Volatility Forecasting with Empirical Similarity: Japanese Stock Market Case

Takayuki Morimoto* Yoshinori Kawasaki†

Abstract

In this study, we compare the forecasting ability of various volatility models through within-sample and out-of-sample forecasting simulations. The models considered here are heterogeneous autoregression models (HAR), a 1/3 model where the weight coefficients are all set to 1/3 in the HAR model (ES0), and an HAR model where the weight coefficients are determined by their empirical similarity. We also test AR(1), ARCH/GARCH and their variants, and models incorporating the realized quarticity (RQ), which are referred to as ARQ, HARQ, and ESQ. For stock data, we picked six index series stocks that are listed on the Tokyo Stock Exchange as well as 24 individual stock series. All these stocks had enough liquidity in the market from April 1, 1999, to December 30, 2013, for our investigation. Minute-by-minute data were created based on high-frequency data. Forecasting evaluation depends on what kind of evaluation function we employ. We make use of Patton's error function. By changing the length of estimation period and the forecasting period and the parameter of Patton's error function, we attempt 27,000 forecasting simulations. We find that ESQ and HARQ are almost comparative in within-sample forecasting, whereas ES0 differs in out-of-sample forecasting experiments. We also tried a model comparison based on the pair-wise testing procedure proposed by Hansen et al. We found similar results, but the details are different between the index series and the individual stock series.

Key Words: Empirical similarity, Realized measures, HARQ, ESQ, Model confidence set

1. Introduction

Making inferences based on analogy is one of the basic methods for predicting future events based on experience (Gilboa et al. 2011). Hume (1748) is famous for discussing analogical reasoning, including doubts about the logical validity of inductive reasoning, which is a way to learn from the past about the future. Generally, in an uncertain situation where one has imperfect information, a decision maker cannot evaluate the probability of a future condition, but it is possible to learn from the past about the future and think based on the similarity. More contemporaneously, in the expected utility theory of von Neumann Morgenstern in decision making under uncertainty, decision makers use the state space that enumerates all possible states and their probability distribution. This is assumed to act to maximize the calculated expected utility of analogous thinking. However, there are many situations where it is impossible to assume that decision makers can fully grasp the state space. One way of thinking about such decisions is that people will decide on the action based on analogy from past experiences. This is the case-based decision-making theory advocated by Gilboa and Schmeidler (1995, 2001). Reasoning based on this similarity is widely applied to decision making in medicine, law, business, politics, and artificial intelligence (Gilboa and Schmeidler, 2001). This case-based decision-making theory assumes reasonable consideration of decision makers to evaluate the current situation by considering similarities with a past situation that was experienced (Gilboa and Schmeidler, 2001). Cases that are similar to a current situation are given greater weight than cases that are not very similar. This is the concept of empirical similarity (ES), based on case-based decision-making

*School of Science and Technology, Kwansei Gakuin University
†The Institute of Statistical Mathematics and The Graduate University for Advanced Studies

theory (Gilboa et al., 2006 and Gilboa et al., 2011), and Gilboa and Schmeidler (2012) provided an econometric framework to estimate similarity functions from data. This makes it possible to measure the distance between cases (i.e., problems and situations) recognized by decision makers. In this paper, we use the concept of ES proposed by Golosnoy et al. (2014) to combine prediction values obtained from different models in a non-stochastic way. In this setting, different predicted values obtained from competing models are evaluated by comparing them to the currently observed state or the realized value. A model giving a more accurate predictor of the past is given a larger weight than models with less accurate predictions. The core idea of Golosnoy et al. (2014) is to measure the empirical distance between the currently observed value and the predicted value obtained from different models. With this similarity distance, we can determine the weight of the model for the next period. Therefore, this model combination method, based on ES, uses information on the immediate predictive power of different models to determine the weight of the combination of prediction models. According to Golosnoy et al. (2014), the following three points are possible advantages of using the model combination method based on this ES over other probabilistic methods.

1. There is no need to calculate the posterior probability of the model and the mean square error (MSE) of the predicted value.

2. We can associate weights of predictive models with the preferences of economic agents.

3. We can clarify from the data how the decision maker evaluates the similarity between the predicted value and the realized value.

In the empirical research for this paper, we analyzed the model combination method based on ES as proposed by Golosnoy et al. (2014) by modeling the daily realized volatility process. For this purpose, we evaluated ES against the combination of HAR (heterogeneous autoregressive) models proposed by Corsi (2009), as in the previous work. The HAR model can estimate results in different past investment periods in terms of volatility. The data used for predictive power evaluation in this empirical study are the daily realized volatility obtained from high-frequency data at 1-minute intervals. These data consist of six stock indices over 15 years from January 1999 to December 2013 and 24 individuals listed on the First Section of the Tokyo Stock Exchange. As for the sampling period of data, 225 estimation periods, including in-sample and out-of-sample simulations from 1999 to 2013, are analyzed. These data include 120 in-sample combinations and 105 out-of-sample combinations. By predicting these in-sample and out-of-samples populations, the predictive power of the model combination method based on this ES is compared with a plurality of general volatility models. Regarding the comparison of predictive power, we use the model confidence set (MCS) proposed by Hansen et al. (2011), to evaluate the predictive power appropriately in the framework of statistical hypothesis tests of the error function values of each model obtained in in-sample and out-of-sample simulations. The MCS enables the best model selection at a given significance level without assuming a true model. Finally, we perform a Mincer-Zarnowitz (MZ) regression, which is one of the general methods to evaluate the predictive power proposed by Mincer and Zarnowitz (1969), on the predicted values of each model. We then compare the obtained adjusted coefficient of determination.

The remainder of this paper is organized as follows. Section 2 explains statistical models of ES in detail, which is the theoretical background of this paper. In Section 3, after explaining the data used for empirical analysis, we compare the predictive power of the model by using an MCS and MZ regression. Section 4 summarizes the results of the empirical analysis in this paper and suggests the direction for future research.

2. Theoretical Background

Here, we explain the theoretical background of ES used in this paper based on Gilboa et al. (2011) and Golosnoy et al. (2014).

2.1 ES

We evaluate the value of variable y_t based on a database constituted by the value $x_t = (x_t^1, \ldots, x_t^d)$ of the relevant variable. For example, let's say that y_t is the price of furniture antiques. Here x_t denotes characteristic values such as style, year of manufacture, size, and so on. To properly evaluate y_t, how should we join past observations x_i and current values? If we follow the idea of Hume (1748), we need an idea of similarity that shows whether a past condition $x_i = (x_i^1, \ldots, x_i^d)$ is similar to x_t or not. In predicting y_t, we give higher weight to observations obtained under more similar conditions than observations obtained under less similar conditions. In the above example, it is reasonable to evaluate the price of this antique by the price of similar antiques sold recently. Furthermore, if historical observations are more similar to the current observations with respect to style, year of manufacture, size, and time of sale, then we place higher weight on the observation when evaluating the current situation.

Formally, we assume a similarity function $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++} = (0, \infty)$. Given a database $(x_i, y_i)_{i \le n}$ and a new data point $x_t = (x_t^1, \ldots, x_t^d) \in \mathbb{R}^d$, a similarity predictor of y_t can be formulated as

$$y_t^s = \frac{\sum_{i<t} s(x_i, x_t) y_i}{\sum_{i<t} s(x_i, x_t)}. \tag{1}$$

Alternatively, if the order of data points in $(x_t, y_t)_{t \le n}$ is arbitrary, it can also be defined as

$$y_t^s = \frac{\sum_{i \ne t} s(x_i, x_t) y_i}{\sum_{i \ne t} s(x_i, x_t)}. \tag{2}$$

For the similarity function s, it can be expressed in an arbitrary functional form if several weak assumptions are satisfied (Lieberman, 2010). For example, Billot et al. (2008) give conditions on the similarity weighted average that is equivalent to the similarity function, which has the form:

$$s(x, x') = \exp(-\|x - x'\|),$$

where $\| \cdot \|$ denotes a norm in $\mathbb{R}^d$. Specifically, if focusing on a norm family defined by the weighted Euclidean distance, then we have

$$s_w\left(x, x'\right) = \exp\left(-d_w(x, x')\right),$$

where $w \in \mathbb{R}_+^d$ is the weighted vector of the distance between the two vectors $x, x' \in \mathbb{R}^d$ given by

$$d_w\left(x, x'\right) = \sum_{j=1}^{d} w_j \left(x_j - x_j'\right)^2. \tag{3}$$

Therefore, in this formulation, the similarity function is a d-dimensional vector of parameters including each predictor.

To perform statistical inference and obtain qualitative results by using a hypothesis test, we can incorporate (1) and (2) into the statistical model. That is, we consider the following models as

$$y_t = \frac{\sum_{i<t} s_w(x_i, x_t) y_i}{\sum_{i<t} s_w(x_i, x_t)} + \varepsilon_t, \tag{4}$$

and

$$y_t = \frac{\sum_{i \neq t} s_w(x_i, x_t) y_i}{\sum_{i \neq t} s_w(x_i, x_t)} + \varepsilon_t, \tag{5}$$

where $\{\varepsilon_t\}$ follows iid $\left(0, \sigma^2\right)$. Then, equation (4) can be interpreted as a certain causal model. For example, we consider the price formation process by an economic agent. This economic agent will determine the price of goods such as real estate and art, according to their similarity with other products whose prices have already been determined. Therefore, we can consider equation (4) as a model of the thought process that involves the economic agent in determining the price. However, equation (5) cannot be interpreted directly in a similar way. Since the distribution of each y_t depends on all other y_t, equation (5) cannot explain the temporal evolution of the process. On the other hand, such interdependence can be interpreted naturally in general geography, sociology, or political science data as an application field of spatial statistics.

2.2 The Relationship Between ES and a Kernel Estimator

For simplicity of explanation, we consider the case where X exists as one dimension, that is, as a variable of $d = 1$. In the nonparametric regression model, we normally assume the following data generation process:

$$y_i = m(x_i) + \varepsilon_i, \quad (i = 1, \dots, n), \quad \varepsilon_i \sim \text{iid}\left(0, \sigma^2\right),$$

where $m : \mathbb{R} \to \mathbb{R}$ is an unknown function relating x and y. The widely used nonparametric estimator of $m(\cdot)$ is a Nadalaya-Watson estimator and is defined as follows:

$$\hat{m}(x_t) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{h}\right)},$$

where $K(x)$ is a nonnegative function that satisfies $\int K(z)dz = 1$ as well as other kernel functions (that is, other regular conditions) and h is a bandwidth parameter. For example, if we choose the Gaussian kernel, then we have

$$\frac{1}{h} K\left(\frac{x_i - x_t}{h}\right) = (2\pi h^2)^{-1/2} \exp\left(-\frac{(x_i - x_t)^2}{2h^2}\right). \tag{6}$$

Since there is a trade-off relationship between variance and bias, selection of h is an important issue in nonparametric statistics. One of the most common criteria for choosing the optimal bandwidth is to minimize the mean integral squared error. That is, the optimum h satisfies

$$h^* = \arg\min_h E_{f_0} \int \left(\hat{m}(x) - m(x)\right)^2 dx,$$

where the expected value E_{f_0} means the expected value under f_0 which is the true distribution of y. If x is countable and we substitute $m(x)$ with y, then we decide h^* by the criterion of minimizing the expected value of the sum of squared errors.

