are essentially linear combinations of the predictor variables. Consequently, similar to the PCA method, it retains the advantage of the HAR-PCA model in terms of economic interpretation. Moreover, since the sPCA method considers the information of the forecasting target, the resulting volatility components are more closely aligned with future volatility, thereby further enhancing the model's interpretability from an economic perspective.

3.4 Alternative models

In order to investigate the performance of the proposed HAR-PCA and HAR-sPCA models, we select the classical HAR, HAR-LASSO, HAR-Ridge, and HAR-ENet models as alternative models. A detailed specification of alternative models can be found in Appendix A.

4. Data

We use 20 indices: the AEX Index (AEX), All Ordinaries (AORD), Bell 20 Index (BFX), Bovespa Index (BVSP), Dow Jones Industrial Average (DJI), CAC 40 (FCHI), FTSE 100 (FTSE), DAX (GDAXI), HANG SENG Index (HSI), IBEX 35 Index (IBEX), Nasdaq 100 (IXIC), Korea Composite Stock Price Index (KS11), IPC Mexico (MXX), Nikkei 225 (N225), NIFTY 50 (NSEI), Russel 2000 (RUT), S&P 500 Index (SPX), Swiss Stock Market Index (SSMI), FT Straits Times Index (STI), Euro STOXX 50 (STOXX50E) for the period from January 3, 2000, to January 24, 2020. The daily 5-minute realized variance of these indices are provided at the Oxford-Man Institute of Quantitative Finance – Realized Library.[1]

Müller *et al*. (1997) proposed that lagged correlation is a powerful tool for studying the relationship between two time series. In order to explore the correlation between the RV series and its lagged series, we examine the ordinary autocorrelation function between them, which is defined as corr(RV_t, RV_{t-k}), and the results are plotted in Figure 1.

Figure 1 portrays the ordinary autocorrelation function diagram for 20 indices, considering the 252-day (equivalent to one year) lag for RV. It reveals two facts. First, over the 252-day lag, almost all the indices demonstrate significant autocorrelation, establishing a foundational basis for forecasting future RV through the integration of lagged heterogeneous RV. Second, within the lag range of 60–70, the ordinary autocorrelation coefficients for all indices are significantly different from zero. From an industry perspective, numerous economic indicators are routinely disclosed on a monthly or quarterly basis, encompassing corporate financial statements, macroeconomic data, and analogous sources. Furthermore,

1 https://oxford-man.ox.ac.uk/research/realized-library/

the real-life investment portfolios frequently undergo adjustments at monthly or quarterly intervals. Therefore, in the context of forecasting volatility, this study focuses on two specific maximum lag periods for heterogeneous RV, specifically lags of 22 and 66 periods. Consequently, the parameter k is set at 22 and 66. Since the HAR model discusses volatility modelling based on the 22 lag periods, we focus on discussing the case with $k = 66$ in the empirical analysis. The case of $k = 22$ is discussed in the robustness analysis presented in Section 6.

Figure 1: Autocorrelation function plot of RV

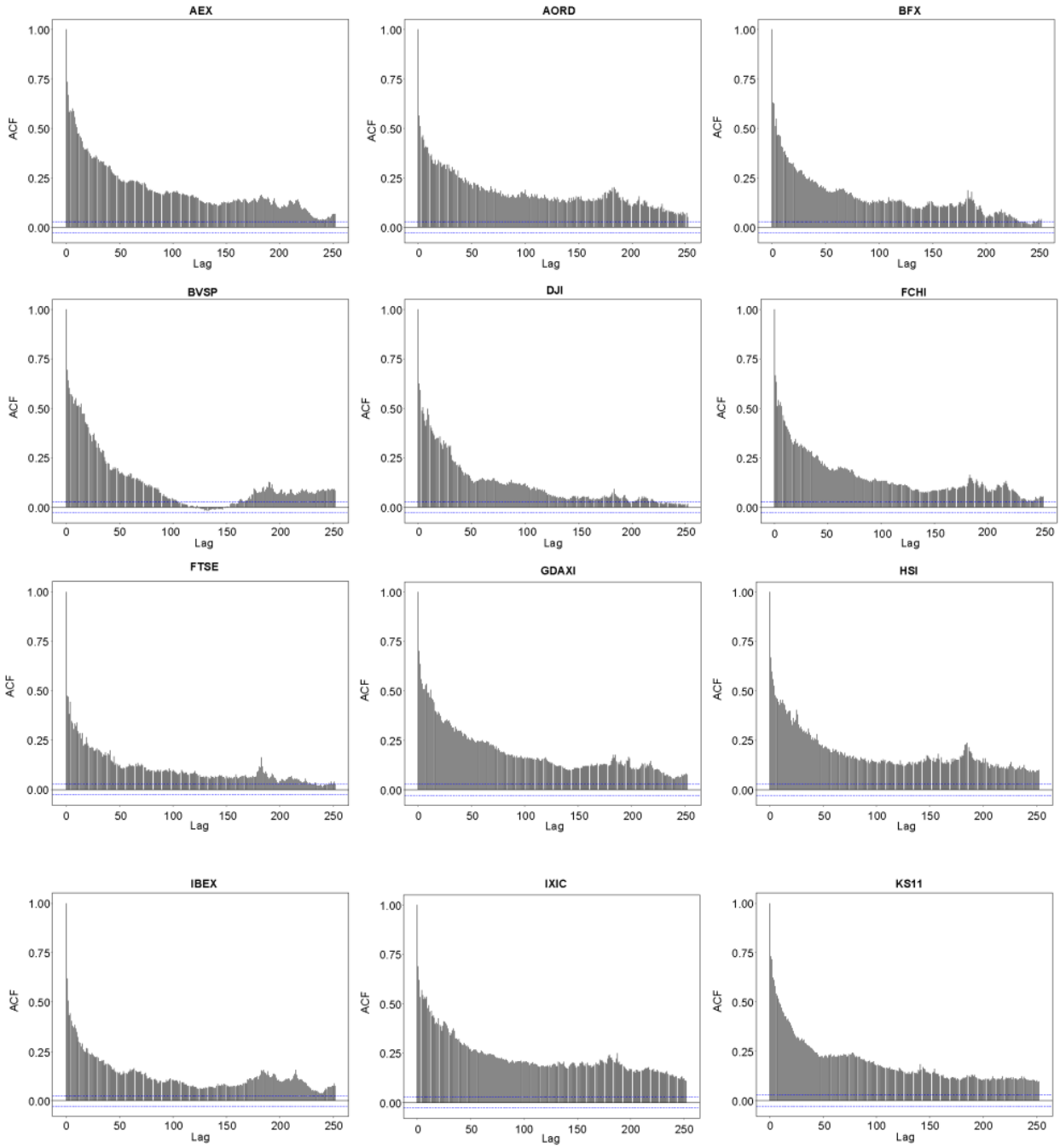
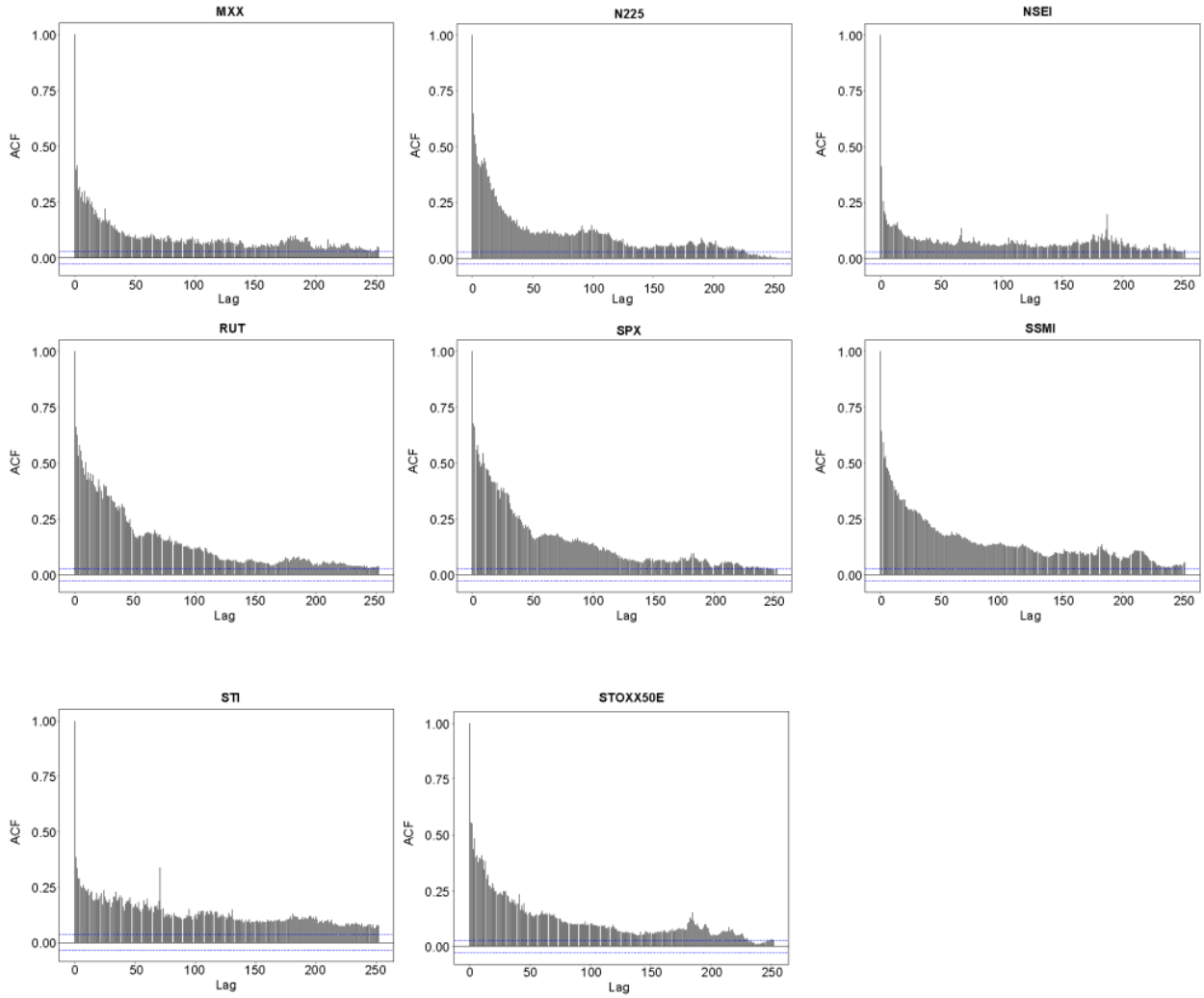


Figure 1: Continuation



Note: The lag period is 252 days.
Source: Authors' own elaboration

Subsequently, we proceed to determine the value of m, the number of principal components. This value is a crucial determinant that significantly impacts the economic implications associated with the HAR-PCA and HAR-sPCA models. Two requirements must be met for parameter *m*. First, it should satisfy the PCA method's stipulation of a cumulative variance contribution rate exceeding 85%. Second, it should provide a rational explanation for the heterogeneous market structure and divergent market traders. In this regard, we refer to the original intention of the HAR model. From the perspective of simplifying the model, they considered three volatility components, which represent three types of investors with different trading frequencies in the market. From the perspectives of economic

interpretability and explanatory power, we aim to make a more purer comparison between the principal components extracted using PCA and sPCA methods and the three volatility components on the right side of the HAR model in terms of improvements in the economic significance of variables (Granger, 2008), thereby enhancing the model's explanatory power (Drerup *et al*., 2017), rather than simply increasing the number of principal components to improve the R^2 value of the fitted equation. Therefore, we also aim to select three principal components to represent investors with different trading frequencies.

Figure 2 and Figure 3 show the average values of the weights of the top four principal components extracted from 20 indices using the PCA and sPCA methods with a fixed window rolling prediction, respectively. Two figures show that the weighting coefficients of the four principal components extracted by the two methods exhibit fundamentally similar trends. Secondly, heterogeneous RV represents the risk level of the stock market during a specific time period. Therefore, the principal components extracted based on it also measure the risk level of a certain type of stock market and reveal the composition of the corresponding market participants. The consistent trend in the principal component weight coefficients across all subgraphs allows for an overall interpretation of the economic significance conveyed in Figure 2 and Figure 3. Given that RV signifies the volatility of investors' trading activities and assumes positive values, a higher RV denotes increased trading frequency among investors. When the principal component has a positive weight on the lagged heterogeneous RV, it means that investors trade in this period, while a negative weight means that investors do not trade. Therefore, we can explain the principal components from the perspective of the trading horizon.

Next, we turn to the meaning of the principal components extracted by the PCA and sPCA methods. It can be observed that the weighting coefficients of the first principal component[2] are positive at each lag period and have similar values, appearing as a straight line in the subgraph. This means that the market power represented by the first principal component plays a continuous and moderate role in the future trading activities of investors, which we summarize as the main effect. The remaining three principal components exhibit an alternating pattern of change. Although they have differentiated weight coefficient changes, they all maintain a large positive weight within the 0–5 days lag period. This indicates that, overall, regardless of the trading frequency of investors, their very short-term trading behavior will have a significant impact on future trading activities. Specifically, the second principal component has a positive weight within a lag period of approximately 0–20 days.