Now, we discuss the relationship between estimation based on the kernel and ES. As explained above, the ES method proposes to predict y_t by

$$y_t = \frac{\sum_{i=1}^{n} s_w(x_i, x_t) y_i}{\sum_{i=1}^{n} s_w(x_i, x_t)},$$

where

$$s_w(x_i, x_t) = \exp(-d_w) = (\pi/w)^{1/2} \left[\frac{1}{(1/\sqrt{2w})} K\left(\frac{x_i - x_t}{1/\sqrt{2w}}\right) \right],$$

D_w is defined in equation (3) and K is given in equaiton (6). Finally, we have

$$\frac{\sum_{i=1}^{n} s_w(x_i, x_t) y_i}{\sum_{i=1}^{n} s_w(x_i, x_t)} = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{1/\sqrt{2w}}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{1/\sqrt{2w}}\right)},$$

and this setting results in $h = 1/\sqrt{2w}$.

2.3 ES for Model Combination

Here we assume that there is a finite set of forecasts $x_t = (x_t^1, \ldots, x_t^d)$ obtained from distinct d models which could be combined to predict the variable of interest y_t. According to Bates and Granger (1969), predictive linear combination is given by

$$\hat{y}_t = \sum_{j=1}^{d} a_{t-1}^j x_{t-1}^j, \tag{7}$$

where nonnegative a_t^j represents the ratio of the jth model satisfying $\sum_{j=1}^{d} a_t^j \equiv 1$. The weight a_t^j in equation (7) can be interpreted in relation to the quantitative evaluation (such as probability) of the likelihood of the model or the predicted value. In Elliott and Timmermann (2004), the smallness of the MSE derived from the model corresponds to the weighting factor. Several approaches have been proposed to properly select this weight a_t^j, but none of them can be considered a general method. Within this context, Golosnoy et al. (2014) formulated the linear combination of prediction based on the ES concept by Gilboa et al. (2006) as follows:

$$y_t = \sum_{j=1}^{p} \phi[y_{t-1}, x_{t-2}^j] x_{t-1}^j + \varepsilon_t, \quad \varepsilon_t \sim (0, \sigma^2).$$

The feature of this formulation is that it is possible to measure the distance between the one step ahead predicted value x_{t-2}^j, which is necessary to obtain the weight $\phi[y_{t-1}, x_{t-2}^j]$, and the corresponding realized value y_t. The linear combination of the prediction, which is the weighted sum of the predicted values $x_t = (x_t^1, \ldots, x_t^d)$, is then given by

$$\hat{y}_t = \sum_{j=1}^{d} \phi[y_{t-1}, x_{t-2}^j] x_{t-1}^j.$$

Furthermore, the weight $\phi[\cdot, \cdot]$ depends on the past values of the observed data. The distance between the proxy variable of the current realized value and the predicted value of the jth model is calculated as

$$\phi[y_t, x_{t-1}^j] = \frac{\theta[y_t, x_{t-1}^j]}{\sum_{k=1}^{d} \theta[y_t, x_{t-1}^k]}.$$

The weight $\phi[y_t, x_{t-1}^j] \in [0, 1]$ can be interpreted as a normalized relative ES having the property of $\sum_{k=1}^{d} \phi[y_t, x_{t-1}^k] \equiv 1$. The $\theta[y_t, x_{t-1}^j]$ on the right side denotes a similarity function, and if the distance between y_t and x^j_{t-1} is shorter, then the $\theta[y_t, x^j_{t-1}]$ implies higher similarity. In this paper, we use the exponential function according to Billot et al. (2008) introduced in the previous section as a similarity function:

$$\theta[y_t, x^j_{t-1}] = \exp\left(-\omega_j(y_t - x^j_{t-1})^2\right), \quad \omega_j \in \mathbb{R}.$$

3. Empirical Analysis

The purpose of the empirical analysis is to evaluate the predictive power of the empirical similarity model introduced in the previous section by using it to predict the daily volatility of stock indices and individual stocks listed on the Tokyo Stock Exchange. For this purpose, we use the ticker data of stock indices and individual stocks provided by Nikkei Media Marketing Co., Ltd., as high-frequency data at 1-minute intervals. The sample period is 15 years from January 4, 1999, to December 30, 2013, and the stock indices and individual stocks used are as follows: Regarding the stock price index, we use six series: Nikkei 225, Nikkei 300, TOPIX, TOPIX Electric Appliances Index, TOPIX Transportation Equipment Index, and TOPIX Banks Index. For individual issues, we use 24 stocks continuously traded on the market from 1999 to 2013, which are among the stocks included in the TOPIX Core 30 as of April 1, 2009. We exclude six stocks which were discontinuously traded, including Seven & i HD (1999-2005), JFE-HD (1999-2002), Mitsubishi UFJ-FG (1999-2001), Mitsui Sumitomo FG (1999-2002), Mizuho FG (1999-2003) and Tokio Marine HD (1999-2002). Note that the parentheses indicate the period of deficiency for each series. Table 1 summarizes the six stock indices and 24 individual stocks adopted for the empirical analysis.

The daily time-series data used in the empirical analysis are the stock price, stocks' logarithmic return, realized volatility (RV), and realized quarticity (RQ). In addition, we formulate and analyze the logarithm and square root of RV and RQ using the HAR model described later. However, we cannot find any significant differences in predictive power, so that we omit these results. Next, we discuss two realized measures of these, RV and RQ.

3.1 Realized Measures

Volatility, one of the most common risk indicators in financial markets, is defined as the variance or standard deviation of the logarithmic return. So, far, many models have been proposed to estimate volatility. However, these models are basically parametric and are designed to estimate daily, weekly, and monthly volatilities, using data taken at the same frequency. In recent years, intraday data of financial asset prices have become widely available, and we can use very frequent data recorded every second or minute to calculate the daily volatility *ex post*.

Here, we outline the estimation method of daily volatility using data that have a daily frequency. Following Bollerslev et al. (2016), we consider the financial asset price process P_t as determined by the stochastic differential equation

$$d\log(P_t) = \mu_t dt + \sigma_t dW_t,$$

where μ_t and σ_t represent drift and instantaneous volatility processes, respectively, and W_t is the standard Brownian motion. It is assumed that the model here does not include jumps or abrupt transitions to facilitate an easy understanding. The main objective of this paper is to estimate and predict latent daily volatility, that is, integrated variance (IV). Specifically, the daily IV is formally defined by

$$IV_t = \int_{t-1}^{t} \sigma_s^2 ds.$$

Table 1: Target issues (6 stock indices and 24 individual stocks)

Stock indices	Individual stocks
TOPIX	JAPAN TOBACCO INC.
Nikkei 225	Shin-Etsu Chemical Co.,Ltd.
Nikkei 300	Takeda Pharmaceutical Company
TOPIX Electric Appliances Index	Astellas Pharma Inc.
TOPIX Transportation Equipment Index	FUJIFILM Holdings Corporation
TOPIX Banks Index	NIPPON STEEL CORPORATION
	KOMATSU LTD.
	Hitachi,Ltd.
	Panasonic Corporation
	SONY CORPORATION
	NISSAN MOTOR CO.,LTD.
	TOYOTA MOTOR CORPORATION
	HONDA MOTOR CO.,LTD.
	CANON INC.
	Nintendo Co.,Ltd.
	MITSUI & CO.,LTD.
	Mitsubishi Corporation
	Nomura Holdings, Inc.
	Mitsubishi Estate Company,Limited
	East Japan Railway Company
	NTT CORPORATION
	KDDI CORPORATION
	NTT DoCoMo,Inc.
	The Tokyo Electric Power Company

This IV cannot be observed directly in the financial market, but RV given by the following formula can be calculated as the sum of squares of the intraday high frequent return. S_t, $r_t = \log S_t - \log S_{t-1}$ and RV_t denote the stock price at time $t = 1, 2, \ldots, T$, its logarithmic return, and realized volatility, respectively. For RV, it is defined as the sum of squares of the intraday return sampled at 1-minute intervals as

$$RV_t = \sum_{i=1}^{n_t} r_{t,i}^2,$$

where $r_{t,i}$ is the ith observed logarithmic return on the t day and n_t represents the number of samples on day t (see, for example, Andersen et al., 2001). By considering the log-valued price process as a continuous martingale part of a semi-martingale, we can regard this RV as a proxy variable of IV. Also, it is known that RV is a consistent and unbiased estimator of IV (McAleer and Medeiros, 2008). Consequently, the RV estimation requires full high-frequency data over 24 hours as a daily volatility measure. However, the Japanese stock market is divided into two sessions by a lunch break, that is, the morning session goes from 09:00 to 11:00 (until 11:30 after 21 November 2011) and the afternoon session goes from 12:30 to 15:00. Thus, we adopt the weighted RV proposed by Masuda and Morimoto (2012), which is a modified version of Hansen and Lunde (2005) and which is adjusted to the Japanese market. The weighted RV with estimated optimal weights λ_1,λ_2,λ_3, and λ_4 is defined by

$$RV_t^{weighted} = \lambda_1 Y_{t,1}^2 + \lambda_2 RV_{t,2} + \lambda_3 Y_{t,3}^2 + \lambda_4 RV_{t,4},$$

where $Y_{t,1}^2$, $RV_{t,2}$, $Y_{t,3}^2$, and $RV_{t,4}$ denote the square of the close-to-open return, the RV in the morning session, the square of the lunch break return, and the RV in the afternoon session, respectively, on the tth day. Hereafter, we denote the weighted realized volatility $RV^{weighted}$ as RV for notational simplicity.

Furthermore, it is known that RV is influenced by an observational error caused by the microstructure noise when using the intraday data sampled at high frequencies. As a method for mitigating this bias, it is conceivable to use RV with a low-frequency interval (see Andersen and Bollerslev, 1997 or Bandi and Russell, 2008), a subsample method (see Zhang et al., 2005), and a kernel method. Among these, in this paper, we use the Newey-West (NW) estimator using the Bartlett kernel following Hansen and Lunde (2005). It is shown by Bamdorff-Nielsen et al. (2008) that this NW estimator is almost the same as the above-described subsample method.

According to Bollerslev et al. (2016), the resulting estimation error in RV may be characterized by the asymptotic (for $\Delta \to 0$) distribution theory of Barndorff-Nielsen and Shephard (2002),

$$RV_t = IV_t + \eta_t, \quad \eta_t \sim MN(0, 2\Delta IQ_t), \tag{8}$$

where $IQ_t \equiv \int_{t-1}^{t} \sigma_s^4 ds$ denotes the integrated quarticity (IQ), and MN represents a mixed normal distribution: that is, a normal distribution that is conditional on the realization of IQ_t. In parallel, for the IV, the IQ may be consistently estimated by the RQ,

$$RQ_t \equiv \frac{M}{3} \sum_{i=1}^{M} r_{t,i}^4.$$

However, since the estimation of IQ includes the estimation of the 4th moment of the intraday return with many noises such as measurement errors, even if the magnitude of the jump is small, the RQ estimator is inevitably unstable. Thus, for example, to reduce even the influence of the jumps, Andersen et al. (2012) propose two robust estimates of IQ called $MinRQ$ and $MedRQ$ which employ the minimum or median of each adjacent return. In this paper, we use the NW estimator that uses the Bartlett kernel for RQ estimation in the same way as for RV estimation to ease the calculation burden.

We select an index and individual stock prices from the above 30 issues and give an overview of the characteristics of these time series because of limited space. Figures 1 and 2 show the stock price, logarithmic return, RV, and RQ for TOPIX and Hitachi over 15 years from January 1999 to December 2013. From these figures, we can see that there are three major peaks in the latter half of the period. It is especially easy to capture the peaks when focusing on the panel of realized volatility. These substantial fluctuations correspond to the financial crisis of September 2008, the Tohoku district Pacific offshore earthquake that occurred in March 2011, and the Nikkei Average major crash on May 23, 2013, respectively.