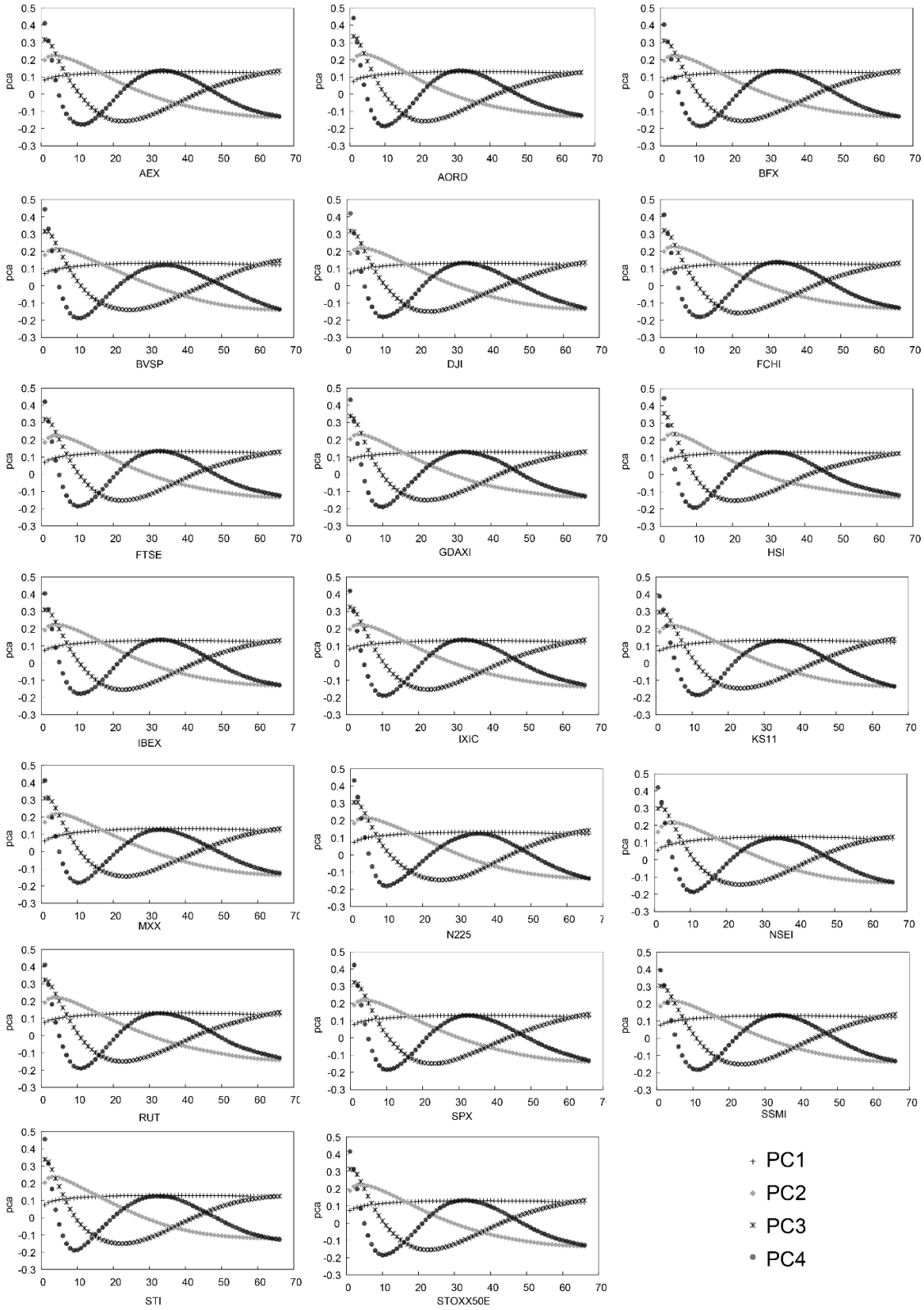
2 For example, "first principal component" denotes the set of first principal components described in the subgraph. The other three principal components have the similar meaning.

This means that the market forces represented by the second principal component will trade at that frequency. Given the maximum time horizon, we express the second principal component as short-term investors within that time horizon. The third principal component has positive weights within a lag period of approximately 40–66 days. This indicates that the market power represented by the third principal component is associated with a longer investment horizon.

Similar to the interpretation of the economic significance of the second principal component, we regard the third principal component as long-term investors within a given maximum time horizon. The fourth principal component has positive weights within a lag period of approximately 20–40 days. This means that the market power represented by the fourth principal component is related to this trading frequency. Similar to before, given the maximum time horizon, we regard the fourth principal component as a medium-term investor within that time horizon. In summary, the first principal component exerts a relatively uniform influence across all lag periods. By contrast, the second, third, and fourth principal components exhibit distinct alternating patterns of variation. These components can be interpreted as capturing the behavior of three distinct investor groups characterized by heterogeneous trading frequencies, each contributing unique dynamics that differentially affect future trading activities. Accordingly, we selected four principal components, which encompassed three types of investors with distinct trading characteristics, and associated them with the explanatory variables on the right side of the HAR model. When m is set to 4, Figure 4 shows that the principal components extracted from both models are almost entirely statistically significant, coupled with a cumulative variance contribution rate exceeding 85%. Therefore, m is conclusively established as 4 to optimally meet these criteria.

The above analysis highlights a common feature among the 20 stock market indices, namely the existence of short-term, medium-term, and long-term trading behavior characteristics within a given maximum time horizon. Both the PCA and sPCA methods outline a consistent principal component structure for lagged heterogeneous RV, facilitating a straightforward and scientifically grounded interpretation. In terms of economic significance, our interpretation of the core concepts of the HAR model provides a fresh perspective, thereby supplementing the heterogeneous market hypothesis theory. Our approach introduces main, short-term, medium-term, and long-term effects, thus enriching the internal structure of the cascade model while containing the original daily, weekly, and monthly effects.

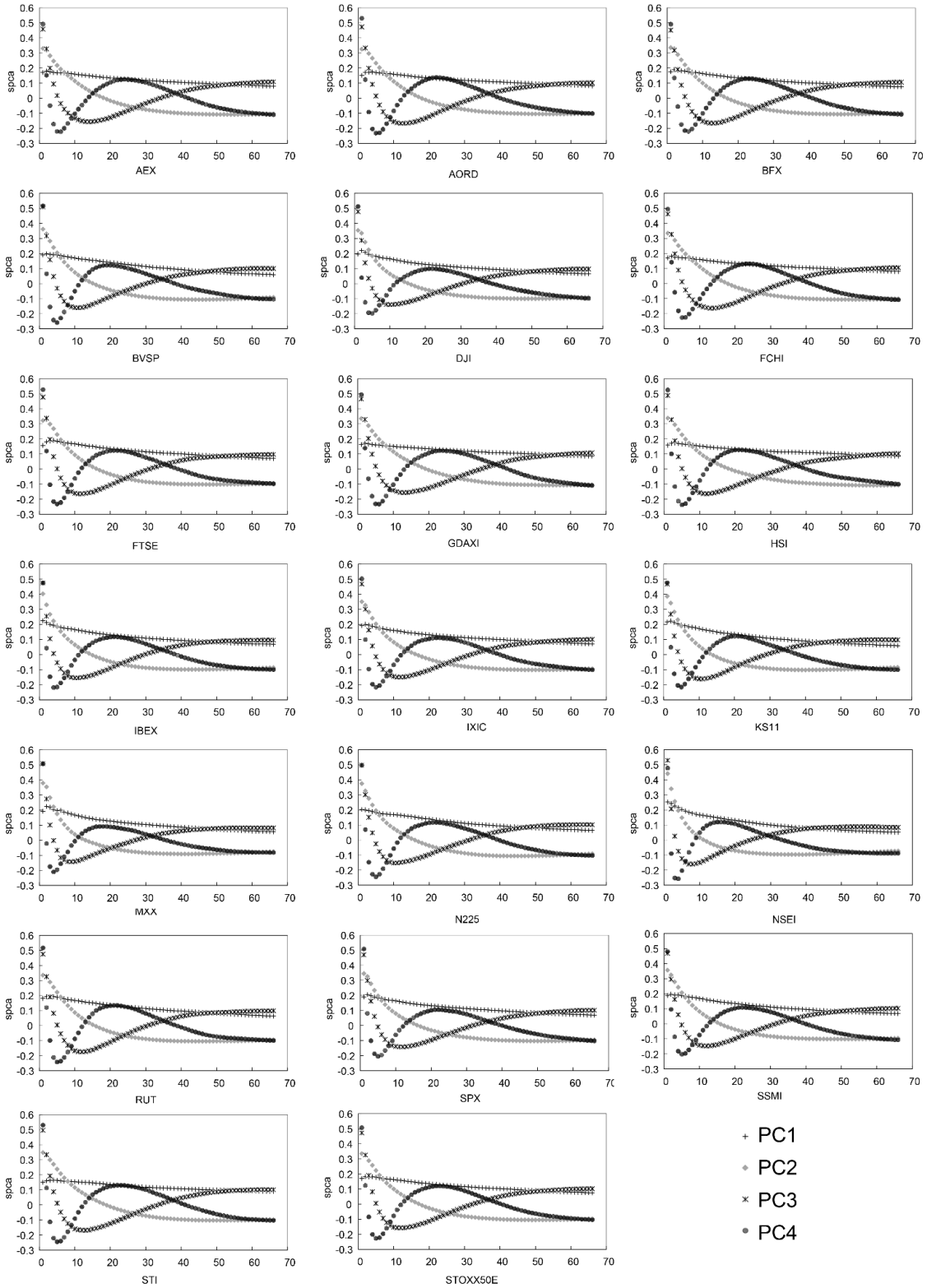
Figure 2: Four heterogeneous principal components extracted by the PCA method



Notes: *k* is set at 66. Four heterogeneous principal components represent the average of all rolling-window results.

Source: Authors' own elaboration

Figure 3: Four heterogeneous principal components extracted by the sPCA method



Note: Conditions are set as in Figure 2.

Source: Authors' own elaborations

5. Empirical study

5.1 Methods of forecasting evaluation

The out-of-sample forecasts for each model are conducted using a fixed-window rolling prediction approach. The scheme is executed as follows: the window size is empirically set at 1,000 in line with prior research. Subsequently, the model is constructed using the first 1,000 samples, and the parameters are retained. Make a one-step prediction based on these parameters. The window moves down one unit at a time. Section 5 uses a rolling window size of 1,000 for one-step forward forecasting.

We use two error metrics to compare the out-of-sample forecasting performance, namely Mean Squared Error (MSE) and Quasi-Likelihood (QLIKE), as shown in Equations (12) and (13) below:

$$\text{MSE} = T^{-1}\sum_{t=1}^{T}(RV_t - \widehat{RV}_t)^2 \tag{12}$$

$$\text{QLIKE} = T^{-1}\sum_{t=1}^{T}(log(\widehat{RV}_t) + \frac{RV_t}{\widehat{RV}_t}) \tag{13}$$

where *RVt* represents the actual value, and $\widehat{RV}_t$ signifies the in-sample fitted value or predicted value. When the in-sample fit value in MSE is replaced by the out-of-sample forecast value, this assessment criterion is called Mean Squared Predictive Error (MSPE).

The R_{oos}^2 test is widely utilized to assess the out-of-sample forecasting accuracy of models and to identify significant differences between these models and the benchmark model (Rapach *et al*., 2010, He *et al*., 2021). The R_{oos}^2 statistic is defined as follows:

$$R_{oos}^2 = 1 - \frac{MSPE_F}{MSPE_B} \tag{14}$$

where $MSPE_i$ is sample average of $(RV_t - \widehat{RV_{t,i}})^2$ for $i = F, B$, and F and B represent the forecasting model under examination and the benchmark model, respectively. A positive R_{oos}^2 statistic indicates that the predictive power of the forecasting model is stronger than that of the benchmark model, and a larger R_{oos}^2 implies better forecasting performance.

Furthermore, to rigorously ascertain if the MSPE of the forecasting model yields a statistically significant enhancement in comparison to the benchmark model, the MSPE adjustment model as formulated by Clark and West (2007) is introduced herein. The null hypothesis posits equal MSPE. The alternative is that the forecasting model has a smaller MSPE than the benchmark model. This examination is conducted by regressing $\widehat{f_{t+\tau}}$

on a constant, followed by scrutiny of the t-statistic associated with the constant coefficient. The $\widehat{f_{t+\tau}}$ is defined as follows:

$$\widehat{f_{t+h}} = (RV_{t+1} - \widehat{RV_{B,t+1}})^2 - [(RV_{t+1} - \widehat{RV_{F,t+1}})^2 - (\widehat{RV_{B,t+1}} - \widehat{RV_{F,t+1}})^2] \quad (15)$$

where RV_{t+1} represents the actual value, and $\widehat{RV_{F,t+1}}$ and $\widehat{RV_{B,t+1}}$ represent out-of-sample forecasts of the model being tested and the benchmark model, respectively. When the MSPE-adjusted statistic is significant and R_{oos}^2 is above zero, it indicates that the forecasting model outperforms the benchmark model from a statistical perspective.

The mentioned testing methods are conventionally employed to compare prediction performances between pairs of models. To facilitate a comprehensive comparison of prediction performance across multiple models, this study employs the model confidence set (MCS) test introduced by Hansen *et al*. (2011). The MCS test entails a series of significance assessments on the set denoted as E_0, which encompasses sequences of loss functions across multiple models, executed at a designated confidence level. Within this process, models with inferior predictive capabilities in E_0 are successively eliminated. This continues until no model remains excluded from E_0. The null and alternative hypotheses of the MCS test are as follows:

$$H_{0,E} : E(d_{ij,t}) = 0 , \quad \forall i, j \in E \subseteq E_0 \quad (16)$$

$$H_{A,E} : E(d_{ij,t}) \neq 0 , \quad \forall i, j \in E \subseteq E_0 , \quad (17)$$

where $d_{ij,t}$ is equal to the difference between the loss functions of model *i* and model *j*.

Different studies have proposed varying significance levels for the MCS test, with Liang *et al*. (2023) opting for 0.25, Noureldin (2022) for 0.05, and Guo *et al*. (2022) for 0.5. Unlike the conventional rationale for setting the significance level in hypothesis testing, the MCS procedure tests the null hypothesis that all candidate models exhibit equal predictive ability (EPA), iteratively eliminating inferior models from the model confidence set. A higher α level increases the likelihood of model exclusion, resulting in a smaller and more selective model confidence set (Hansen *et al*., 2011). To ensure that the surviving models have better performance, this study aligns with Wei *et al*. (2017) and Guo *et al*. (2022) by setting a significance level α of 0.5. The test is based on two error metrics of MSE and QLIKE. The MCS test results are derived by computing T_R, and T_{SQ}, which are defined as follows:

$$T_{R,E} = \max_{i,j \in E} \frac{|\bar{d}_{ij}|}{\sqrt{\widehat{var}(\bar{d}_{ij})}} \quad (18)$$

$$T_{SQ,E} = \max_{i,j \in E} \frac{(\bar{d}_{ij})^2}{\widehat{var}(d_{ij})} \tag{19}$$

where $\bar{d}_{ij} = \frac{1}{T}\sum_{t=1}^{T} d_{ij,t}$.

5.2 In-sample fit

In accordance with the analysis outlined in Section 4, the principal components derived from PCA and sPCA methods exhibit strong distinguishing characteristics and good economic implications. To further verify that they have explanatory power for RV in the following day, this section examines the in-sample fitting outcomes of the HAR, HAR-PCA, and HAR-sPCA models. The in-sample fitting results are obtained based on the full sample.

The results are presented in Figure 4 and Table 1. We find that the parameters of most of the HAR, HAR-PCA, and HAR-SPCA models constructed using 20 stock market indices are statistically significant. Furthermore, the adjusted R-squared for both HAR-PCA and HAR-SPCA models surpasses that of the HAR model across nearly all indices. The results confirm the economic rationality of the HAR model and demonstrate the effectiveness of the heterogeneous principal components in explaining the RV for the following day. The results provide a statistical foundation for the analysis in Section 4.

Figure 4: Coefficients from the in-sample fit for HAR, HAR-PCA, and HAR-sPCA models

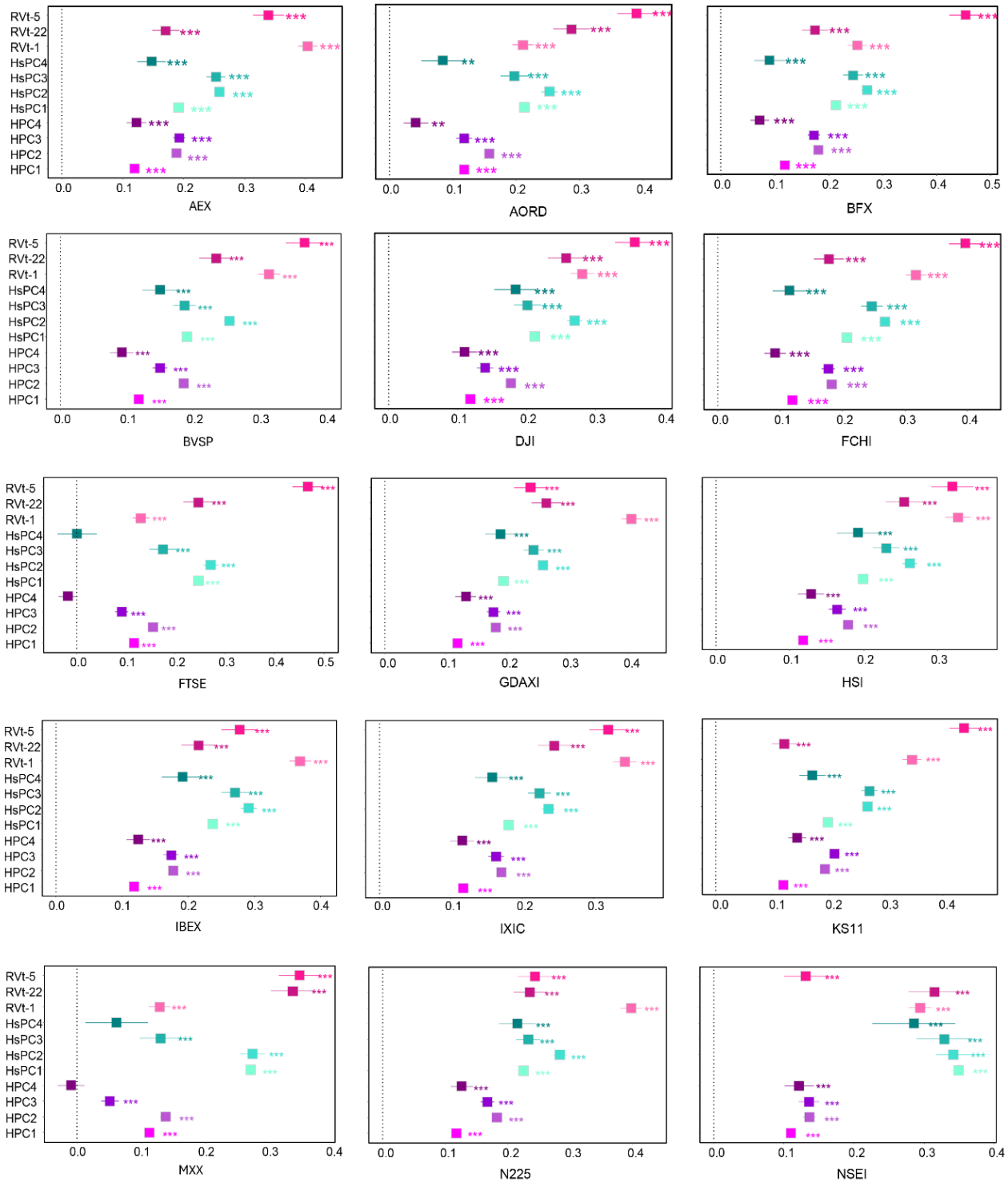
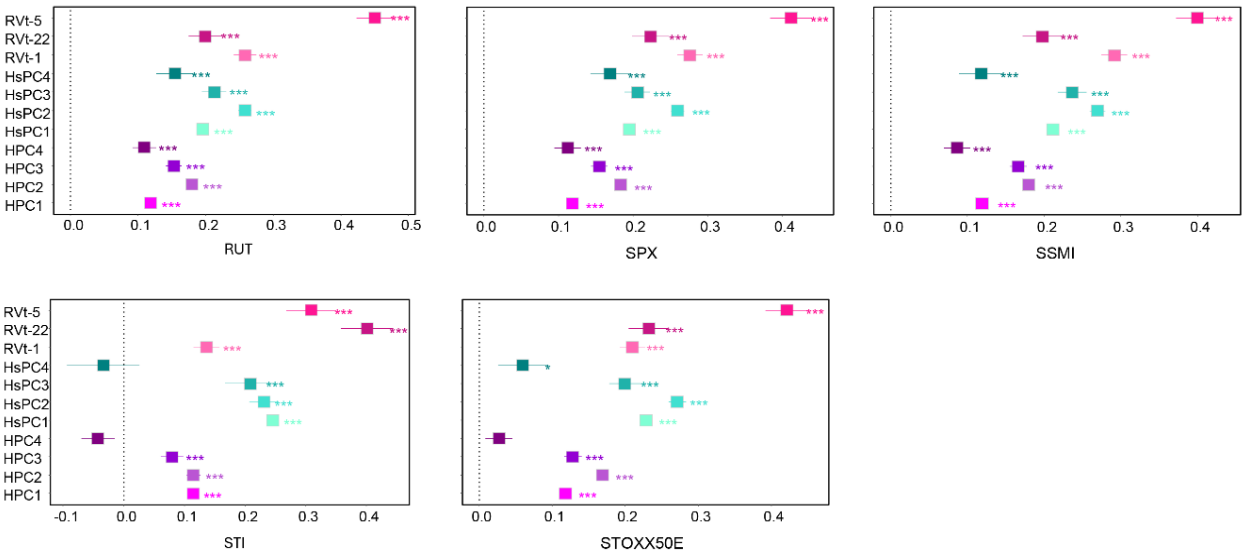


Figure 4: Continuation



Notes: The dots in the figure represent the coefficient estimates obtained from the in-sample fitting, with the corresponding 95% confidence intervals indicated. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Source: Authors' own elaboration

Table 1: Goodness-of-fit performance of the HAR, HAR-PCA, and HAR-sPCA models

	HAR		HAR-PCA		HAR-sPCA	
	Adj. R^2	RSE	Adj. R^2	RSE	Adj. R^2	RSE
AEX	0.597	1.245	0.598	1.245	0.598	1.244
AORD	0.419	0.605	0.423	0.603	0.423	0.603
BFX	0.485	0.974	0.499	0.962	0.5	0.961
BVSP	0.577	1.642	0.581	1.634	0.582	1.633
DJI	0.474	1.881	0.481	1.868	0.482	1.867
FCHI	0.516	1.552	0.523	1.539	0.524	1.539
FTSE	0.334	2.207	0.34	2.197	0.341	2.195
GDAXI	0.551	1.927	0.555	1.919	0.556	1.917
HIS	0.517	1.124	0.523	1.118	0.523	1.117
IBEX	0.435	1.551	0.432	1.554	0.433	1.553
IXIC	0.573	1.363	0.575	1.36	0.575	1.36
KS11	0.596	1.346	0.606	1.329	0.607	1.327
MXX	0.255	1.501	0.267	1.488	0.268	1.487
N225	0.482	1.179	0.484	1.177	0.487	1.174
NSEI	0.203	3.614	0.202	3.617	0.204	3.613
RUT	0.534	1.021	0.534	1.021	0.534	1.021
SPX	0.545	1.622	0.554	1.607	0.554	1.606
SSMI	0.49	1.112	0.496	1.105	0.496	1.105
STI	0.274	0.472	0.283	0.4691	0.283	0.469
STOXX50E	0.405	2.399	0.418	2.373	0.418	2.372

Note: The results in this table correspond to Figure 4, with the adjusted R-squared (Adj. R^2) and RSE used to evaluate the goodness-of-fit of the HAR, HAR-PCA, and HAR-sPCA models. RSEs are multiplied by 10,000.

Source: Authors' own calculations

5.3 Out-of-sample forecasting

In this section, we conduct a comprehensive evaluation of out-of-sample prediction capabilities across several models. The R_{oos}^2 test serves as a method for assessing differences among pairwise models. The MCS test serves as a powerful tool for comparing performance across multiple models.

Table 2 presents the R_{oos}^2 for each model and the MSPE-adjusted model results. The HAR model is regarded as a benchmark. The analysis demonstrates that under most indices, the R_{oos}^2 values for the HAR-PCA and HAR-sPCA models consistently surpass those of other models. This underscores the enhanced predictive capabilities of the HAR-PCA and HAR-sPCA models.

Table 2: R_{oos}^2 test for one-day-ahead out-of-sample forecasts

	HAR-PCA	HAR-sPCA	HAR-Lasso	HAR-Ridge	HAR-ENet		HAR-PCA	HAR-sPCA	HAR-Lasso	HAR-Ridge	HAR-ENet
AEX	**5.000 (1.220)**	**4.290 * (1.282)**	**0.010 * (1.466)**	−124.290 * (1.282)	3.570 ** (1.686)	**IXIC**	**2.760 * (1.570)**	**2.760 * (1.572)**	−13.100 (0.622)	−104.830 * (1.500)	**4.140 * (1.391)**
AORD	**0.480 ** (1.933)**	**0.240 ** (1.726)**	−4.280 (−0.555)	−70.780 (0.919)	−2.610 (−0.200)	**KS11**	**8.210 * (1.620)**	**7.690 ** (1.690)**	−6.670 (0.934)	−94.360 * (1.458)	**1.030 (1.215)**
BFX	**10.990 ** (1.880)**	**10.090 ** (1.998)**	−4.500 * (1.610)	−63.060 * (1.556)	**1.800 ** (1.787)**	**MXX**	**2.050 ** (2.312)**	**2.050 ** (2.033)**	−1.640 * (1.540)	−33.200 ** (1.536)	**0.410 ** (1.773)**
BVSP	**1.850 ** (2.226)**	**1.540 ** (2.245)**	−6.460 (−0.870)	−121.230 (1.182)	−2.770 (−0.452)	**N225**	**1.160 ** (1.698)**	**1.160 ** (1.692)**	−5.780 (1.063)	−84.390 ** (1.901)	−0.580 * (1.392)
DJI	**3.510 ** (1.700)**	**3.510 ** (1.749)**	−23.970 (0.794)	−57.640 ** (1.824)	−0.830 (1.247)	**NSEI**	**58.850 (1.037)**	**56.730 (1.024)**	−15.960 (0.023)	**62.120 (1.059)**	**5.960 (0.758)**
FCHI	**8.920 (1.249)**	**7.060 (1.218)**	−15.990 (0.740)	−67.660 * (1.436)	−2.600 (1.144)	**RUT**	**−0.700 (0.810)**	**−0.700 (0.695)**	−9.860 (0.961)	−86.620 * (1.330)	**2.110 * (1.631)**
FTSE	**0.830 (0.780)**	−0.330 (0.482)	−53.300 (−0.383)	−27.230 ** (1.938)	−13.530 (−0.333)	**SPX**	**3.780 ** (1.824)**	**3.780 ** (1.918)**	−30.270 (0.328)	−78.110 * (1.626)	−1.620 (0.822)
GDAXI	**8.490 ** (2.144)**	**5.970 ** (1.960)**	−14.150 (1.092)	−92.770 * (1.557)	−8.810 (1.229)	**SSMI**	**5.760 ** (1.840)**	**4.320 ** (1.748)**	−10.070 (0.359)	−61.870 ** (1.692)	−3.600 (0.901)
HSI	**0.650 ** (1.823)**	**0.650 ** (1.770)**	−2.600 (0.962)	−100.000 (0.939)	−2.600 (0.909)	**STI**	**3.350 * (1.342)**	**2.380 * (1.294)**	**1.410 * (1.557)**	−56.760 (0.479)	4.110 (0.972)
IBEX	**5.070 (1.258)**	**3.720 (1.201)**	−2.700 (1.253)	−53.720 * (1.2868)	−1.350 (1.158)	**STOX X50E**	**2.680 (1.174)**	**1.070 (0.721)**	−77.180 (−0.078)	−25.500 ** (1.876)	−9.400 (0.435)

Notes: MSPE- adjusted statistics are enclosed in parentheses. ** and * denote that the predictive accuracy test rejects the null hypothesis at the 5% and 10% statistical significance levels. Bold indicates R_{oos}^2 values above zero. Bold italics highlight instances where R_{oos}^2 is above zero and the MSPE-adjusted statistic is significant.