In addition, Tables 2 and 3 provide the descriptive statistics of the stock price, logarithmic return, RV, and squared RQ, which range 15 years from January 1999 to December 2013 for TOPIX and Hitachi. From these tables, we can see that the kurtosis of the logarithmic return r_t is 3 or more, which is one of the stylized facts for the financial time series. The skewness for both is negative, which is an interesting result, as the peaks of the return distribution are biased to the right, that is, in the positive direction. But, both of the skews are negative, which means that the peak of the return distribution is biased to the right, that is, in the positive direction. This seems to be an interesting result. In both cases, the absolute value of the maximum value and the minimum value exceeds 10%, and the averages are almost equal to 0.

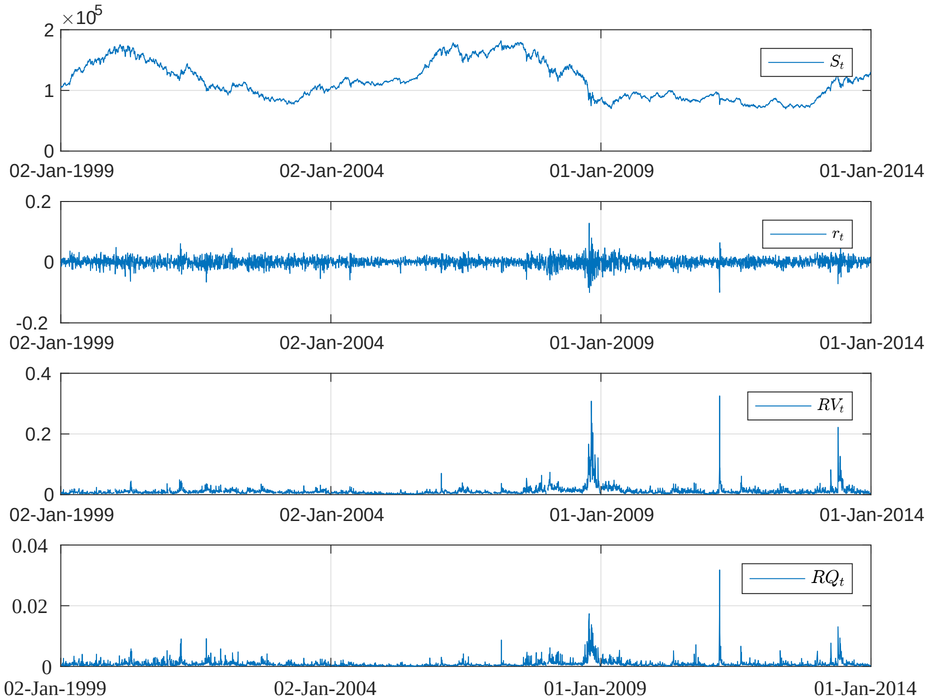

Figure 1: Stock price, logarithmic return, RV and RQ (TOPIX)

3.2 Models

In this subsection, we introduce 17 time-series models used for empirical analysis. The HAR model proposed by Corsi (2009) combines volatility measures sampled at different frequencies in a simple linear regression framework. The standard HAR model for the daily volatility process v_t is given by

$$v_t = \alpha_0 + \omega_1 v_{t-1}^{(d)} + \omega_2 v_{t-1}^{(w)} + \omega_3 v_{t-1}^{(m)} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2),$$

where $v_{t-1}^{(d)} = v_{t-1}$, $v_{t-1}^{(w)}$ and $v_{t-1}^{(m)}$ are daily, weekly and monthly average volatility measures, respectively. These are defined as

$$v_t^{(w)} = 5^{-1} \sum_{i=1}^{5} v_{t-i+1} \text{ and } v_t^{(m)} = 22^{-1} \sum_{i=1}^{22} v_{t-i+1}.$$

By substituting unobservable v_t by rv_t, we can estimate the HAR model with the framework of ordinary least squares (OLS) regression. According to Golosnoy et al. (2014), the economic interpretation of volatility components relate the long-term component $v^{(m)}$ to the fundamental macroeconomic uncertainty factors. The medium-term component $v^{(w)}$ reflects the current market uncertainty concerning the processing of news, and the short-term component $v^{(d)}$ accounts for the speculative momentum uncertainty.

Next, we introduce the model in which the constant term of the HAR model is 0 and the other three parameters are fixed to the 1/3 value as follows:

$$v_t = \frac{1}{3} v_{t-1}^{(d)} + \frac{1}{3} v_{t-1}^{(w)} + \frac{1}{3} v_{t-1}^{(m)} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2),$$

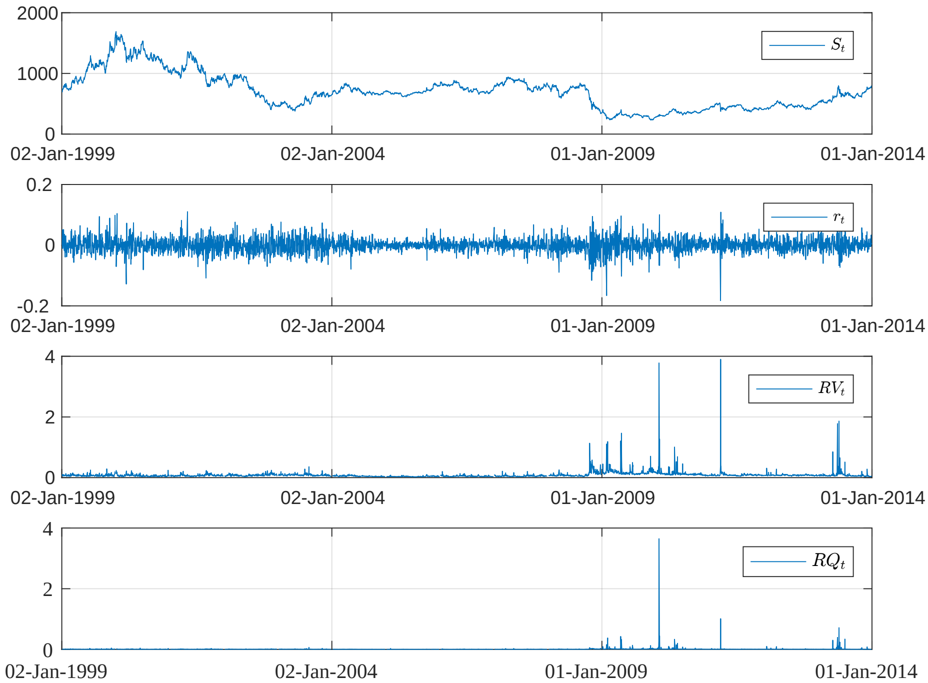

Figure 2: Stock price, logarithmic return, RV and RQ (Hitachi)

In this paper, we refer to this model as the ES0 model, whereas it is called the "1/3 model" in Golosnoy et al. (2014). Although this ES0 model does not directly use empirical similarity, we can regard this model as a special case of the ES1 model as described below. That is, if we put the parameter of the ES1 model as $\theta[v_{t-1}, v_{t-2}] = \theta[v_{t-1}, v_{t-2}^{(w)}] = \theta[v_{t-1}, v_{t-2}^{(m)}] = 1/3$, then the ES1 model is reduced to ES0. Since all the parameters of the ES0 model are constantly 1/3 as described above, there is no need to estimate the parameter values from the data. Then, the ES0 model appears only at the stage of predictive power evaluation later.

The third model is the empirical similarity model ES1, which plays a central role in this paper. The ES1 model was derived from the concept of empirical similarity estimates for the weight for the predictors of volatility model components, based on the past data observed directly. If the predictor of volatility obtained from model h belonging to the model set $\mathcal{H}$ can be denoted by $v_t^{(h)}$, then the ES model for volatility prediction is given by

$$v_t = \sum_{h\in\mathcal{H}} \phi[v_{t-1}, v_{t-2}^{(h)}] \cdot v_{t-1}^{(h)} + \varepsilon_t = \frac{\sum_{h\in\mathcal{H}} \theta[v_{t-1}, v_{t-2}^{(h)}] \cdot v_{t-1}^{(h)}}{\sum_{h\in\mathcal{H}} \theta[v_{t-1}, v_{t-2}^{(h)}]} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2),$$

where $\sum_{h\in\mathcal{H}} \phi\left[v_{t-1}, v_{t-2}^{(h)}\right] \equiv 1$. The similarity function is defined by $\theta\left[v_t, v_{t-1}^{(h)}\right] = e^{-w_h\left(v_t - v_{t-1}^{(h)}\right)^2}$ and the function measures the distance between the current volatility state v_t and the hth model's predictor $v_t^{(h)}$. Thus, we can predict v_{t+1} by using the weight $\phi\left[v_t, v_{t-1}^{(h)}\right] \in [0, 1]$ and the predictor $v_t^{(h)}$ of the model.

As in Golosnoy et al. (2014), we use the HAR model as a benchmark in the study so that we focus on combining the three components with ES models. Our objective is

Table 2: Descriptive statistics (TOPIX) *1999-2013*

	S_t	r_t	RV_t	$\sqrt{RQ}_t$
Mean	115815.4168	0.0001	0.0114	0.0008
Median	111634.0000	0.0003	0.0080	0.0005
Max.	181697.0000	0.1286	0.3255	0.0318
Min.	69551.0000	−0.1001	0.0012	0.0000
SD	30461.9413	0.0141	0.0155	0.0013
Ske.	0.4294	−0.3543	9.7217	8.9328
Kur.	1.9932	8.8653	144.9977	143.5033
Size	3685	3684	3664	3664

Table 3: Descriptive statistics (Hitachi) *1999-2013*

	S_t	r_t	RV_t	$\sqrt{RQ}_t$
Mean	701.3720	0.0000	0.0884	0.0095
Median	683.0000	0.0000	0.0686	0.0051
Max.	1690.0000	0.1105	3.9003	3.6477
Min.	233.0000	−0.1827	0.0203	0.0015
SD	286.7948	0.0233	0.1249	0.0668
Ske.	0.7732	−0.1371	18.2170	45.8112
Kur.	3.3460	6.7904	481.5358	2424.5612
Size	3685	3684	3663	3663

to evaluate how the relative distance between the current volatility and the weighted sum of volatility sampled at different time periods is determined from historical data. In other words, we would like to analyze how the economic agents with different investment periods evaluate the weights of these volatility processes using ES. The empirical similarity model having the HAR component, represented hereafter as the ES1 model, is given as

$$v_t = \frac{\theta[v_{t-1}, v_{t-2}]v_{t-1} + \theta[v_{t-1}, v_{t-2}^{(w)}]v_{t-1}^{(w)} + \theta[v_{t-1}, v_{t-2}^{(m)}]v_{t-1}^{(m)}}{\theta[v_{t-1}, v_{t-2}] + \theta[v_{t-1}, v_{t-2}^{(w)}] + \theta[v_{t-1}, v_{t-2}^{(m)}]} + \epsilon_t, \quad \epsilon_t \sim (0, \sigma^2), \tag{9}$$

where

$$\begin{aligned}
\theta[v_{t-1}, v_{t-2}] &= \exp\left(-\omega_1(v_{t-1} - v_{t-2})^2\right), \\
\theta[v_{t-1}, v_{t-2}^{(w)}] &= \exp\left(-\omega_2(v_{t-1} - v_{t-2}^{(w)})^2\right), \\
\theta[v_{t-1}, v_{t-2}^{(m)}] &= \exp\left(-\omega_3(v_{t-1} - v_{t-2}^{(m)})^2\right).
\end{aligned}$$

The ES1 model can be interpreted as a combination of predictive models, assuming a simple weighted average of volatilities sampled at different frequencies. Component v_{t-1} is a predictor obtained from the volatility on the previous day, whereas $v_{t-1}^{(w)}$ and $v_{t-1}^{(m)}$ are predictors of the moving average in the previous 1 week and 1 month, respectively. Consequently, the daily volatility v_t in the equation (9) is expressed as a weighted average of past daily realized volatilities. As is apparent from the equation (9), the ES1 model is characterized by having one parameter less than the HAR model; that is, there is no constant term.

Next, we introduce three models that incorporate the RQ that introduced in the previous section. Bollerslev et al. (2016) suppose that the dynamic dependencies in IV may be described by an autoregressive (AR) model of order 1,

$$IV_t = \phi_0 + \phi_1 IV_{t-1} + u_t, \;\; u_t \sim \text{iid}(0, \sigma_u^2).$$

Let $\eta_t \sim \text{iid}(0, \sigma_\eta^2)$ be a measurement error of IV_t. Thus, a simple AR (1) model for IV_t incorporating η_t is given by

$$IV_t + \eta_t = \beta_0 + \beta_1(IV_{t-1} + \eta_{t-1}) + u_t. \tag{10}$$

The formal theoretic justification for applying the autoregressive model to RV is given by Andersen et al. (2003). Andersen et al. (2004) also show that the predictive power of IV can be significantly improved by using a simple discrete-time autoregressive model rather than a continuous time-based model for RV. If we assume that u_t and η_t are both $i.i.d.$, so that $\text{Cov}(RV_t, RV_{t-1}) = \phi_1 \text{Var}(IV_t)$ and $\text{Var}(RV_t) = \text{Var}(IV_t) + 2\Delta IQ$, then, we have

$$\beta_1 = \phi_1 \left(1 + \frac{2\Delta IQ}{\text{Var}(IV_t)}\right)^{-1}. \tag{11}$$

Therefore, the coefficient β_1 of RV is smaller than the coefficient ϕ_1 of the IV, due to the so-called attenuation bias. For details of the attenuation bias, for example, see Wooldridge (2015). From equation (11), β_1 varies depending on the variance $2\Delta IQ$ of the measurement error. That is, if $2\Delta IQ = 0$, then $\beta_1 = \phi_1$, but if $2\Delta IQ$ is large, then β_1 goes to zero. In general, β_1 in equation (11) assumes that the variance of the measurement error is a constant. In practice, however, the variance with respect to an estimation error of RV changes through time in practice. On days when IQ is small, RV has a higher predictive power for IV, and conversely, RV has relatively weak predictive power for IV on days with a large IQ. Therefore, it is more realistic to assume an autoregressive coefficient that changes through time such as $\beta_{1,t}$, rather than assuming that the coefficient of AR is constant.

Equation (10) can be viewed as an AR(1) model for RV as $RV_t = \beta_0 + \beta_1 RV_{t-1} + u_t$ from the relationship of $RV_t = IV_t + \eta_t$ in equation (8). Bollerslev et al. (2016) implement a more flexible and robust specification that allows the time-varying AR parameter to depend linearly on an estimate of $IQ^{1/2}$ as

$$RV_t = \beta_0 + \underbrace{(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2})}_{\beta_{1,t}} RV_{t-1} + u_t.$$

The specification is called the ARQ model. The model can easily be estimated using the standard OLS method, rendering both estimating and forecasting straightforward and fast. Importantly, the value of the autoregressive $\beta_{1,t}$ parameter varies with the estimated measurement error variance. If RQ is constant over time, the ARQ model reduces to a standard AR(1) model; see Bollerslev et al. (2016) for a more detailed description of the ARQ model.

Bollerslev et al. (2016) consider that the AR(1) model in equation (10) is too simplistic to satisfactorily describe the long-run dependencies in most RV series. Instead, the heterogeneous autoregression (HAR) model of Corsi (2009) has arguably emerged as the most popular model for daily RV based forecasting,

$$RV_t = \beta_0 + \underbrace{(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2})}_{\beta_{1,t}} RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t.$$

The specification is called the HARQ model. Here, the coefficient of the daily RV only changes through time as a function of $RQ^{1/2}$. For models that include time varying coefficients of weekly and monthly RV, see Bollerslev et al. (2016). Further, we can also add

$RQ^{1/2}_{t-1|t-5}, RQ^{1/2}_{t-1|t-22}$ to the explanatory variable of the HARQ model above as a natural extension. However, Bollerslev et al. (2016) report that the prediction power of the model referred to as HARQ-Full does not improve unconditionally compared with the HARQ model above, since it is practically difficult to accurately estimate weekly and monthly variances of the measurement errors. Based on the results, we omit the analysis of HARQ-Full model in the paper.

According to Golosnoy et al. (2014), the ES model can be used for any combination of volatility predictors. Therefore, it is also possible to consider a new model combining the past daily volatility v_{t-1}, the HAR predictor $v^{(har)}_{t-1}$, and the HARQ predictor $v^{(harq)}_{t-1}$. We call this model the "ESQ" model here. The specification of the model is given by

$$v_t = \frac{\theta[v_{t-1}, v_{t-2}]v_{t-1} + \theta[v_{t-1}, v^{(har)}_{t-2}]v^{(har)}_{t-1} + \theta[v_{t-1}, v^{(harq)}_{t-2}]v^{(harq)}_{t-1}}{\theta[v_{t-1}, v_{t-2}] + \theta[v_{t-1}, v^{(har)}_{t-2}] + \theta[v_{t-1}, v^{(harq)}_{t-2}]} + \epsilon_t, \ \ \epsilon_t \sim (0, \sigma^2),$$

and as previously defined, we have

$$\theta[v_{t-1}, v_{t-2}] = \exp\left(-\omega_1(v_{t-1} - v_{t-2})^2\right),$$
$$\theta[v_{t-1}, v^{(har)}_{t-2}] = \exp\left(-\omega_2(v_{t-1} - v^{(har)}_{t-2})^2\right),$$
$$\theta[v_{t-1}, v^{(harq)}_{t-2}] = \exp\left(-\omega_3(v_{t-1} - v^{(harq)}_{t-2})^2\right).$$

In addition, we also analyze the following models from the viewpoint of consistency in the model comparison. First, we introduce a model simply combining daily volatility v_t and the HAR predictor $v^{(har)}_t$,

$$v_t = \frac{\theta[v_{t-1}, v_{t-2}]v_{t-1} + \theta[v_{t-1}, v^{(har)}_{t-2}]v^{(har)}_{t-1}}{\theta[v_{t-1}, v_{t-2}] + \theta[v_{t-1}, v^{(har)}_{t-2}]} + \epsilon_t.$$

The second is a model that combines $v_t, v^{(w)}_t, v^{(m)}_t, q_t = RQ^{1/2}_t RV_t$ as it is without modeling,

$$v_t = \frac{\theta[v_{t-1}, v_{t-2}]v_{t-1} + \theta[v_{t-1}, v^{(w)}_{t-2}]v^{(w)}_{t-1} + \theta[v_{t-1}, v^{(m)}_{t-2}]v^{(m)}_{t-1} + \theta[v_{t-1}, q_{t-2}]q_{t-1}}{\theta[v_{t-1}, v_{t-2}] + \theta[v_{t-1}, v^{(w)}_{t-2}] + \theta[v_{t-1}, v^{(m)}_{t-2}] + \theta[v_{t-1}, q_{t-2}]} + \epsilon_t,$$

where

$$\theta[v_{t-1}, q_{t-2}] = \exp\left(-\omega_4(v_{t-1} - q_{t-2})^2\right).$$

We refer to these models as ES1a and ES1b, respectively.

We introduce a simple AR (1) model

$$v_t = \alpha_0 + \omega_1 v_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2),$$

and eight types of GARCH models as benchmarks of predictive power, in addition to the five models that are the center of the analysis, of which the GARCH models are GARCH(1,1), GJR(1,1,1), EGARCH(1,1,1), IGARCH(1,1), AGARCH(1,1), NAGARCH(1,1), APARCH(1,1,1), and ZARCH(1,1,1). However, the eight types of GARCH models are overwhelmingly disadvantageous compared with other AR1, HAR, ES0, ES1, ES1a, ES1b, ARQ, HARQ, and ESQ models in the predictive power comparison, since the GARCH models estimate and predict volatility by not directly using RV_t series. Therefore, we note that comparison of predictive power within the GARCH models can be useful information, but the results of comparative study between the eight types of GARCH models and the other nine models are just for reference.

3.3 Estimation

In this section, we estimate and predict volatility using stock price data for six stock indices and 24 individual stocks introduced in the previous section. The sample period and indices for these stocks range from January 4, 1999, to December 30, 2013. Table 4 shows the periods used for estimation and prediction. As can be seen from the table, periods for estimation and prediction sum up to 225 including in-sample and out-of-sample simulations from 1999 to 2013. In the table, we omit the first two digits of year because of limited space. The breakdown of the 225 combinations comprises 120 in-samples and 105 out-of-samples. Furthermore, we divide the 120 estimation periods of in-samples into 15 for 1 year, 50 for more than 2 years but less than 5 years, and 55 for more than 5 years but less than 15 years. Likewise, we divide the 105 estimation periods of out-of-sample simulations into 14 for 1 year, 46 for more than 2 but less than 5 years and 45 for more than 5 years but less than 14 years. Note that the prediction period of the out-of-sample is set as the most recent one year of the estimation period.

When outliers exist in the estimation period, nonlinear models such as ES1, ES1a, ES1b, and ESQ are susceptible to the influence of the outliers. Therefore, we apply Cook's distance to the RV series to detect and exclude the outliers, so that we use data without outliers when estimating the models. The rate of outlier detection is approximately 5% (with a minimum value of 3.2% and a maximum value of 6.6%), although there exists difference across series. For the definition of Cook's distance, see, for example, Weisberg (2014).

We use the Statistics and Machine Learning Toolbox of MATLAB for parameter estimation of the ES1, ES1a, ES1b, HAR, AR1, ARQ, HARQ, and ESQ models and the MFE Toolbox of Prof. Kevin Sheppard for the eight types of GARCH models.

3.4 Prediction

In this section, we compare and analyze the predictive power of empirical similarity models and other time-series models in terms of volatility, which is the main contribution of the paper. To do this, we estimate the parameters of 16 models, excluding the ES0 model, using time-series data of 30 stocks in the estimation period shown in Table 4. Then, we compare the predictive power between models using the error function described below in the corresponding prediction period of Table 4. Note that we only list the model ranking by MCS and the result of the MZ regression for the ES1a and ES1b models because of limited space, and the MCS and MZ are described below.

With respect to the predictors of the out-of-sample, for example, when the estimation and prediction periods are 99-99 and 00-00 respectively, we estimate parameters using the data for 1999 and sequentially make predictions for 2000. However, we do not perform so-called rolling window prediction, which estimates and predicts parameters daily. There are two reasons for that:

1. First, the number of calculations is simply too large. The objective of the study is to compare the predictive power of the models in terms of volatility against various (i.e., 30) stocks and multiple models (i.e., 17) over long and short periods of time (the estimation and prediction periods number 225). Therefore, it is not practical to estimate parameters and execute 1-day ahead prediction day by day for 114,750 ($30 \times 17 \times 225$) combinations if we consider its calculation time.