Source: Authors' own calculations

Table 3 and Table 4 display the outcomes of the MCS test evaluating MSE and QLIKE for each model, utilizing data from 20 indices. If a p-value for a forecasting model exceeds the significance level of 0.5, this model can be included in MCS. Moreover, a higher p-value

indicates enhanced predictive accuracy of the model. Under the MSE indicator, nearly all indices favor the HAR-PCA model. Results under QLIKE indicate a preference for the HAR-sPCA model. Consequently, the findings from the three tables collectively affirm that the HAR-PCA and HAR-sPCA models demonstrate superior forecasting performance compared to other models.

Table 3: The Model Confidence Set test for one-day-ahead out-of-sample forecasts under the MSE criterion

Model	HAR		HAR-PCA		HAR-sPCA		HAR-Lasso		HAR-Ridge		HAR-ENet	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
AEX	0.250	0.151	**1.000**	**1.000**	0.277	0.270	0.277	0.154	0.147	0.110	0.277	0.172
AORD	0.273	0.320	**1.000**	**1.000**	0.273	0.320	0.203	0.213	0.167	0.099	0.273	0.282
BFX	0.234	0.101	**1.000**	**1.000**	0.234	0.132	0.234	0.105	0.066	0.078	0.234	0.105
BVSP	0.286	0.257	**1.000**	**1.000**	0.286	0.257	0.286	0.223	0.257	0.091	0.286	0.235
DJI	0.310	0.297	**0.576**	**0.576**	**1.000**	**1.000**	0.310	0.251	0.174	0.100	0.496	0.431
FCHI	0.116	0.109	**1.000**	**1.000**	0.335	0.335	0.116	0.109	0.022	0.075	0.116	0.109
FTSE	**0.736**	**0.656**	**1.000**	**1.000**	**0.736**	**0.656**	**0.650**	**0.509**	0.027	0.077	**0.730**	**0.535**
GDAXI	0.161	0.120	**1.000**	**1.000**	0.205	0.205	0.158	0.099	0.048	0.066	0.161	0.120
HSI	0.385	0.268	**1.000**	**1.000**	**0.557**	**0.557**	0.385	0.268	0.300	0.118	0.385	0.268
IBEX	0.243	0.332	**1.000**	**1.000**	0.243	0.332	0.242	0.234	0.027	0.040	0.243	0.332
IXIC	0.150	0.177	**0.603**	**0.603**	0.499	**0.509**	0.150	0.137	0.150	0.085	**1.000**	**1.000**
KS11	0.342	**0.503**	**1.000**	**1.000**	**0.517**	**0.517**	0.281	0.151	0.169	0.092	0.286	0.453
MXX	0.300	0.327	**1.000**	**1.000**	**0.758**	**0.758**	0.241	0.220	0.156	0.131	0.300	0.352
N225	0.493	**0.555**	**0.851**	**0.851**	**1.000**	**1.000**	0.334	0.279	0.172	0.077	0.493	**0.555**
NSEI	0.228	**0.511**	**0.645**	**0.645**	0.228	**0.602**	0.228	**0.602**	**1.000**	**1.000**	0.228	**0.602**
RUT	0.280	0.387	**0.549**	**0.549**	0.280	0.387	0.280	0.241	0.160	0.089	**1.000**	**1.000**
SPX	0.170	0.096	0.170	0.116	**1.000**	**1.000**	0.170	0.091	0.091	0.080	0.170	0.108
SSMI	0.103	0.086	**1.000**	**1.000**	0.168	0.120	0.103	0.079	0.103	0.070	0.168	0.102
STI	0.071	0.107	**0.574**	**0.574**	0.105	**0.557**	0.065	0.064	0.000	0.000	**1.000**	**1.000**
STOXX50E	0.087	0.085	**1.000**	**1.000**	0.326	0.326	0.087	0.085	0.015	0.070	0.087	0.085

Notes: The significance level of the MCS is set at 0.5. MCS confidence sets are computed using two widely adopted statistics T_R and T_{SQ}.

Source: Authors' own calculations

Table 4: The Model Confidence Set test for one-day-ahead out-of-sample forecasts under the QLIKE criterion

Model	HAR		HAR-PCA		HAR-sPCA		HAR-Lasso		HAR-Ridge		HAR-ENet	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
AEX	0.142	0.136	0.142	0.136	**1.000**	**1.000**	0.000	0.001	0.000	0.000	0.023	0.054
AORD	0.000	0.002	**1.000**	**1.000**	0.084	0.084	0.000	0.002	0.000	0.002	0.000	0.002
BFX	0.133	0.116	0.284	0.284	**1.000**	**1.000**	0.007	0.008	0.004	0.001	0.007	0.023
BVSP	**1.000**	**1.000**	0.024	0.029	0.024	0.029	0.024	0.029	0.001	0.005	0.001	0.011
DJI	0.202	0.206	0.202	0.206	**1.000**	**1.000**	0.202	0.155	0.000	0.001	0.202	0.155
FCHI	0.215	0.129	0.215	0.129	**1.000**	**1.000**	0.215	0.126	0.000	0.005	0.215	0.129
FTSE	0.217	0.122	**1.000**	**1.000**	0.217	0.165	0.217	0.104	0.005	0.015	0.217	0.122
GDAXI	**0.547**	**0.540**	**0.812**	**0.812**	**1.000**	**1.000**	0.491	0.397	0.001	0.013	0.491	0.332
HSI	**0.815**	**0.861**	**0.815**	**0.861**	**1.000**	**1.000**	0.163	0.144	0.016	0.013	0.376	0.452
IBEX	0.388	0.405	0.388	0.405	**1.000**	**1.000**	0.146	0.110	0.010	0.018	0.146	0.110
IXIC	**0.562**	**0.562**	0.068	0.158	**1.000**	**1.000**	0.068	0.100	0.000	0.000	0.068	0.100
KS11	**0.615**	**0.655**	0.096	0.328	**1.000**	**1.000**	0.096	0.112	0.002	0.004	**0.615**	**0.655**
MXX	0.002	0.002	0.460	0.460	**1.000**	**1.000**	0.002	0.002	0.002	0.000	0.002	0.002
N225	**1.000**	**1.000**	**0.524**	0.500	**0.524**	0.500	0.358	0.193	0.000	0.001	**0.524**	0.500
NSEI	**1.000**	**1.000**	0.245	0.191	**0.908**	**0.908**	0.245	0.131	0.000	0.001	0.245	0.131
RUT	0.019	0.023	0.019	0.023	**1.000**	**1.000**	0.019	0.023	0.009	0.003	0.019	0.023
SPX	0.160	0.084	0.160	0.084	**1.000**	**1.000**	0.160	0.084	0.002	0.007	0.160	0.084
SSMI	**0.919**	**0.919**	0.181	0.339	**1.000**	**1.000**	0.065	0.046	0.002	0.001	0.031	0.024
STI	0.001	0.004	**1.000**	**1.000**	0.001	0.004	0.000	0.000	0.000	0.000	0.001	0.001
STOXX50E	0.144	0.110	0.210	0.210	**1.000**	**1.000**	0.144	0.110	0.002	0.010	0.144	0.110

Note: All other conditions adhere to those specified in Table 3.

Source: Authors' own calculations

6. Robustness check

6.1 Alternative forecasting window

Figure B.1 shows the results of the MCS test for each model under the MSE and QLIKE error measure metrics after considering the change in the size of the rolling window, based on the 20 indices. It is found that the HAR-PCA model enters MCS the most times under the MSE criterion, and the *p*-value is mostly 1. Under the QLIKE criterion, the HAR-sPCA model enters MCS more often. The performance of the two models is robust.

6.2 Alternative maximum lag periods

Figure B.2 presents the MCS test results for each model using data from 20 indices, applying MSE and QLIKE error metrics, with *k* set at 22. The results show that under the MSE criterion, the HAR-PCA model enters more into MCS, and most of the *p*-values are equal to one. Under the QLIKE criterion, the HAR-sPCA model is more frequently included in the MCS, and most of the *p*-values are equal to 1. The HAR-PCA and HAR-sPCA models exhibit the most robust performance.

6.3 Alternative numbers of principal components

To strengthen the rationale for selecting four principal components, we conducted comparative analyses across models using 3, 4, and 5 principal components. The results are summarized in Figure B.3. According to the MSE criterion, 13 indices identified the HAR-PCA model with four principal components as the optimal specification, as it entered the MCS confidence set most frequently. The HAR-sPCA model with four principal components followed closely, entering the MCS confidence set six times – ranking second in frequency. Under the QLIKE metric, the HAR-sPCA model with four principal components entered the MCS confidence set 17 times, which was the highest among all models. Taken together, these findings provide robust empirical support for selecting four principal components in both the HAR-PCA and HAR-sPCA models.

6.4 Alternative significance levels for the MCS test

Given the variation in significance levels employed across the existing literature, we adopt multiple significance levels for the MCS test to strengthen the robustness of our conclusions. We perform the tests at significance levels of 0.05, 0.10, and 0.25 (Ma *et al*., 2019, Noureldin, 2022, Liang *et al*., 2023), with the results summarized in Figure B.4, Figure B.5, and Figure

B.6. Under the MSE criterion, the HAR-PCA model ranks first 13 times at significance levels of 0.05, 0.1, and 0.25, and is included in the MCS confidence set 19 to 20 times, being the model with the highest ranking and the most frequent inclusion in the MCS confidence set. The HAR-sPCA model follows closely behind. According to the QLIKE criterion, at significance levels of 0.05, 0.10, and 0.25, the HAR-sPCA model achieves the highest rank on 13 to 14 indices and is included in the MCS confidence set 16 to 17 times, making it the best-performing model in terms of both ranking and inclusion frequency. The above analysis indicates that the previous results are robust.

7. Longer-Run Forecasting

This section further studies the attributes of the HAR-PCA and HAR-sPCA models in longer-run forecasting. We adapt the forecast horizon of the HAR model from one day ahead to encompass one week ahead and one month ahead. We replace the dependent variable in all models from next-day realized variance ($RV_{t+1}^{(d)}$) to next-week average realized variance ($RV_{t+5|t+1}$) and next-month average realized variance ($RV_{t+22|t+1}$), as a function of the information set that is available at time *t*. Furthermore, since the HAR-sPCA model is a supervised learning model, the principal components in the HAR-sPCA model will be re-estimated based on these new forecasting targets.

Table 5 and Table 6 show the results of the MCS test under the MSE and QLIKE error criteria when using each model for one-week-ahead forecasting of the 20 indices. The results show that the HAR-PCA model has the most superior performance under the MSE criterion, and all the indices except DJI are included in MCS. Under the QLIKE criterion, the HAR-PCA and HAR-sPCA models have similar performance, most of the indexes are included in MCS, and more than half of the p-values are equal to 1, which is significantly better than other models. Therefore, HAR-PCA and HAR-sPCA models have superior performance when conducting a one-week-ahead forecast.

Table 5: The Model Confidence Set test for one-week-ahead out-of-sample forecasts under the MSE criterion

Model	HAR		HAR-PCA		HAR-sPCA		HAR-Lasso		HAR-Ridge		HAR-ENet	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
AEX	**0.579**	**0.541**	**1.000**	**1.000**	**0.579**	**0.541**	0.221	0.250	0.175	0.085	**0.579**	**0.541**
AORD	0.404	0.332	**1.000**	**1.000**	**0.604**	**0.604**	0.290	0.241	0.238	0.111	0.404	0.332
BFX	0.183	0.128	**1.000**	**1.000**	0.190	0.150	0.183	0.104	0.106	0.076	0.190	0.144
BVSP	**0.675**	**0.610**	**1.000**	**1.000**	**0.675**	**0.668**	0.318	0.424	0.298	0.191	**0.675**	**0.610**
DJI	0.238	0.154	0.238	0.190	**1.000**	**1.000**	0.238	0.095	0.079	0.073	0.238	0.108
FCHI	0.142	0.120	**1.000**	**1.000**	0.207	0.207	0.142	0.120	0.093	0.072	0.142	0.120
FTSE	0.218	0.321	**1.000**	**1.000**	0.449	0.426	0.218	0.309	0.029	0.076	0.449	0.391
GDAXI	0.248	0.152	**1.000**	**1.000**	0.248	0.152	0.104	0.091	0.057	0.066	0.248	0.106
HSI	**0.519**	0.411	**1.000**	**1.000**	**0.531**	**0.531**	0.386	0.334	0.294	0.179	**0.519**	0.429
IBEX	0.222	0.370	**1.000**	**1.000**	0.222	0.370	0.045	0.131	0.045	0.049	0.222	0.370
IXIC	0.133	0.106	**1.000**	**1.000**	0.135	0.135	0.133	0.100	0.133	0.076	0.133	0.106
KS11	0.279	0.451	**1.000**	**1.000**	0.279	0.451	0.279	0.362	0.075	0.072	0.258	0.205
MXX	0.324	0.413	**0.954**	**0.954**	**1.000**	**1.000**	0.324	0.361	0.141	0.129	0.324	0.343
N225	**0.630**	**0.620**	**1.000**	**1.000**	**0.729**	**0.729**	0.060	0.166	0.060	0.101	**0.678**	**0.680**
NSEI	0.315	**0.593**	**1.000**	**1.000**	0.315	**0.593**	0.315	**0.593**	0.315	**0.593**	0.315	**0.593**
RUT	0.282	0.175	**1.000**	**1.000**	0.282	0.180	0.211	0.140	0.211	0.101	0.282	0.175
SPX	0.455	0.386	**0.543**	**0.543**	**1.000**	**1.000**	0.455	0.252	0.125	0.082	0.455	0.312
SSMI	0.185	0.083	**1.000**	**1.000**	0.185	0.105	0.133	0.072	0.133	0.072	0.185	0.083
STI	0.051	0.018	**0.577**	**0.577**	0.051	0.092	0.051	0.092	0.000	0.000	**1.000**	**1.000**
STOXX50E	0.313	0.334	**1.000**	**1.000**	0.487	0.368	0.090	0.092	0.004	0.065	0.487	0.368

Note: All other conditions are set as in Table 3.

Source: Authors' own calculations

Table 6: The Model Confidence Set test for one-week-ahead out-of-sample forecasts under the QLIKE criterion

Model	HAR		HAR-PCA		HAR-sPCA		HAR-Lasso		HAR-Ridge		HAR-ENet	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
AEX	0.036	0.029	**0.638**	**0.638**	**1.000**	**1.000**	0.161	0.123	0.002	0.001	0.161	0.094
AORD	0.005	0.026	**1.000**	**1.000**	**0.534**	**0.503**	0.295	0.325	0.005	0.012	**0.534**	**0.503**
BFX	0.037	0.014	**1.000**	**1.000**	**0.730**	**0.730**	0.037	0.041	0.011	0.004	0.037	0.022
BVSP	**1.000**	**1.000**	0.105	0.105	0.053	0.084	0.053	0.084	0.040	0.024	0.053	0.084
DJI	**1.000**	**1.000**	**0.510**	**0.507**	**0.515**	**0.515**	**0.510**	0.474	0.002	0.049	**0.510**	**0.507**
FCHI	0.143	0.046	0.170	0.170	1.000	1.000	0.143	0.047	0.005	0.005	0.143	0.047
FTSE	0.103	0.064	**1.000**	**1.000**	0.236	0.199	0.236	0.125	0.012	0.004	0.103	0.064
GDAXI	**0.533**	0.328	**1.000**	**1.000**	**0.533**	0.328	**0.533**	0.290	0.000	0.009	**0.533**	0.296
HSI	0.430	**0.606**	**0.879**	**0.879**	**0.663**	**0.766**	0.430	**0.606**	0.026	0.032	**1.000**	**1.000**
IBEX	0.028	0.018	**0.656**	**0.656**	**1.000**	**1.000**	0.028	0.018	0.025	0.002	0.028	0.018
IXIC	0.276	0.426	0.276	0.426	**1.000**	**1.000**	0.098	0.092	0.001	0.001	0.127	0.184
KS11	0.246	**0.574**	0.246	**0.574**	**0.729**	**0.729**	**1.000**	**1.000**	0.007	0.006	0.367	**0.574**
MXX	**0.994**	**0.995**	**0.994**	**0.995**	**0.547**	**0.862**	**0.994**	**0.995**	0.009	0.016	**1.000**	**1.000**
N225	**0.547**	**0.547**	**1.000**	**1.000**	0.355	0.462	0.355	0.327	0.000	0.003	0.355	0.462
NSEI	0.482	0.443	0.482	0.449	**1.000**	**1.000**	0.073	0.075	0.000	0.000	0.024	0.027
RUT	0.044	0.112	0.044	0.112	**1.000**	**1.000**	0.044	0.103	0.028	0.031	0.044	0.112
SPX	**1.000**	**1.000**	**0.633**	**0.601**	**0.745**	**0.701**	0.470	0.428	0.004	0.018	0.745	0.701
SSMI	0.279	0.183	0.279	0.183	**1.000**	**1.000**	0.182	0.087	0.006	0.005	0.182	0.128
STI	0.000	0.001	**1.000**	**1.000**	0.000	0.002	0.000	0.002	0.000	0.000	0.000	0.002
STOXX50E	0.051	0.036	**1.000**	**1.000**	0.462	0.462	0.051	0.036	0.003	0.002	0.014	0.017

Note: All other conditions are set as in Table 3.

Source: Authors' own calculations

Table 7 and Table 8 show the results of the MCS test for each model with one-month-ahead forecasts for the 20 indices under the MSE and QLIKE error criteria. The results show that under the MSE criterion, the HAR-PCA and HAR-sPCA models have significantly better performance than the other models, they are included in the MCS more frequently, and most of the p-values are equal to one. Under the QLIKE criterion, the HAR-PCA model has the most superior performance, and only six indices are not included in the MCS. Therefore, HAR-PCA and HAR-sPCA models have superior performance when conducting a one-month-ahead forecast.

Table 7: The Model Confidence Set test for one-month-ahead out-of-sample forecasts under the MSE criterion

Model	HAR		HAR-PCA		HAR-sPCA		HAR-Lasso		HAR-Ridge		HAR-ENet	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
AEX	0.264	0.263	**1.000**	**1.000**	**0.688**	**0.586**	0.246	0.234	0.201	0.122	**0.688**	**0.586**
AORD	**0.715**	**0.704**	**0.960**	**0.960**	**1.000**	**1.000**	0.434	0.339	0.264	0.133	**0.520**	**0.521**
BFX	0.351	0.160	0.351	0.160	1.000	1.000	0.351	0.148	0.151	0.088	0.351	0.160
BVSP	0.211	0.301	**1.000**	**1.000**	0.211	0.301	0.211	0.209	0.211	0.105	0.211	0.301
DJI	**1.000**	**1.000**	0.414	0.430	**0.523**	**0.523**	0.371	0.414	0.127	0.099	0.414	0.430
FCHI	0.168	0.110	0.168	0.110	**1.000**	**1.000**	0.168	0.100	0.117	0.068	0.168	0.110
FTSE	0.195	0.145	0.195	0.145	**1.000**	**1.000**	0.195	0.121	0.025	0.070	0.195	0.145
GDAXI	0.131	0.104	**1.000**	**1.000**	0.463	0.463	0.131	0.104	0.088	0.055	0.227	0.186
HSI	**0.568**	**0.591**	**0.546**	**0.591**	**0.568**	**0.591**	**0.546**	**0.591**	0.272	0.200	**1.000**	**1.000**
IBEX	0.145	0.087	0.145	0.087	1.000	1.000	0.145	0.079	0.099	0.046	0.145	0.087
IXIC	**0.512**	0.296	**1.000**	**1.000**	**0.512**	0.365	0.462	0.226	0.138	0.082	**0.512**	0.296
KS11	0.197	0.213	**1.000**	**1.000**	0.197	0.213	0.197	0.213	0.110	0.082	0.197	0.213
MXX	0.313	0.340	**1.000**	**1.000**	**0.660**	**0.702**	0.313	0.340	0.208	0.129	**0.675**	**0.702**
N225	0.180	0.116	**1.000**	**1.000**	0.180	0.116	0.180	0.082	0.180	0.074	0.180	0.116
NSEI	0.253	0.237	0.401	0.401	**1.000**	**1.000**	0.133	0.131	0.253	0.237	0.253	0.237
RUT	0.360	0.394	**1.000**	**1.000**	0.360	0.394	0.318	0.201	0.318	0.127	0.360	0.394
SPX	**0.552**	0.350	**0.552**	0.350	**1.000**	**1.000**	**0.504**	0.284	0.138	0.105	**0.552**	0.316
SSMI	0.303	0.257	**0.830**	**0.830**	**1.000**	**1.000**	0.139	0.059	0.139	0.059	0.194	0.117
STI	0.002	0.020	0.443	**0.542**	**1.000**	**1.000**	0.219	0.187	0.000	0.000	0.443	**0.542**
STOXX50E	0.236	0.279	**1.000**	**1.000**	**0.948**	**0.948**	0.236	0.220	0.006	0.061	0.236	0.279

Note: All other conditions are set as in Table 3.

Source: Authors' own calculations

Table 8: The Model Confidence Set test for one-month-ahead out-of-sample forecasts under the QLIKE criterion

Model	HAR		HAR-PCA		HAR-sPCA		HAR-Lasso		HAR-Ridge		HAR-ENet	
	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}	T_R	T_{SQ}
AEX	0.003	0.004	**0.823**	**0.823**	**1.000**	**1.000**	**0.591**	**0.556**	0.003	0.001	0.432	0.401
AORD	0.225	0.103	0.411	0.411	**1.000**	**1.000**	0.314	0.221	0.063	0.038	0.309	0.199
BFX	0.042	0.029	**1.000**	**1.000**	0.472	0.472	0.103	0.067	0.030	0.017	0.103	0.067
BVSP	**0.748**	**0.748**	**1.000**	**1.000**	0.289	0.290	0.430	0.368	0.146	0.059	0.289	0.290
DJI	0.053	0.115	0.053	0.115	0.053	0.115	0.053	0.115	0.046	0.071	**1.000**	**1.000**
FCHI	0.055	0.023	**1.000**	**1.000**	**0.827**	**0.827**	0.140	0.098	0.021	0.007	0.215	0.213
FTSE	0.107	0.053	**0.952**	**0.952**	**1.000**	**1.000**	0.105	0.039	0.037	0.015	0.107	0.066
GDAXI	0.108	0.103	**1.000**	**1.000**	0.453	0.453	0.442	0.340	0.006	0.005	0.442	0.338
HSI	0.280	0.232	**1.000**	**1.000**	0.368	0.368	0.050	0.029	0.050	0.008	0.065	0.079
IBEX	0.005	0.003	0.386	0.386	**1.000**	**1.000**	0.005	0.003	0.005	0.003	0.005	0.003
IXIC	**0.706**	**0.593**	**1.000**	**1.000**	**0.922**	**0.899**	**0.812**	**0.809**	0.025	0.032	**0.922**	**0.899**
KS11	0.011	0.079	**1.000**	**1.000**	0.011	0.079	0.011	0.079	0.011	0.013	0.011	0.079
MXX	0.151	0.158	**0.801**	**0.801**	0.232	0.433	0.232	0.433	0.020	0.027	**1.000**	**1.000**
N225	0.026	0.012	**1.000**	**1.000**	0.054	0.054	0.019	0.007	0.012	0.001	0.013	0.007
NSEI	0.488	0.411	**1.000**	**1.000**	**0.905**	**0.905**	0.488	0.336	0.001	0.002	**0.528**	**0.506**
RUT	**1.000**	**1.000**	0.379	0.397	0.379	0.397	0.379	0.397	0.050	0.080	**0.566**	**0.566**
SPX	0.398	0.471	0.342	0.471	0.342	0.471	0.342	0.471	0.024	0.061	**1.000**	**1.000**
SSMI	0.046	0.054	**1.000**	**1.000**	0.742	0.742	0.046	0.054	0.029	0.013	0.239	0.209
STI	0.077	0.048	**0.812**	**0.812**	**1.000**	**1.000**	0.251	0.254	0.000	0.000	0.106	0.127
STOXX50E	0.077	0.031	0.404	0.404	**1.000**	**1.000**	0.247	0.188	0.013	0.006	0.123	0.103

Notes: All other conditions are set as in Table 3.