2. Again, the objective of the study is not to compare the strict predictive power between models, but to rank models for a wide variety of data as well as estimation and prediction periods. Therefore, when considering cost effectiveness, it is more

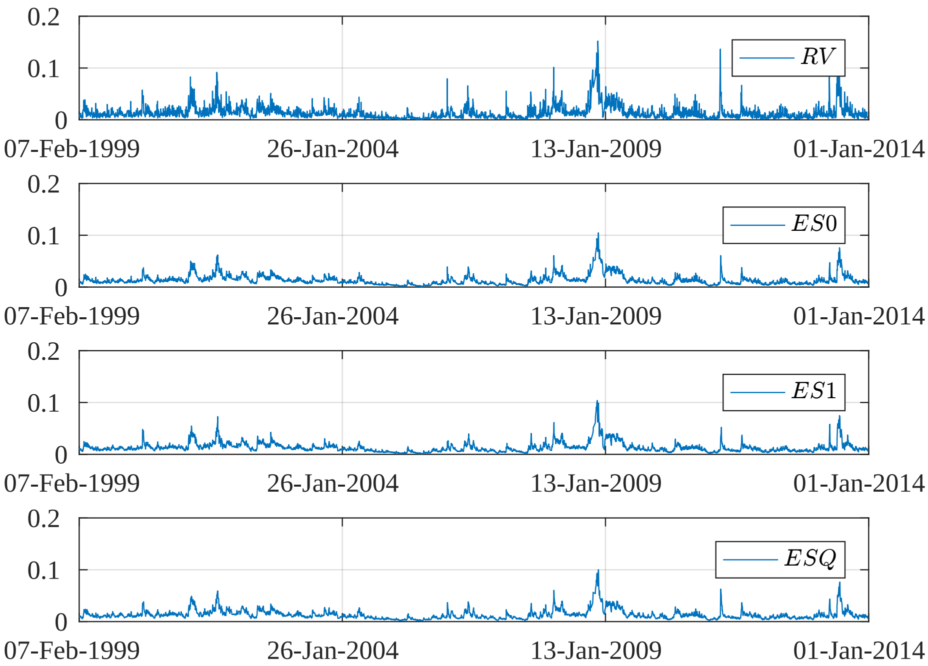

Figure 3: RV and predictors obtained from models (Nikkei 225)

efficient to fix the parameter estimation and perform a 1-day ahead prediction than it is to execute the rolling window prediction, which requires much calculation time.

Figures 3 and 4 show in-sample RVs for Nikkei 225 and KDDI as well as predictors of volatility calculated from the ES0, ES1, HAR, HARQ, and ESQ models from February 1999 to December 2013. The reason that the plots begin in February 1999 is because we exclude the predictors for the first 22 days because the HAR model requires the average of the volatility over the latest 22 days to make a forecast. First, looking at Figure 3, we can see that significant fluctuations are noticeable in the second half of 2008 due to the financial crisis of 2007-2008. Comparing RV over the fluctuations with the predictor calculated from each model, we can see that the ES0, ES1, HAR, HARQ and ESQ models underestimate the volatility. Overall, we can see that the predictor from each model is lower than RV even in periods other than the one during which the financial crisis occurred. Second, looking at Figure 4, we can see that the behavior of volatility significantly differs between the Nikkei 225 and KDDI. Especially, long-term volatility clustering for KDDI is prominent from 1999 to 2000. This long-term volatility clustering coincides approximately with the period from the formal announcement of the merger of KDD, DDI, and IDO in December 1999 to the actual merger in October 2000. It is interesting that individual stocks behave differently from the stock price indices as events specific to each stock affect their movements. Comparing RV with the predictor calculated from each model, we can see that the ES0, ES1, HAR, HARQ, and ESQ models underestimate the volatility as a whole similar to the Nikkei 225 plot. Considering that the scale of the y axis of RV in the plot of Nikkei 225 is up to 0.2, we can see that the volatility of individual stocks may very high compared with the stock price indices.

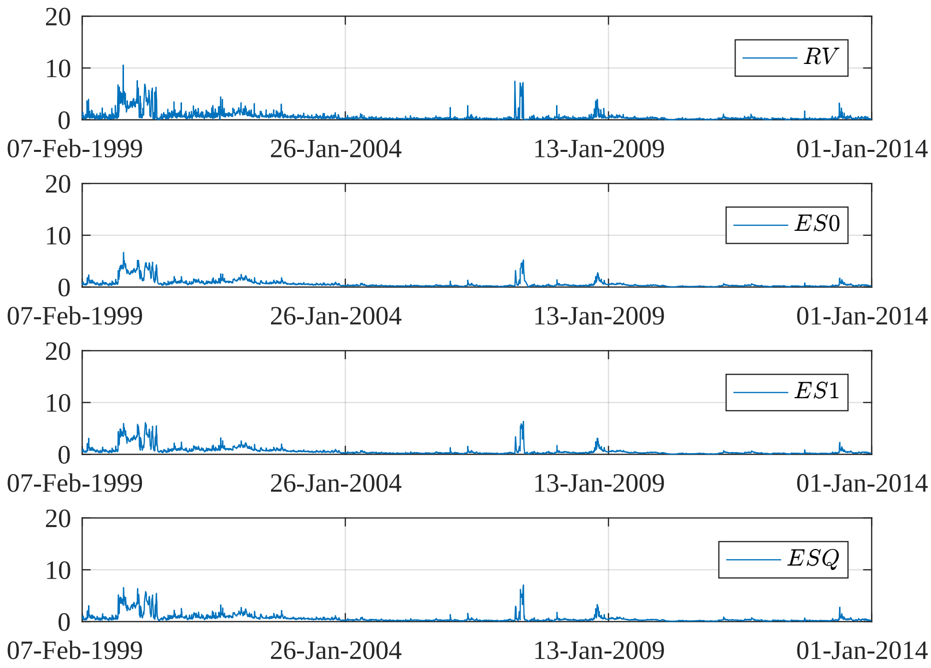

Figure 4: RV and predictors obtained from models (KDDI)

3.4.1 Descriptive Statistics of Error Functions

Which predictor has the highest predictive power among the predictor calculated for each model? To answer the question, we use the class of error functions proposed by Patton (2011) and compare the predictive power among the models. The error functions are robust in the presence of noise in proxy variables of RV, and we can use it for ranking predictive models. By parameterizing with a certain $b \in \mathbb{R}$, the class of the error functions is defined as

$$\mathcal{L}(rv, \hat{v}, b) = \begin{cases} \frac{1}{(b+1)(b+2)}(rv^{b+2} - \hat{v}^{b+2}) - \frac{1}{b+1}\hat{v}^{b+1}(rv - \hat{v}) & \text{for } b \notin \{-1, -2\}, \\ \hat{v} - rv + rv \cdot \log(rv/\hat{v}) & \text{for } b = -1, \\ \frac{rv}{\hat{v}} - \log \frac{rv}{\hat{v}} - 1 & \text{for } b = -2, \end{cases}$$

where rv is a volatility measure and $\hat{v}$ is the corresponding predictor. The error functions correspond to the quasi-likelihood (QLIKE) when $b = -2$, while corresponding to the MSE measure when $b = 0$. According to Patton and Sheppard (2009), QLIKE, which is a likelihood-based error function, is robust to noise, so that QLIKE is a preferable error function for comparing the predictive power of volatility compared to MSE. For a large positive value b, the error functions result from overestimation, whereas for a negative value b, the error functions increase and underestimate true values (Patton, 2011).

In the study, we use four kinds of values: $b \in \{1, 0, -1, -2\}$. The number of empirically calculated error functions is 27,000 ($225 \times 30 \times 4$), which is a combination of 30 stocks, 4 kinds of b, and 225 kinds of estimation and prediction periods in Table 4. Table 5 shows a part of the average errors calculated using the error function with $b = -2$, that is, QLIKE. Looking at the table, we can see that the results change in different estimation and prediction periods. It is also difficult to evaluate models by comparing the error functions one by one based on the results extending 27,000.

Therefore, we broadly classify the estimation and prediction periods into in-sample and out-of-sample, and we examine descriptive statistics to grasp the characteristics of the overall error functions. Tables 6 and 7 show the descriptive statistics of each error function for in-sample and out-of-sample, respectively. Focusing on the average values, from Table 6, ES1 is the lowest when $b = 1$, and HARQ is the lowest when $b = 0, -1, \text{or } -2$; in contrast, ARQ is the highest when $b = 0$ and AR1 is the highest when $b = 1, -1, \text{or } -2$. From the table, we can see that the average values of error functions for in-sample simulations are ordered as $\text{HARQ} < \text{ES1} < \text{HAR} \approx \text{ESQ} < \text{ES0} < \text{ARQ} < \text{AR1}$ as the overall tendency. From Table 6, ES1 is the lowest when $b = 0$, and ES0 is the lowest when $b = 1, -1, \text{and } -2$; in contrast, ARQ is the highest when $b = 1$, and AR1 is the highest when $b = 0, -1, \text{or } -2$. From the table, we can see that the average values of error functions for out-of-sample simulations are ordered as $\text{ES0} < \text{ES1} < \text{HAR} \approx \text{HARQ} \approx \text{ESQ} < \text{ARQ} < \text{AR1}$ as the overall tendency. Next, focusing on the maximum values, the results differ depending on b for both Tables 6 and 7. One thing we can say is that the differences between the results of all models, including AR1, are not so large.

Finally, focusing on the standard deviations in Table 6, ARQ or AR1 is the highest for all b, but there are not many differences among the results of the other ES0, ES1, HAR, HARQ, and ESQ models. Meanwhile, from Table 6, for out-of-sample simulations, we can recognize that ES0 has the lowest standard deviation in all b. As described above, ES0 is a model in which parameters other than constant terms of the HAR model are fixed to 1/3 without estimating parameters. Surprisingly, in out-of-sample simulations, ES0 is the least dispersive and can more stably predict volatility than other models. It is consistent with the results suggested in the previous study that ES0 exerts its power in out-of-sample prediction. The fact that ES0, that is, the 1/3 weighting model shows good results in out-of-sample simulations can be associated with the empirical finding that no models overcome the $1/N$ weights method when deciding the optimal portfolio selection (DeMiguel Et al., 2009). In other words, the result implies that uninformed decision makers tend to predict volatility by equally weighting volatilities observed daily, weekly, and monthly.

3.4.2 Model Comparison Based on the MCS

We cannot perform model comparisons simply by looking at the descriptive statistics of error functions for in-sample and out-of-sample simulations seen above. To judge the results of error functions appropriately in the framework of the statistical hypothesis tests and to compare the predictive power of models, we introduce the MCS proposed by Hansen et al. (2011). By using MCS, it is possible to select the best among the models at a given significance level without premising on a specific statistical model. We explain the outline of MCS following Hamid and Heiden (2015). First, we prepare a set $\mathcal{M}_0 = \{1, \ldots, m_0\}$ of candidate models, where $m_0 = 17$ in the study. Second, for all model pairs, we evaluate the superiority of the models based on the differences of the error functions L, obtained from each model. That is, for models i and j $(i, j = 1, \ldots, m_0)$ and for all time $t = 1, \ldots, T$, we evaluate

$$d_{ij,t} = L(rv_{it}, \widehat{rv}_{it}) - L(rv_{jt}, \widehat{rv}_{jt}).$$

Finally, against the model set $\mathcal{M} \in \mathcal{M}_0$, we test the null hypothesis

$$H_0 : E[d_{ij,t}] = 0, \quad \forall i, j \in \mathcal{M}, \quad i > j,$$

for each $d_{ij,t}$, where the initial value is set to $\mathcal{M} = \mathcal{M}_0$. If the null hypothesis H_0 is rejected at a given significance level (e.g., 10%), then, the model with the lowest predictive power is excluded from the model set. We continue the above method until H_0 cannot be rejected. Following Hansen et al. (2011), for evaluating H_0, we use range statistics such as

$$T_{R,k} = \max_{i,j \in \mathcal{M}} |t_{ij}| = \max_{i,j \in \mathcal{M}} \frac{|\bar{d}_{ij}|}{\sqrt{\widehat{var}(d_{ij})}},$$

where $\bar{d}_{ij} = \frac{1}{T}\sum_{t=1}^{T} d_{ij}$ and $\widehat{var}(\bar{d}_{ij})$ are obtained by using the block bootstrap method. The model i^* with the worst predictive power excluded from the model set $\mathcal{M}$ is chosen by a criterion,

$$i^* = \arg\max_{i \in \mathcal{M}} \frac{\bar{d}_i}{\sqrt{\widehat{var}(\bar{d}_i)}},$$

where $\bar{d}_i = \frac{1}{m-1}\sum_{j \in \mathcal{M}} \bar{d}_{ij}$ and m is the number of models included in model set $\mathcal{M}$. In the study, we perform a block boot strap method of block length 17 with 10,000 iterations. We set the significance level to 90% and use the MFE Toolbox mentioned above in our empirical analysis.