Source: Authors' own calculations

8. Conclusions

Within the HAR framework, this paper constructs HAR-PCA and HAR-sPCA models based on the 20 stock market indices using PCA and sPCA methods by relaxing the maximum lag period. We extracted four heterogeneous principal components from the heterogeneous lag terms by employing these two models. Our analyses of the coefficients of these heterogeneous principal components reveal insights into their main, short-term, medium-term, and long-term effects. Additionally, this finding leads us to classify the market participants into short-term, medium-term, and long-term traders. Further, we confirm that they have good economic explanatory power. Our construction of HAR-PCA and HAR-sPCA models offers a more scientific validation of the heterogeneity among market traders, as proposed by the heterogeneous market hypothesis. By evaluating the out-of-sample forecasts, we find that the HAR-PCA and HAR-sPCA models have better predictive power than the competing models under the R_{oos}^2 test and the MCS test. In other words, PCA and sPCA methods seem to significantly improve the forecasting effect of the HAR model in predicting the volatility of stock indices. Furthermore, our results are robust to various settings. Our findings extend the current study on volatility forecasting models and provide further support for the classical theory of heterogeneous market hypothesis.

Acknowledgement

Funding: There was no funding, either externally or internally, towards this study.

Conflicts of interest: The authors hereby declare that this article was neither submitted nor published elsewhere.

AI usage statement: The authors confirm that no artificial intelligence (AI) or AI-assisted tools were used in the creation of this manuscript.

References

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

Andersen, T. G., Bollerslev, T. and Diebold, F. X. (2007). Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. *The Review of Economics and Statistics*, 89(4), 701-720.

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

Audrino, F., Huang, C. and Okhrin, O. (2018). Flexible HAR model for realized volatility. *Studies in Nonlinear Dynamics & Econometrics*, 23(3), 20170080.

Baillie, R. T., Bollerslev, T. and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1), 3-30.

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

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

Bollerslev, T., Tauchen, G. and Zhou, H. (2009). Expected Stock Returns and Variance Risk Premia. *Review of Financial Studies*, 22(11), 4463-4492.

Bollerslev, T., Patton, A. J. and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1), 1-18.

Branco, R. R., Rubesam, A. and Zevallos, M. (2024). Forecasting realized volatility: Does anything beat linear models? *Journal of Empirical Finance*, 78, 101524.

Buncic, D. and Gisler, K. I. M. (2016). Global equity market volatility spillovers: A broader role for the United States. *International Journal of Forecasting*, 32(4), 1317-1339.

Chen, X. and Ghysels, E. (2011). News—good or bad—and its impact on volatility predictions over multiple horizons. *The Review of Financial Studies*, 24(1), 46-81.

Christensen, K., Siggaard, M. and Veliyev, B. (2022). A Machine Learning Approach to Volatility Forecasting. *Journal of Financial Econometrics*, 21(5), 1680-1727.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291-311.

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

Ding, Y., Kambouroudis, D. and McMillan, D. G. (2021). Forecasting realised volatility: Does the LASSO approach outperform HAR? *Journal of International Financial Markets, Institutions and Money*, 74, 101386.

Drerup, T., Enke, B. and Von Gaudecker, H.-M. (2017). The precision of subjective data and the explanatory power of economic models. *Journal of Econometrics*, 200(2), 378-389.

Dudek, G., Fiszeder, P., Kobus, P., et al. (2024). Forecasting cryptocurrencies volatility using statistical and machine learning methods: A comparative study. *Applied Soft Computing*, 151, 111132.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4), 987-1007.

Granger, C. W. (2008). Non-linear models: Where do we go next-Time varying parameter models? *Studies in Nonlinear Dynamics & Econometrics*, 12(3), 1-11.

Granger, C. W. J. and Ding, Z. (1996). Varieties of long memory models. *Journal of Econometrics*, 73(1), 61-77.

Guo, Y., He, F., Liang, C., et al. (2022). Oil price volatility predictability: New evidence from a scaled PCA approach. *Energy Economics*, 105, 105714.

Hansen, P. R., Lunde, A. and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453-497.

He, M., Zhang, Y., Wen, D., et al. (2021). Forecasting crude oil prices: A scaled PCA approach. *Energy Economics*, 97, 105189.

Huang, D., Jiang, F., Li, K., et al. (2022). Scaled PCA: A new approach to dimension reduction. *Management Science*, 68(3), 1678-1695.

Hull, J. and White, A. (1987). The Pricing of Options on Assets with Stochastic Volatilities. *The Journal of Finance*, 42(2), 281-300.

Liang, C., Xu, Y., Chen, Z., et al. (2023). Forecasting China's stock market volatility with shrinkage method: Can Adaptive Lasso select stronger predictors from numerous predictors? *International Journal of Finance & Economics*, 28(4), 3689-3699.

Ma, F., Guo, Y., Chevallier, J., et al. (2022). Macroeconomic attention, economic policy uncertainty, and stock volatility predictability. *International Review of Financial Analysis*, 84, 102339.

Ma, Y., Ji, Q. and Pan, J. (2019). Oil financialization and volatility forecast: Evidence from multidimensional predictors. *Journal of Forecasting*, 38(6), 564-581.

Müller, U., Dacorogna, M., Dav, R., et al. (1993). Fractals and Intrinsic Time—A Challenge to Econometricians. *39th International AEA Conference on Real Time Econometrics. 14-15 October 1993*. Luxembourg.

Müller, U. A., Dacorogna, M. M., Davé, R. D., et al. (1997). Volatilities of different time resolutions—Analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2-3), 213-239.

Niu, Z., Ma, F. and Zhang, H. (2022). The role of uncertainty measures in volatility forecasting of the crude oil futures market before and during the COVID-19 pandemic. *Energy Economics*, 112, 106120.

Noureldin, D. (2022). Volatility Prediction Using a Realized-Measure-Based Component Model. *Journal of Financial Econometrics*, 20(1), 76-104.

Patton, A. J. and Sheppard, K. (2015). Good Volatility, Bad Volatility: Signed Jumps and The Persistence of Volatility. *The Review of Economics and Statistics*, 97(3), 683-697.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572.

Rapach, D. E., Strauss, J. K. and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2), 821-862.

Wei, Y., Liu, J., Lai, X., et al. (2017). Which determinant is the most informative in forecasting crude oil market volatility: Fundamental, speculation, or uncertainty? *Energy Economics*, 68, 141-150.

Appendix A

HAR-LASSO model

The form of the HAR-LASSO model is as follows:

$$\begin{gathered} RV_{t+1d}^{(d)} = \beta_0 + \sum_{i=1}^{k} \beta_i RV_t^{(i)} + \varepsilon_{t+1d} \\ \beta = \arg\min_{\beta} \left\{ \sum_{t=p}^{T} (RV_{t+1d}^{(d)} - \beta_0 - \sum_{i=1}^{k} \beta_i RV_t^{(i)})^2 + \lambda_1 \sum_{i=1}^{k} |\beta_i| \right\} \end{gathered} \tag{A.1}$$

where T and p represent the upper and lower limits of the prediction interval, respectively, and λ_1 is obtained by a 5-fold cross-validation process.

HAR-Ridge model

The form of the HAR-Ridge model is as follows:

$$\begin{gathered} RV_{t+1d}^{(d)} = \beta_0 + \sum_{i=1}^{k} \beta_i RV_t^{(i)} + \varepsilon_{t+1d} \\ \beta = \arg\min_{\beta} \left\{ \sum_{t=p}^{T} (RV_{t+1d}^{(d)} - \beta_0 - \sum_{i=1}^{k} \beta_i RV_t^{(i)})^2 + \lambda_2 \sum_{i=1}^{k} \beta_i^2 \right\} \end{gathered} \tag{A.2}$$

where λ_2 is obtained by a 5-fold cross-validation process.

HAR-ENet model

The form of the HAR-ENet model is as follows:

$$\begin{gathered} RV_{t+1d}^{(d)} = \beta_0 + \sum_{i=1}^{k} \beta_i RV_t^{(i)} + \varepsilon_{t+1d} \\ \beta = \arg\min_{\beta} \left\{ \begin{array}{c} \sum_{t=p}^{T} (RV_{t+1d}^{(d)} - \beta_0 - \sum_{i=1}^{k} \beta_i RV_t^{(i)})^2 + \lambda_1 \sum_{i=1}^{k} |\beta_i| + \\ \lambda_2 \sum_{i=1}^{k} \beta_i^2 \end{array} \right\} \end{gathered} \tag{A.3}$$

where $\alpha = \dfrac{\lambda_1}{\lambda_1 + \lambda_2}$ and $\lambda = \lambda_1 + \lambda_2$, thereby:

$$RV_{t+1d}^{(d)} = \beta_0 + \sum_{i=1}^{k} \beta_i RV_t^{(i)} + \varepsilon_{t+1d} \tag{A.4}$$

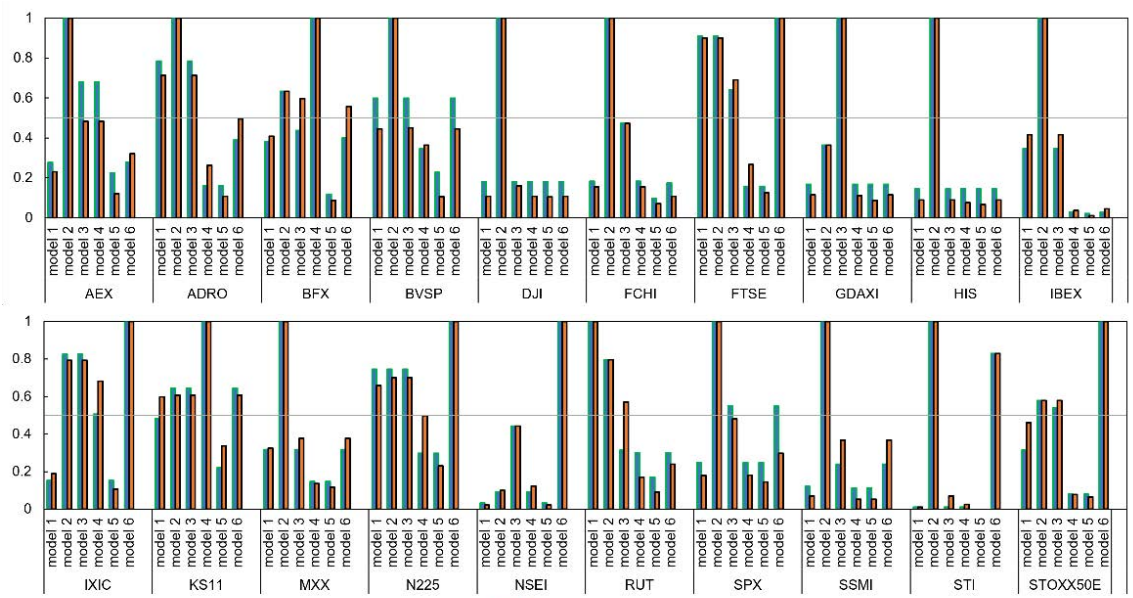
$$\beta = \arg\min_{\beta} \left\{ \begin{array}{c} \sum_{t=p}^{T} (RV_{t+1d}^{(d)} - \beta_0 - \sum_{i=1}^{k} \beta_i RV_t^{(i)})^2 + \\ \lambda [\alpha \sum_{i=1}^{k} |\beta_i| + (1-\alpha) \sum_{i=1}^{k} \beta_i^2] \end{array} \right\} \tag{A.5}$$

The values of α and λ are selected by a 5-fold cross-validation process.

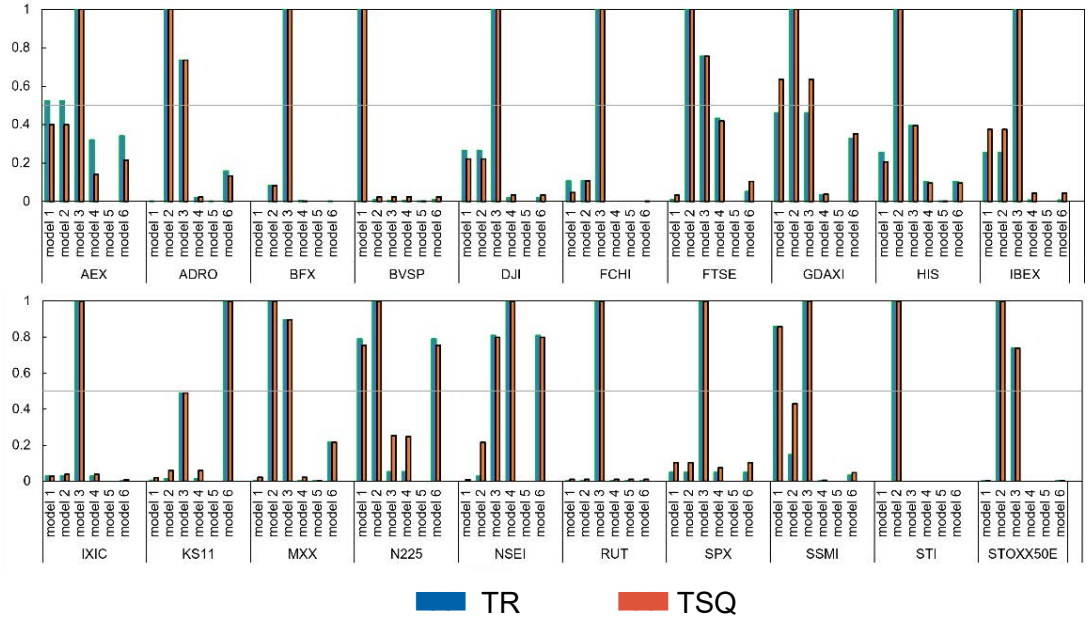
Appendix B

Figure B.1: The results of the MCS test for alternative fixed window sizes under the MSE and QLIKE criterion