Next, we investigate the results of our empirical analysis using MCS. Like with error functions, we can execute MCS as many as $27,000$ ways, which is the number of combinations of 30 stocks, four kinds of b, and estimation and prediction periods of 225. Then, we excerpt a part of the results of MCS obtained from actual data and present it in Table 8. To ensure fairness, we present the results of the periods during which ES 0, ES 1, HAR, HARQ, and ESQ are the best. Table 8 lists the ranking of predictive power of models by MCS from 6th place to 1st place in the rightmost column. The parentheses under the model names represent the P values. In other words, the model with the P $= 1.00$ is the best, and after that, we rank predictive power of models depending on the P values.

Furthermore, Figures 5 to 7 show the cumulative frequencies of the models with the best predictive power based on the MCS criterion; that is, the models with P $= 1.00$ in the predictive power ranking obtained by MCS. Figure 5 shows the cumulative frequency of the best MCS models for in-sample simulations with estimation periods of 1 year, less than 5 years and more than 5 years, where the left-hand panels are stock indices and the right-hand panels are individual stocks. At first glance, in the left-side panels for stock indices, ESQ has the highest frequency being the best model during the estimation period of 1 year, and HARQ has the highest in the other periods. Either way we can see that HARQ and ESQ are overwhelmingly more predictive than the other models. On the right-side panels for individual stocks, we can see that ES1 is relatively good during the estimation period of more than 5 years, although the overall trend is similar to the stock indices. Figure 6 shows the cumulative frequency of the best MCS models for out-of-sample simulations with the same settings as in Figure 5. The characteristic feature of the plot is that the frequency of ES0 is the highest in all estimation periods of individual stocks in the out-of-sample simulations, as suggested by the result of the error functions in Table 7. In the left-side panels for stock indices, HAR has the highest frequency of the best model in the estimation period of more than 5 years, and ESQ has the highest in the other periods. Figure 7 summarizes Figures 5 and 6 and contains in-sample simulations, out-of-sample simulations, stock indices, and individual stocks for all estimation periods. The left- and right-side panels show cumulative frequencies for stock indices and individual stocks, respectively. The top and bottom panels are for in-sample and out-of-sample simulations, respectively. From the left-side panels for stock indices of the figure, we can see that HARQ has the highest frequency of being the best model for in-sample simulations, and ESQ has the highest for out-of-sample. From the right-side panels for individual stocks, we can see that HARQ again has the highest frequency of the best model for in-sample simulations, and ES0 has the highest for out-of-sample simulations. As described above, the eight types of GARCH models are overwhelmingly worse compared with other AR1, HAR, ES0, ES1, ES1a, ES1b, ARQ,

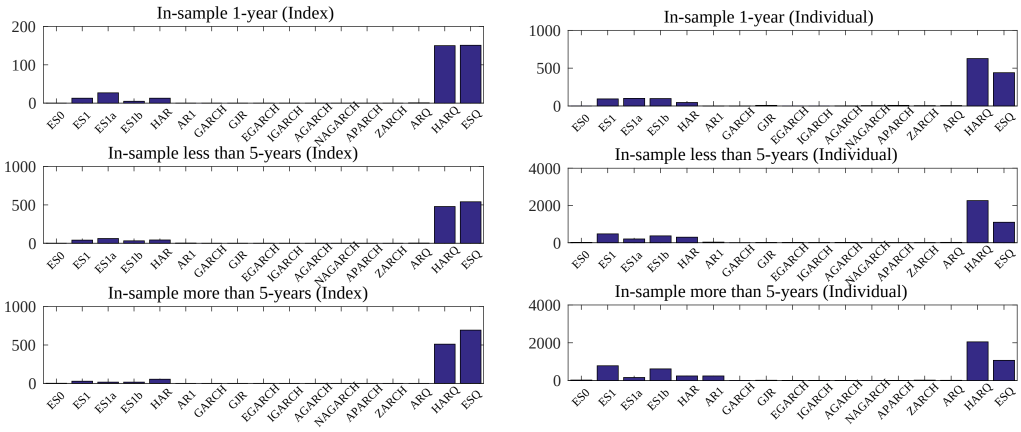

Figure 5: Cumulative frequencies of the best MCS models (in-sample)

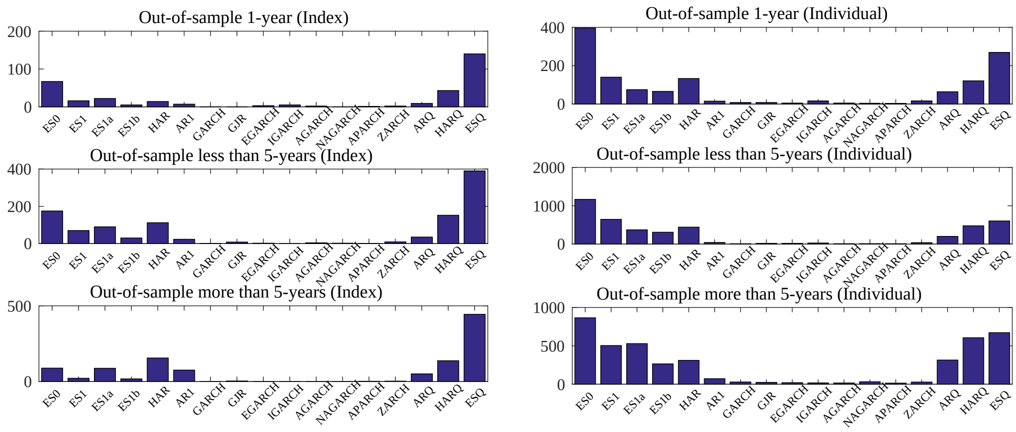

Figure 6: Cumulative frequencies of the best MCS models (out-of-sample)

HARQ, and ESQ models in a predictive power comparison since the GARCH models estimate and predict volatility by not directly using the RV_t series. Hence, note that it is not appropriate to directly compare the predictive power of the GARCH models with other AR1, HAR, ES0, ES1, ES1a, ES1b, ARQ, HARQ, and ESQ models.

3.4.3 Model Comparison Based on the MZ Regression

Finally, we report the results of the MZ regression on predictors obtained from models. According to Patton and Sheppard (2009), MZ regression is one of the general methods for evaluating the predictive power of volatility, which is proposed by Mincer and Zarnowitz (1969). The MZ regression is formulated by

$$\hat{\sigma}_t^2 = \alpha + \beta h_t + e_t,$$

where $\hat{\sigma}_t^2$ is a proxy variable for volatility, α and β are parameters, h_t is a predictor obtained from each model, and e_t is an error term. Regarding $\hat{\sigma}_t^2$, we consistently use RV as a proxy variable for volatility in the study. We performed the MZ regression on the data set of stock indices and individual stocks for in-sample and out-of-sample simulations. Table 9 shows

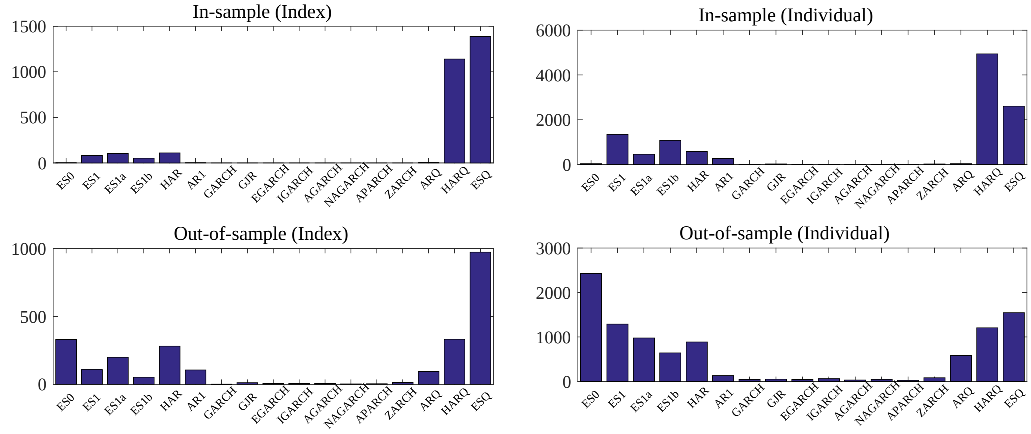

Figure 7: Cumulative frequencies of the best MCS models (all estimation periods)

the result of the adjusted coefficient of determination, R^2. Underlined values in the table are the maximum R^2 in each row. From the table of stock indices, we can see that ESQ has the maximum R^2 in all in-sample estimation periods, whereas ES0 has the maximum R^2 for 1 year and less than 5 years out-of-sample estimation period data. From the table of individual stocks, we can see that ES1 has the maximum R^2 for less than 5 and more than 5 years for in-sample estimation periods, whereas ES0 has the maximum R^2 in all out-of-sample estimation periods. In summary, ES0 has the highest R^2 in almost all out-of-sample estimation periods except for the more than 5 years estimation period of stock indices. This is consistent with the previous results of error functions and MCS.

Based on the result of the MZ regression, the overall trend can be summarized as follows: for in-sample simulations, the model using the information on the fourth moment of stock returns has better predictive power but such an effect is not observed in the out-of-sample simulations. As we often observe volatility clustering in financial asset returns, there are times when it is turbulent and times when it is calm. The various models analyzed in the study have no way to predict when large variability will come in the future, although RQ can be an important explanatory element to account for such differences in the in-sample period.

In situations where we perform an extrapolation type point prediction, there is no big difference when fourth moments are incorporated into the models. Thus, we may enjoy the benefit of assuming a fat-tailed distribution only when using low-frequency data. If the predicted period is short, it happens that such an assumption is successful, and the level of predictive power may increase. However, the longer the period becomes, the lower is the predictive power. This may serve as an interpretation that ES0 dominates the average in out-of-sample simulations.