(a) MSE criterion



(b) QLIKE criterion

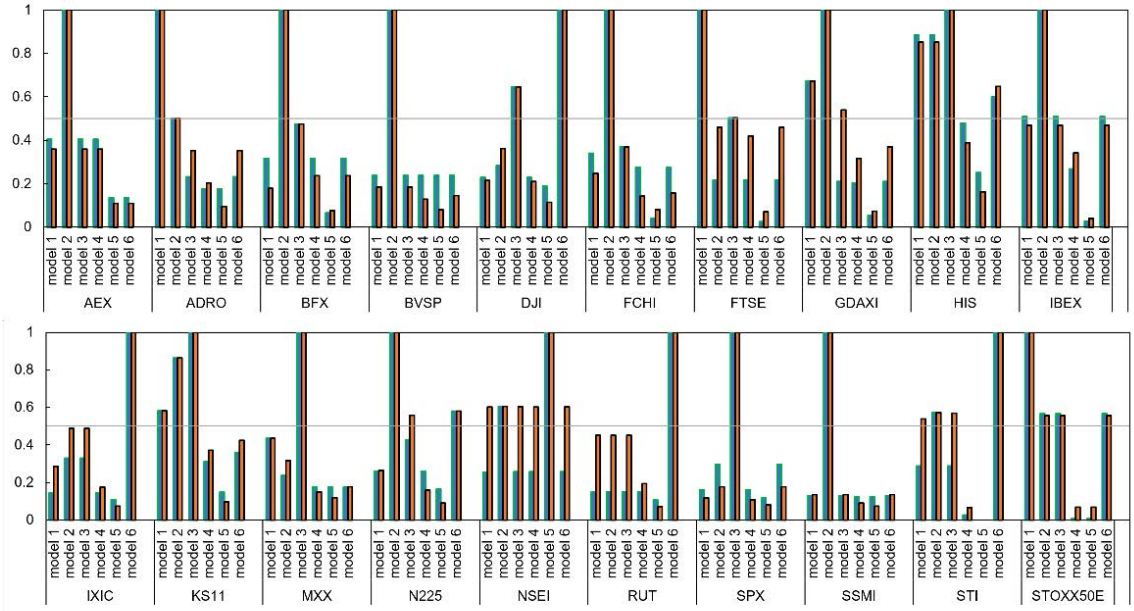


Notes: Fixed window size is set at 2,000. The grey line represents the threshold, which is 0.5. Models 1, 2, 3, 4, 5, and 6 are located on the horizontal coordinates of the figure and correspond to the HAR, HAR-PCA, HAR-sPCA, HAR-Lasso, HAR-Ridge, and HAR-ENet models, respectively. All other conditions are consistent with those outlined in Table 3.

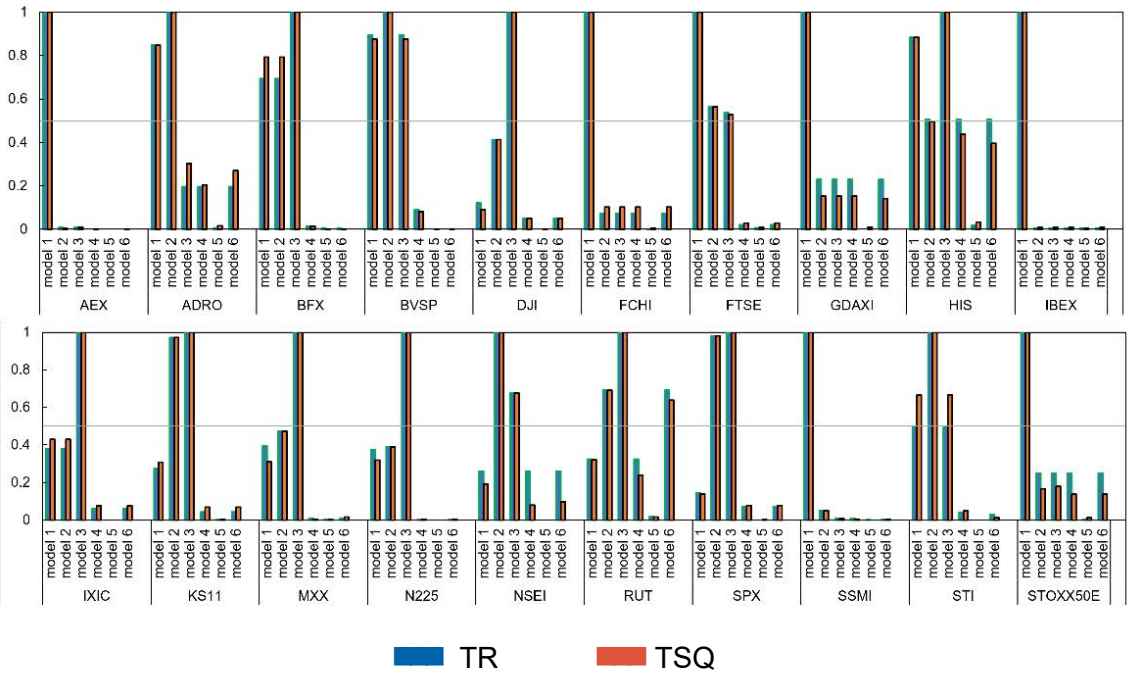
Source: Authors' own elaboration

Figure B.2: The results of the MCS test for alternative maximum lag period under the MSE and QLIKE criterion

(a) MSE criterion



(b) QLIKE criterion

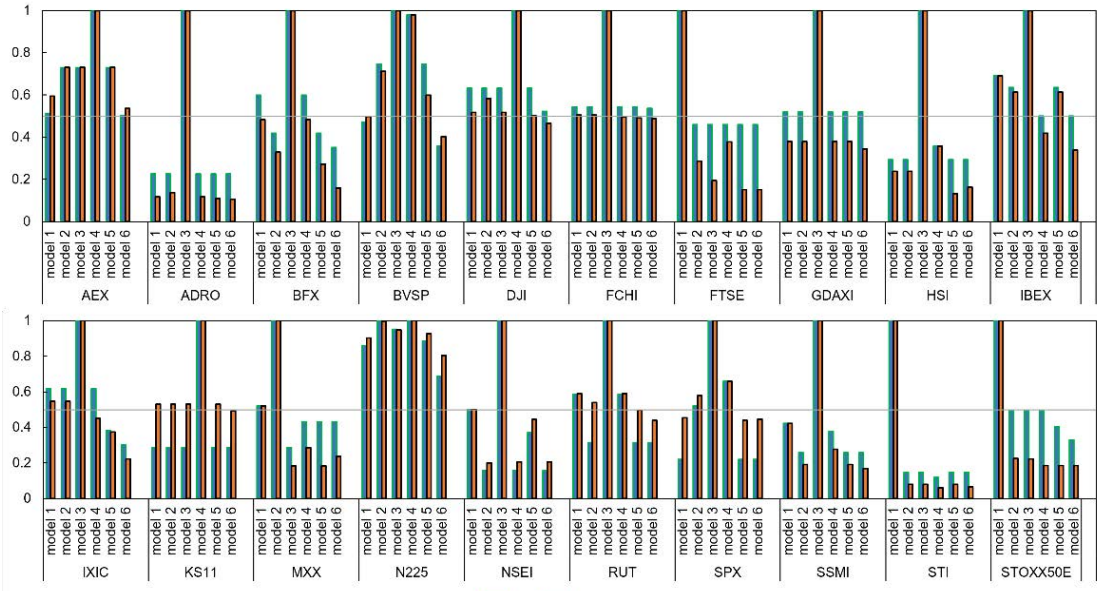


Notes: Maximum lag period is set at 22. The meanings of grey lines, model 1, model 2, model 3, model 4, model 5, and model 6 are consistent with those delineated in Figure B.1. All other conditions are consistent with those outlined in Table 3.

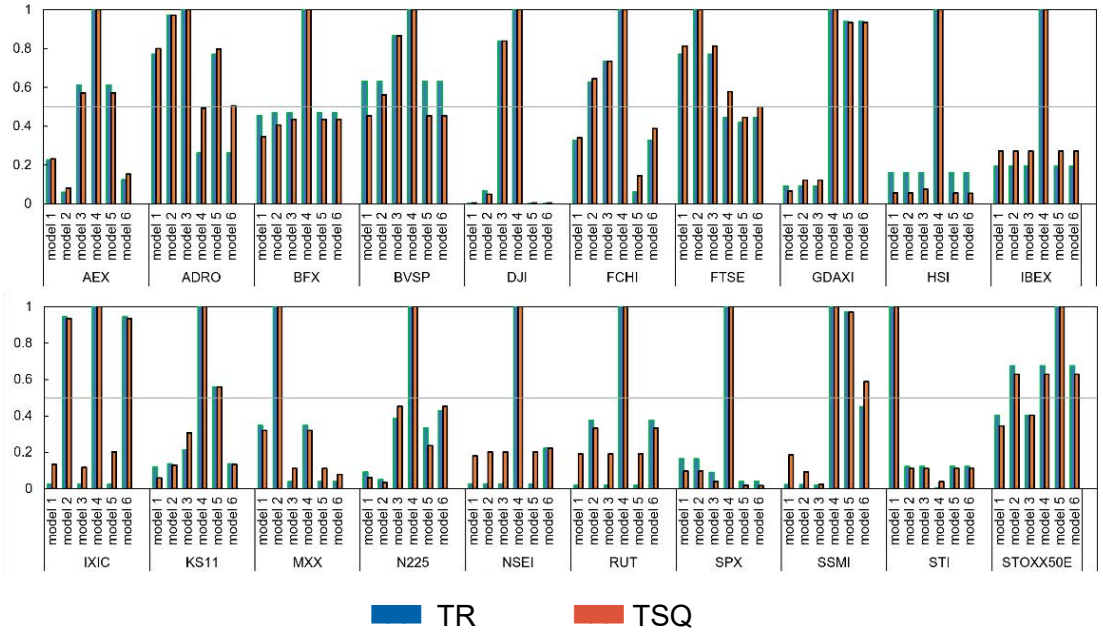
Source: Authors' own elaboration

Figure B.3: The results of the MCS test for different numbers of principal components under the MSE and QLIKE criteria

(a) MSE criterion



(b) QLIKE criterion

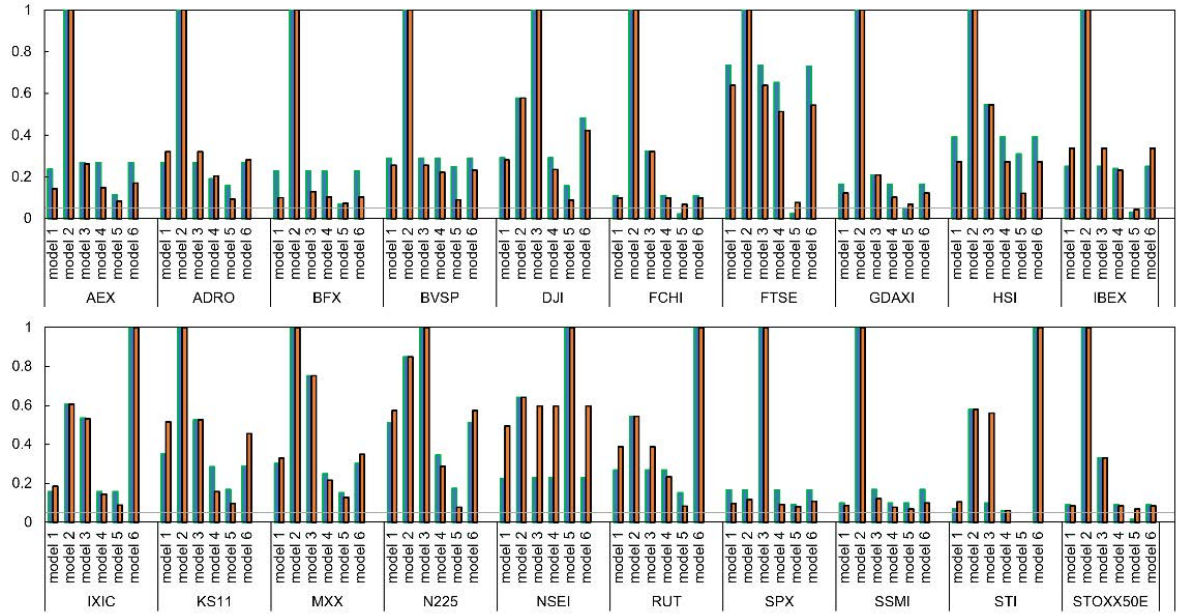


Notes: The number of principal components is set to 3, 4, and 5, respectively. The gray lines indicate the threshold values, which are fixed at 0.5. Models 1 through 6 are positioned along the horizontal axis and correspond to the following configurations: HAR-PCA with 3 principal components, HAR-sPCA with 3 principal components, HAR-PCA with 4 principal components, HAR-sPCA with 4 principal components, HAR-PCA with 5 principal components, and HAR-sPCA with 5 principal components. All other experimental conditions remain consistent with those specified in Table 3.

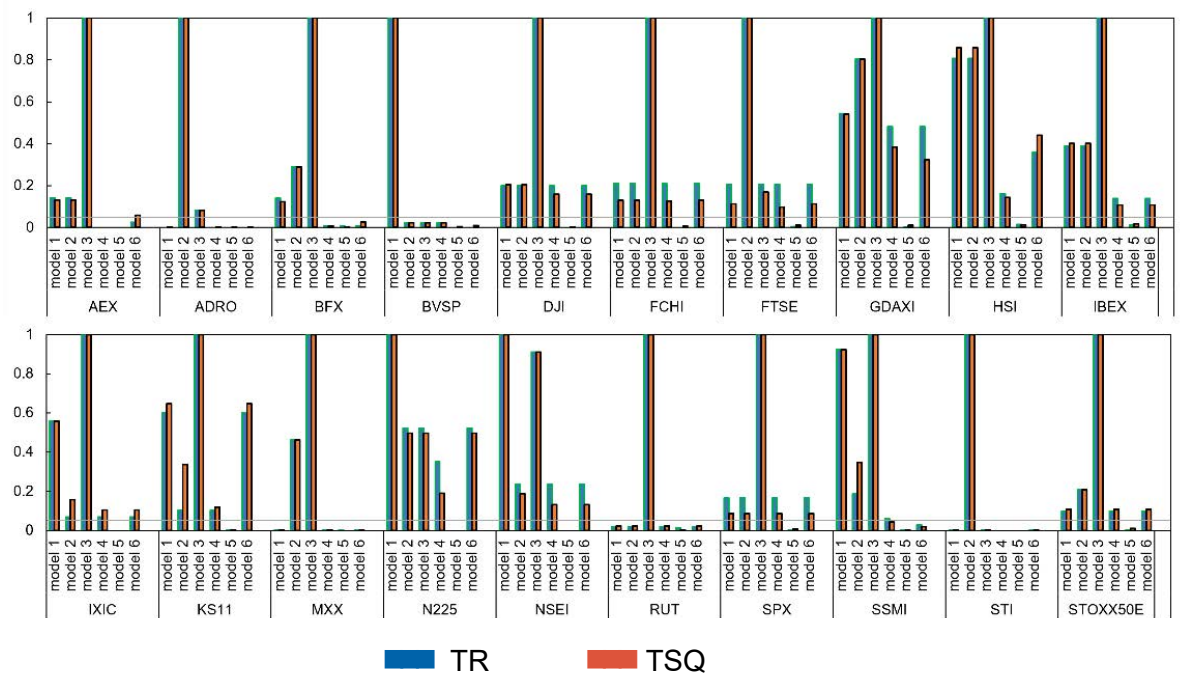
Source: Authors' own elaboration

Figure B.4: The results of the MCS test with a significant alpha of 0.05 under the MSE and QLIKE criteria

(a) MSE criterion



(b) QLIKE criterion

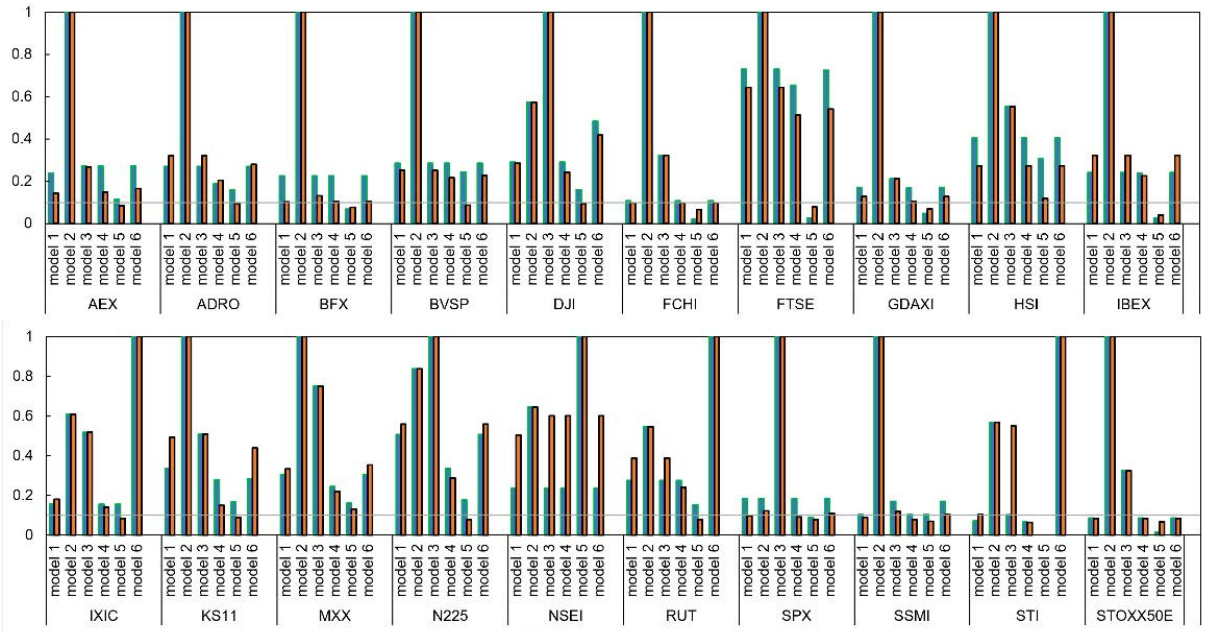


Notes: The significant alpha of the MCS test is set to 0.05, presented with a grey line. Model 1, model 2, model 3, model 4, model 5, and model 6 are consistent with those delineated in Figure B.1. All other conditions are consistent with those outlined in Table 3.

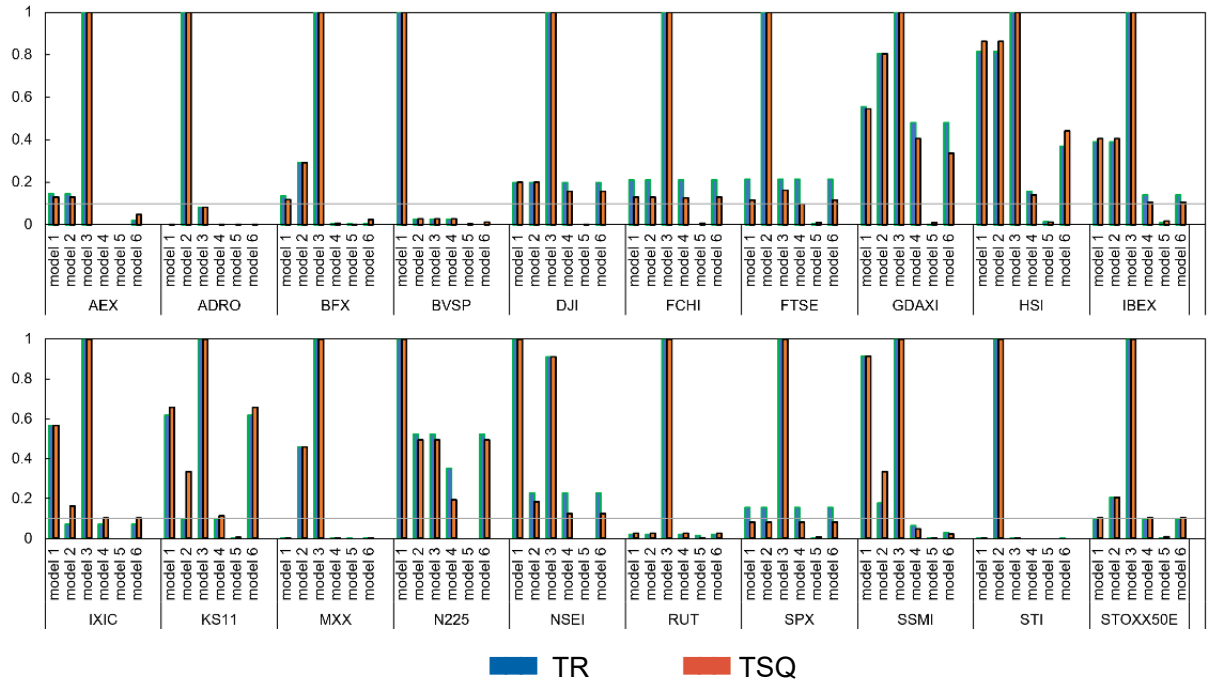
Source: Authors' own elaboration

Figure B.5: The results of the MCS test with a significant alpha of 0.1 under the MSE and QLIKE criteria

(a) MSE criterion



(b) QLIKE criterion

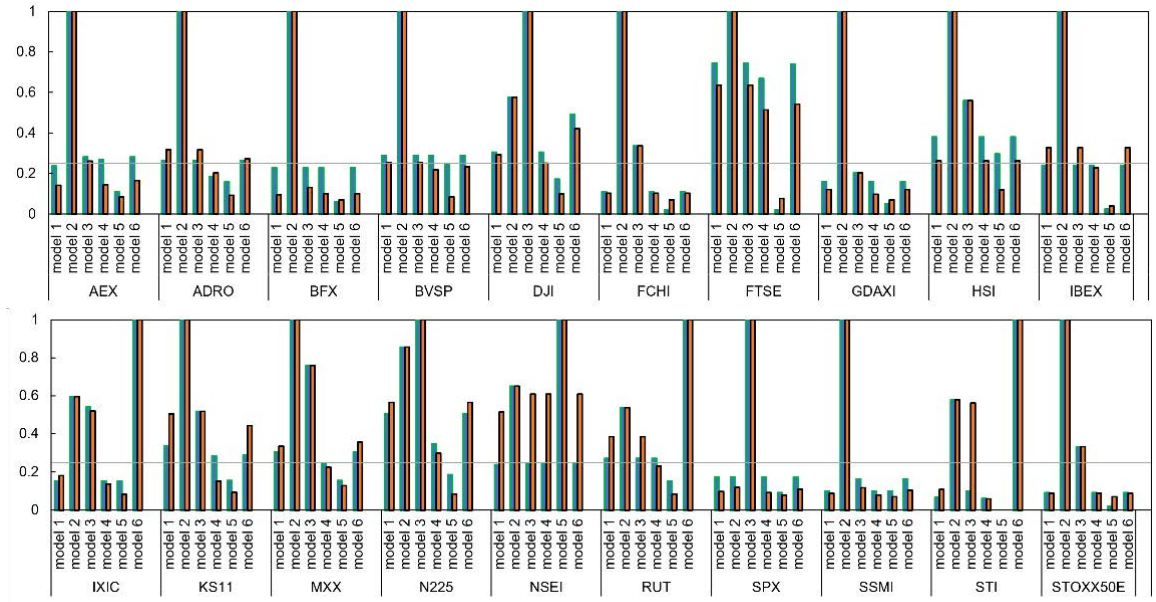


Notes: The significant alpha of the MCS test is set to 0.1, presented with a grey line. Model 1, model 2, model 3, model 4, model 5, and model 6 are consistent with those delineated in Figure B. 1. All other conditions are consistent with those outlined in Table 3.

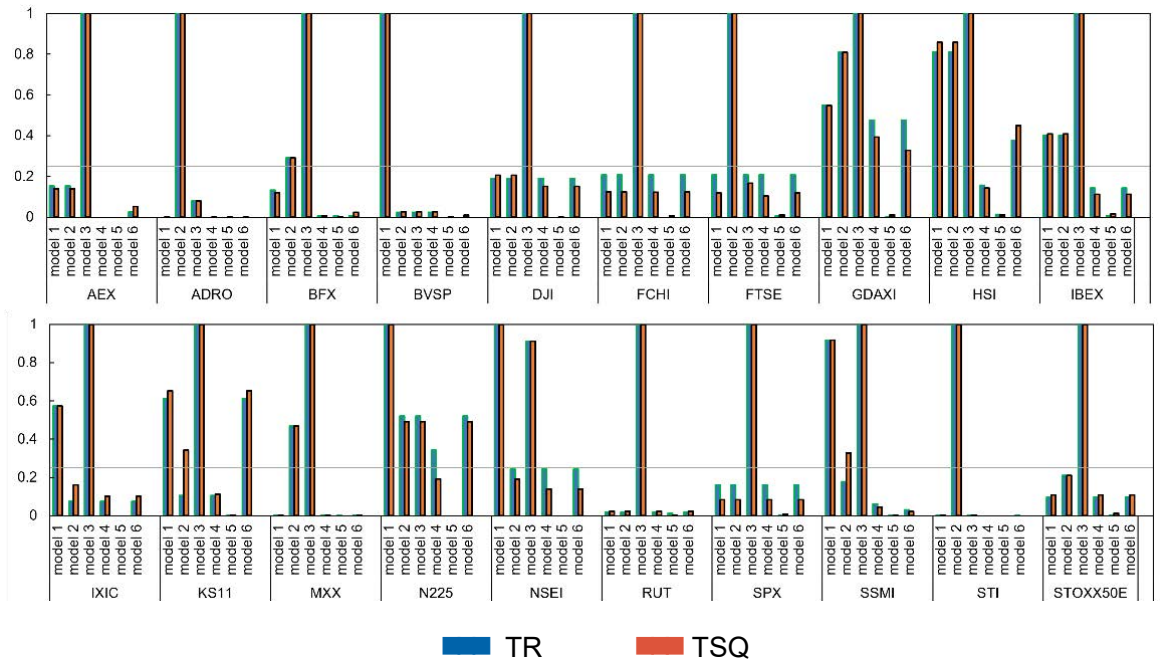
Source: Authors' own elaboration

Figure B.6: The results of the MCS test with a significant alpha of 0.25 under the MSE and QLIKE criteria

(a) MSE criterion



(b) QLIKE criterion



Notes: The significant alpha of the MCS test is set to 0.25, presented with a grey line. Model 1, model 2, model 3, model 4, model 5, and model 6 are consistent with those delineated in Figure B.1. All other conditions are consistent with those outlined in Table 3.

Source: Authors' own elaboration