4. Conclusions

In this paper, we focused on the framework of empirical similarity-based on case-based decision-making theory (Gilboa and Schmeidler, 1995, 2001) advocated by Gilboa et al. (2006). We conducted an empirical analysis of volatility prediction, using an empirical similarity-based time-series model proposed by Golosnoy et al. (2014). Regarding the predictive power comparison of models, we first ranked the predictive power of models in multiple estimation and prediction periods by using MCS with four error functions. We then analyzed cumulative frequencies of the best predictive power models. Because of the empirical analysis, HARQ for both stock indices and individual stocks in the in-sample had the highest frequencies as best models. ESQ for stock indices and ES0 for individual stocks had the highest frequencies in the out-of-sample simulations. Next, we performed the MZ regression to compare the predictive power of models in estimation and prediction periods of various combinations over multiple stocks. Looking at the result of the analysis based on the adjusted coefficient of determination, R^2, obtained from the MZ regression, ESQ for stock indices and ES1 for individual stocks were the best models in the in-sample. ES0 was the best model for both stock indices and individual stocks in the out-of-sample simulations. In the results of the MCS and MZ regression, the tendency that the predictive power of ES0 in out-of-sample data was relatively high compared to other models is consistent with the results of the previous work of Golosnoy et al. (2014). In addition, incorporating the realized measure RQ, calculated from high-frequency data such as HARQ and ESQ, was generally favorable in analyses of error functions, including the MCS and MZ regression. We once again confirmed the abundance of information derived from the aggregation of high-frequency financial data.

In this paper, we do not explicitly consider the asymmetry of volatility, although asymmetry is a very important factor in predicting volatility, so that we leave this for our future work.

Finally, as far as we know, we have applied ES models for volatility analyses in the Japanese stock market. We have done this systematically, using volatility models incorporating the realized measures of RQ, such as HARQ and ESQ. Therefore, the study of volatility models incorporating empirical similarity and the realized measure, RQ, has just begun. We would like to further explore and develop this area of study.

Acknowledgements

This study is partly supported by the Institute of Statistical Mathematics (ISM) cooperative research program (2015-ISM-CRP-2012) and JSPS KAKENHI Grant Number 25245033, 15H01943, 15H03337, 16K00067, 26380279, and 15K03406. We also would like to thank Dr. Moritz Heiden for his valuable comments.

REFERENCES

Andersen, T. G., and Bollerslev, T. (1997). Intraday Periodicity and Volatility Persistence in Financial Markets. *Journal of Empirical Finance* **4**, 115–158.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, **61**, 43–76.

Andersen, T. G., Bollerslev, T., Diebold, F.X. and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, **71**, 579–625.

Andersen, T. G., Bollerslev, T. and Meddahi, N. (2004). Analytical evaluation of volatility forecasts. *International Economic Review*, **45**, 1079–1110.

Andersen, T. G., Dobrev, D. and Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, **169**, 75–93.

Bandi, F. M. and Russell, J. R. (2004). Microstructure Noise, Realized Variance, and Optimal Sampling. *Review of Economic Studies*, **75**, 339–369.

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**, 253–280.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008). Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise. *Econometrica*, **76**, 1481–1536.

Bates, J. M., Granger, C. W. J. (1969). The combination of forecasts. *Operation Research Quarterly*, **20**, 451–468.

Billot, A., Gilboa, I. and Schmeidler, D. (2008). An axiomatization of an exponential similarity function. *Mathematical Social Sciences*, **55**, 107–115.

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

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

DeMiguel, V., Garlappi, L. and Uppal, R. (2009). Optimal versus naive diversification: how inefficient is the $1/N$ portfolio strategy? *Review of Financial Studies*, **22**, 1915–1953.

Elliott, G., Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics*, **122**, 47–79.

Gilboa, I. and Schmeidler, D. (1995). Case-Based Decision Theory. *Quarterly Journal of Economics*. **110**, 605–639.

Gilboa, I. and Schmeidler, D. (2001). *A Theory of Cased-based Decisions*, Cambridge University Press, Cambridge, UK.

Gilboa, I.,and Schmeidler, D. (2012). *Case-Based Predictions*, World Scientific, Singapore.

Gilboa, I., Lieberman, O. and Schmeidler, D. (2006). Empirical similarity. *Review of Economics and Statistics*, **88**, 433–444.

Gilboa, I., Lieberman, O. and Schmeidler, D. (2011). A similarity-based approach to prediction. *Journal of Econometrics*, **162**, 124–131.

Golosnoy, V., Hamid, A. and Okhrin, Y. (2014). The empirical similarity approach for volatility prediction. *Journal of Banking & Finance*, **40**, 321–329.

Hamid, A. and Heiden, M. (2015). Forecasting volatility with empirical similarity and Google Trends. *Journal of Economic Behavior & Organization*, **117**, 62–81.

Hansen, P. R. and Lunde, A. (2005). A Realized Variance for the Whole Day Based on Intermittent High-Frequency Data. *Journal of Financial Econometrics*, **3**, 525–554.

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

Hume, D. (1748). *Enquiry into the Human Understanding*, Clarendon Press, Oxford, UK.

Lieberman, O. (2010). Asymptotic theory for empirical similarity models. *Econometric Theory* **26**, 1032–1059.

Masuda, H. and Morimoto, T. (2012). Optimal Weight For Realized Variance Based On Intermittent High-Frequency Data. *Japanese Economic Review*, **63**, 497–527.

McAleer, M. and Medeiros, M. C. (2008). Realized volatility: A review. *Econometric Reviews*, **27**, 10–45.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In: J. A. Mincer (ed.), *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, Studies in Business Cycles, NBER, 3–46.

Patton, A. J. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In: T. G. Andersen, R. A. Davis, J. P. Kreiss and T. Mikosch (eds.), *Handbook of Financial Time Series*, Springer, Berlin, 801–838.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, **160**, 246–256.

Weisberg, S. (2014). *Applied Linear Regression (4th edition)*, Wiley, Hoboken, NJ.

Wooldridge, J. M. (2015). *Introductory Econometrics: A Modern Approach (6th Edition)*, Cengage Learning, Boston, MA.

Zhang, L., Mykland, P. A. and Ait-Sahalia, Y. (2005). A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data. *Journal of the American Statistical Association*, **100**, 1394–1411.

Table 4: All estimation and prediction periods for empirical analysis

Esti.	Pred.	Esti.	Pred.	Esti.	Pred.	Esti.	Pred.	Esti.	Pred.
99-99	99-99	00-08	00-08	02-06	07-07	04-09	04-09	07-09	07-09
99-99	00-00	00-08	09-09	02-07	02-07	04-09	10-10	07-09	10-10
99-00	99-00	00-09	00-09	02-07	08-08	04-10	04-10	07-10	07-10
99-00	01-01	00-09	10-10	02-08	02-08	04-10	11-11	07-10	11-11
99-01	99-01	00-10	00-10	02-08	09-09	04-11	04-11	07-11	07-11
99-01	02-02	00-10	11-11	02-09	02-09	04-11	12-12	07-11	12-12
99-02	99-02	00-11	00-11	02-09	10-10	04-12	04-12	07-12	07-12
99-02	03-03	00-11	12-12	02-10	02-10	04-12	13-13	07-12	13-13
99-03	99-03	00-12	00-12	02-10	11-11	04-13	04-13	07-13	07-13
99-03	04-04	00-12	13-13	02-11	02-11	05-05	05-05	08-08	08-08
99-04	99-04	00-13	00-13	02-11	12-12	05-05	06-06	08-08	09-09
99-04	05-05	01-01	01-01	02-12	02-12	05-06	05-06	08-09	08-09
99-05	99-05	01-01	02-02	02-12	13-13	05-06	07-07	08-09	10-10
99-05	06-06	01-02	01-02	02-13	02-13	05-07	05-07	08-10	08-10
99-06	99-06	01-02	03-03	03-03	03-03	05-07	08-08	08-10	11-11
99-06	07-07	01-03	01-03	03-03	04-04	05-08	05-08	08-11	08-11
99-07	99-07	01-03	04-04	03-04	03-04	05-08	09-09	08-11	12-12
99-07	08-08	01-04	01-04	03-04	05-05	05-09	05-09	08-12	08-12
99-08	99-08	01-04	05-05	03-05	03-05	05-09	10-10	08-12	13-13
99-08	09-09	01-05	01-05	03-05	06-06	05-10	05-10	08-13	08-13
99-09	99-09	01-05	06-06	03-06	03-06	05-10	11-11	09-09	09-09
99-09	10-10	01-06	01-06	03-06	07-07	05-11	05-11	09-09	10-10
99-10	99-10	01-06	07-07	03-07	03-07	05-11	12-12	09-10	09-10
99-10	11-11	01-07	01-07	03-07	08-08	05-12	05-12	09-10	11-11
99-11	99-11	01-07	08-08	03-08	03-08	05-12	13-13	09-11	09-11
99-11	12-12	01-08	01-08	03-08	09-09	05-13	05-13	09-11	12-12
99-12	99-12	01-08	09-09	03-09	03-09	06-06	06-06	09-12	09-12
99-12	13-13	01-09	01-09	03-09	10-10	06-06	07-07	09-12	13-13
99-13	99-13	01-09	10-10	03-10	03-10	06-07	06-07	09-13	09-13
00-00	00-00	01-10	01-10	03-10	11-11	06-07	08-08	10-10	10-10
00-00	01-01	01-10	11-11	03-11	03-11	06-08	06-08	10-10	11-11
00-01	00-01	01-11	01-11	03-11	12-12	06-08	09-09	10-11	10-11
00-01	02-02	01-11	12-12	03-12	03-12	06-09	06-09	10-11	12-12
00-02	00-02	01-12	01-12	03-12	13-13	06-09	10-10	10-12	10-12
00-02	03-03	01-12	13-13	03-13	03-13	06-10	06-10	10-12	13-13
00-03	00-03	01-13	01-13	04-04	04-04	06-10	11-11	10-13	10-13
00-03	04-04	02-02	02-02	04-04	05-05	06-11	06-11	11-11	11-11
00-04	00-04	02-02	03-03	04-05	04-05	06-11	12-12	11-11	12-12
00-04	05-05	02-03	02-03	04-05	06-06	06-12	06-12	11-12	11-12
00-05	00-05	02-03	04-04	04-06	04-06	06-12	13-13	11-12	13-13
00-05	06-06	02-04	02-04	04-06	07-07	06-13	06-13	11-13	11-13
00-06	00-06	02-04	05-05	04-07	04-07	07-07	07-07	12-12	12-12
00-06	07-07	02-05	02-05	04-07	08-08	07-07	08-08	12-12	13-13
00-07	00-07	02-05	06-06	04-08	04-08	07-08	07-08	12-13	12-13
00-07	08-08	02-06	02-06	04-08	09-09	07-08	09-09	13-13	13-13

Esti. and Pred. denote estimation and prediction periods respectively.

Table 5: Error functions for $b = -2$ (excerpt)

Nikkei 225	Esti.	Pred.	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
	99-11	99-11	0.156	0.156	0.154	0.186	0.180	0.153	0.153
	99-11	12-12	0.186	0.185	0.179	0.201	0.203	0.183	0.173
	99-12	99-12	0.159	0.159	0.156	0.187	0.183	0.156	0.155
	99-12	13-13	0.205	0.207	0.203	0.220	0.230	0.205	0.202
	99-13	99-13	0.162	0.162	0.160	0.190	0.186	0.160	0.159
Astellas	Esti.	Pred.	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
	99-11	99-11	0.025	0.025	0.025	0.032	0.032	0.025	0.025
	99-11	12-12	0.009	0.009	0.011	0.025	0.024	0.011	0.010
	99-12	99-12	0.024	0.024	0.024	0.031	0.030	0.024	0.024
	99-12	13-13	0.028	0.028	0.028	0.031	0.030	0.028	0.028
	99-13	99-13	0.024	0.024	0.023	0.030	0.029	0.023	0.024
Komatsu	Esti.	Pred.	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
	99-11	99-11	0.176	0.175	0.173	0.223	0.211	0.171	0.171
	99-11	12-12	0.107	0.107	0.116	0.212	0.182	0.114	0.107
	99-12	99-12	0.174	0.173	0.170	0.219	0.208	0.168	0.169
	99-12	13-13	0.240	0.246	0.263	0.373	0.335	0.257	0.255
	99-13	99-13	0.180	0.179	0.177	0.227	0.214	0.175	0.176
Hitachi	Esti.	Pred.	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
	99-11	99-11	0.129	0.128	0.132	0.187	0.159	0.128	0.128
	99-11	12-12	0.030	0.030	0.030	0.030	0.032	0.029	0.031
	99-12	99-12	0.123	0.123	0.126	0.177	0.151	0.122	0.123
	99-12	13-13	0.212	0.207	0.212	0.230	0.211	0.205	0.212
	99-13	99-13	0.127	0.127	0.130	0.179	0.153	0.126	0.126
Toyota	Esti.	Pred.	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
	99-11	99-11	0.079	0.079	0.079	0.101	0.094	0.078	0.079
	99-11	12-12	0.079	0.079	0.077	0.100	0.090	0.077	0.080
	99-12	99-12	0.083	0.083	0.083	0.105	0.098	0.082	0.083
	99-12	13-13	0.102	0.102	0.105	0.114	0.088	0.109	0.097
	99-13	99-13	0.100	0.099	0.097	0.117	0.108	0.096	0.097

Esti. and Pred. denote estimation and prediction periods respectively.

Table 6: Descriptive statistics of error functions (in-sample)

$b = 1$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.0145	0.0142	0.0144	0.0162	0.0156	0.0147	0.0146
Median	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Max.	0.8704	0.9900	0.9717	0.9532	0.9082	0.9982	0.9041
SD	0.0743	0.0734	0.0747	0.0838	0.0801	0.0766	0.0758
Ske.	6.6617	6.9163	6.9393	6.7303	6.7412	7.0681	6.7948
Kur.	51.19	57.44	57.50	52.10	52.44	60.28	54.26
Size	3324	3324	3324	3324	3324	3324	3324
$b = 0$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.0083	0.0080	0.0080	0.0092	0.0093	0.0079	0.0080
Median	0.0004	0.0004	0.0004	0.0005	0.0005	0.0004	0.0004
Max.	0.9818	0.9703	0.9486	0.6279	0.9609	0.9093	0.9434
SD	0.0397	0.0384	0.0381	0.0411	0.0434	0.0377	0.0384
Ske.	10.2256	10.4631	10.2925	7.6756	9.1738	9.9895	10.1998
Kur.	163.47	172.57	166.33	77.96	125.25	154.31	161.72
Size	3324	3324	3324	3324	3324	3324	3324
$b = -1$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.0110	0.0107	0.0107	0.0129	0.0125	0.0106	0.0107
Median	0.0042	0.0041	0.0041	0.0050	0.0049	0.0041	0.0041
Max.	0.7353	0.7070	0.6926	0.7427	0.7270	0.6898	0.6713
SD	0.0272	0.0262	0.0261	0.0307	0.0303	0.0259	0.0261
Ske.	10.1021	10.0234	9.7591	8.4654	8.3573	9.7409	9.4669
Kur.	188.98	186.43	177.16	129.45	125.12	176.75	162.61
Size	3326	3326	3326	3326	3326	3326	3326
$b = -2$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.1161	0.1153	0.1148	0.1375	0.1317	0.1138	0.1139
Median	0.1090	0.1082	0.1077	0.1287	0.1229	0.1082	0.1074
Max.	0.6328	0.6067	0.9892	0.9444	0.9738	0.9847	0.9598
SD	0.0747	0.0750	0.0757	0.0857	0.0839	0.0750	0.0739
Ske.	1.0967	1.1221	1.7024	0.9060	1.2516	1.7417	1.4951
Kur.	5.8412	5.9414	13.2327	5.9489	9.2506	13.8119	11.3362
Size	3326	3326	3326	3326	3326	3326	3326

Table 7: Descriptive statistics of error functions (out-of-sample)

$b=1$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.0026	0.0028	0.0027	0.0035	0.0040	0.0029	0.0029
Median	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Max.	0.4941	0.4660	0.5605	0.6676	0.9178	0.4867	0.5220
SD	0.0175	0.0179	0.0185	0.0228	0.0276	0.0190	0.0192
Ske.	14.1394	12.9544	15.3115	14.7657	17.6783	12.7591	13.4480
Kur.	286.85	232.25	349.62	322.32	472.63	223.54	256.69
Size	2840	2840	2840	2840	2840	2840	2840
$b=0$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.0050	0.0049	0.0056	0.0063	0.0058	0.0052	0.0054
Median	0.0002	0.0002	0.0002	0.0003	0.0003	0.0002	0.0002
Max.	0.5086	0.7682	0.8326	0.8262	0.8362	0.6840	0.8029
SD	0.0344	0.0363	0.0422	0.0432	0.0388	0.0375	0.0399
Ske.	12.3288	15.1551	13.2819	13.7566	15.3827	14.0863	13.9322
Kur.	166.77	264.39	193.23	212.51	272.53	220.87	218.47
Size	2840	2840	2840	2840	2840	2840	2840
$b=-1$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.0091	0.0096	0.0098	0.0130	0.0129	0.0101	0.0098
Median	0.0035	0.0037	0.0039	0.0054	0.0051	0.0037	0.0037
Max.	0.3381	0.5879	0.4847	0.5723	0.9602	0.6940	0.5902
SD	0.0224	0.0286	0.0264	0.0318	0.0375	0.0302	0.0280
Ske.	10.0364	13.3357	10.7361	9.3700	12.9680	13.2420	11.8349
Kur.	131.31	225.72	147.35	118.62	239.53	238.43	181.30
Size	2841	2841	2841	2841	2841	2841	2841
$b=-2$	ES0	ES1	HAR	AR1	ARQ	HARQ	ESQ
Mean	0.1225	0.1243	0.1275	0.1697	0.1624	0.1276	0.1245
Median	0.1003	0.1016	0.1035	0.1351	0.1291	0.1016	0.1016
Max.	0.5506	0.7121	0.5686	0.9904	0.9495	0.8868	0.7665
SD	0.1004	0.1028	0.1042	0.1415	0.1367	0.1071	0.1026
Ske.	1.2581	1.3651	1.2641	1.4917	1.5307	1.5091	1.3682
Kur.	4.4922	5.2335	4.4802	5.6011	6.0364	6.4795	5.3569
Size	2841	2841	2841	2841	2841	2841	2841

Table 8: Model ranking by MCS (Nikkei 225, excerpt)

Estimation	Prediction	1st	2nd	3rd	4th	5th	6th
1999-2004	2005-2005	ES0	ES1	ES1b	ESQ	ES1a	HARQ
		(1.000)	(0.945)	(0.572)	(0.141)	(0.053)	(0.024)
2001-2004	2005-2005	ES0	ES1	ES1b	ESQ	ES1a	HARQ
		(1.000)	(0.066)	(0.047)	(0.047)	(0.047)	(0.047)
2002-2003	2004-2004	ES0	ES1b	ES1	ESQ	ES1a	HAR
		(1.000)	(0.226)	(0.226)	(0.203)	(0.160)	(0.057)
Estimation	Prediction	1st	2nd	3rd	4th	5th	6th
2000-2004	2005-2005	ES1	ES0	ESQ	ES1a	ES1b	HARQ
		(1.000)	(0.850)	(0.142)	(0.090)	(0.090)	(0.029)
2001-2002	2003-2003	ES1	ES1b	ES0	ESQ	ES1a	HAR
		(1.000)	(0.179)	(0.179)	(0.179)	(0.176)	(0.097)
2008-2008	2009-2009	ES1	ES0	ES1b	ESQ	HAR	ES1a
		(1.000)	(0.816)	(0.313)	(0.313)	(0.003)	(0.003)
Estimation	Prediction	1st	2nd	3rd	4th	5th	6th
1999-2009	2010-2010	HAR	ES1b	ES1	ESQ	ES0	HARQ
		(1.000)	(0.939)	(0.250)	(0.250)	(0.225)	(0.017)
2002-2009	2010-2010	HAR	ESQ	ES0	HARQ	ES1b	ES1
		(1.000)	(0.713)	(0.672)	(0.007)	(0.007)	(0.007)
2009-2011	2012-2012	HAR	ESQ	HARQ	ES1b	ES1	ES1a
		(1.000)	(0.498)	(0.438)	(0.438)	(0.135)	(0.135)
Estimation	Prediction	1st	2nd	3rd	4th	5th	6th
1999-2006	2007-2007	HARQ	ESQ	HAR	ES1	ES0	ES1a
		(1.000)	(0.924)	(0.257)	(0.104)	(0.104)	(0.104)
2001-2006	2007-2007	HARQ	ESQ	HAR	ES1a	ES1b	ES1
		(1.000)	(0.835)	(0.388)	(0.388)	(0.388)	(0.388)
2001-2012	2013-2013	HARQ	ESQ	ES1a	ARQ	HAR	ES1b
		(1.000)	(0.963)	(0.963)	(0.963)	(0.555)	(0.555)
Estimation	Prediction	1st	2nd	3rd	4th	5th	6th
1999-2007	1999-2007	ESQ	HARQ	HAR	ES1	ES0	ES1a
		(1.000)	(0.506)	(0.391)	(0.034)	(0.028)	(0.001)
2000-2007	2000-2007	ESQ	HARQ	HAR	ES1	ES1a	ES0
		(1.000)	(0.340)	(0.288)	(0.067)	(0.004)	(0.004)
2001-2004	2001-2004	ESQ	HARQ	HAR	ES1a	ES0	ES1b
		(1.000)	(0.692)	(0.692)	(0.044)	(0.044)	(0.044)

Table 9: Adjusted coefficients of determination R^2 obtained from MZ regression

Stock indices	ES0	ES1	ES1a	ES1b	HAR	AR1	ARQ	HARQ	ESQ
In-sample									
1 year	0.293	0.320	0.323	0.313	0.312	0.250	0.269	0.322	<u>0.330</u>
less than 5 years	0.436	0.451	0.450	0.449	0.444	0.368	0.385	0.450	<u>0.457</u>
more than 5 years	0.547	0.554	0.554	0.552	0.552	0.468	0.485	0.557	<u>0.560</u>
Out-of-sample									
1 year	<u>0.302</u>	0.254	0.270	0.259	0.291	0.258	0.191	0.247	0.263
less than 5 years	<u>0.317</u>	0.300	0.305	0.304	0.317	0.272	0.241	0.290	0.300
more than 5 years	0.324	0.308	<u>0.326</u>	0.312	0.325	0.294	0.277	0.316	0.325
Individual stocks	**ES0**	**ES1**	**ES1a**	**ES1b**	**HAR**	**AR1**	**ARQ**	**HARQ**	**ESQ**
In-sample									
1 year	0.284	0.308	0.305	0.304	0.299	0.240	0.259	<u>0.309</u>	0.308
less than 5 years	0.432	<u>0.447</u>	0.438	0.444	0.440	0.364	0.385	0.446	0.443
more than 5 years	0.538	<u>0.548</u>	0.540	0.546	0.543	0.456	0.480	0.547	0.545
Out-of-sample									
1 year	<u>0.289</u>	0.265	0.261	0.268	0.275	0.245	0.208	0.244	0.261
less than 5 years	<u>0.295</u>	0.280	0.282	0.284	0.289	0.252	0.231	0.273	0.280
more than 5 years	<u>0.286</u>	0.275	0.280	0.278	0.281	0.249	0.236	0.273	0.277