COLLANA DEL
DIPARTIMENTO DI ECONOMIA

FORECAST COMBINATIONS FOR REALIZED VOLATILITY IN PRESENCE OF STRUCTURAL BREAKS

Davide De Gaetano

ISSN 2279-6916 Working papers
(Dipartimento di Economia Università degli studi Roma Tre) (online)

Working Paper n° 208, 2016

I Working Papers del Dipartimento di Economia svolgono la funzione di divulgare tempestivamente, in forma definitiva o provvisoria, i risultati di ricerche scientifiche originali. La loro pubblicazione è soggetta all'approvazione del Comitato Scientifico.

Per ciascuna pubblicazione vengono soddisfatti gli obblighi previsti dall'art. l del D.L.L. 31.8.1945, n. 660 e successive modifiche.

Copie della presente pubblicazione possono essere richieste alla Redazione.

REDAZIONE:
Dipartimento di Economia
Università degli Studi Roma Tre
Via Silvio D'Amico, 77 - 00145 Roma
Tel. 0039-06-57335655   fax 0039-06-57335771
E-mail: dip_eco@uniroma3.it
http://dipeco.uniroma3.it

DIPARTIMENTO DI ECONOMIA

FORECAST COMBINATIONS FOR REALIZED VOLATILITY IN PRESENCE OF STRUCTURAL BREAKS

Davide De Gaetano

Forecast Combinations for Realized Volatility in Presence of Structural Breaks

Davide De Gaetano

Department of Economics
University of Roma Tre (Italy)

`davide.degaetano@uniroma3.it`

Abstract

In this paper the problem of instability due to changes in the parameters of some Realized Volatility (RV) models has been addressed. The analysis is based on 5-minute RV of four U.S. stock market indices. Three different representations of the log-RV have been considered and, for each of them, the parameter instability has been detected by using the recursive estimates test. In order to analyse how instabilities in the parameters affect the forecasting performance, an out-of-sample forecasting exercise has been performed. In particular, several forecast combinations, designed to accommodate potential structural breaks, have been considered. All of them are based on different estimation windows, with alternative weighting schemes, and do not take into account explicitly estimated break dates. The model confidence set has been used to compare the forecasting performances of the proposed approaches. Our analysis gives empirical evidences of the effectiveness of the combinations which make adjustments for accounting the possible most recent break point.

Keywords: Forecast combinations, Structural breaks, Realized volatility.

JEL Code: C530, C580, G170.

Acknowledgements: The author is grateful to prof. M.Barbieri for her useful suggestions and comments. The usual disclaimer applies.

1 Introduction

Modelling and forecasting the volatility is an important issue in empirical finance. Traditional approaches are based on the univariate GARCH class of models or stochastic volatility models. More recently, Realized Volatility (RV) has become very popular; it uses improved measures of ex post volatility constructed from high frequency data and provides an efficient estimate of the unobserved volatility of financial markets. In contrast with the GARCH approach, in which the volatility is treated as a latent variable, RV can be considered as an observable proxy and, as a consequence, it can be used in time series model to generate forecasts.
Many authors (see for example Liu and Maheau, 2008) have highlighted the evidence for structural breaks in RV. Their presence in the data generating process can induce instability in the model parameters. Ignoring structural breaks and assuming wrongly that the structure of a model remains fixed over time, has clear adverse implications. The first finding is inconsistency of the parameter estimates. Moreover, structural changes are likely to be responsible for most major forecast failures of time invariant series models.
In this paper, three different model specifications of the log-RV have been considered. The first is the Heterogeneous Auto-Regressive model (HAR-RV) proposed by Corsi (2004); this model is able to capture many of the features of volatility including long memory, fat tails and self-similarity. The second is the Leverage Heterogeneous Auto-Regressive model (LHAR-RV) proposed by Corsi and Renó (2010); it is able to approximate both long-range dependence and the leverage effect. The last is the Asymmetric Heterogeneous Auto-Regressive model (AHAR-RV) which is a simplified version of the model proposed by Liu and Maheu (2008). In the spirit of the EGARCH model, it allows for asymmetric effects from positive and negative returns. These models, which have become very popular in the econometric literature on RV, have very parsimonious linear structures and, as a consequence, they are extremely easy to implement and to estimate. Moreover, they have good performances in approximating many features which characterize the dynamic of RV (Corsi, 2009).
The aim of this paper is to investigate empirically the relevance of structural breaks for forecasting RV of a financial time series. The presence of structural breaks in the considered RV-representations has been investigated and verified by resorting to a fluctuation test for parameter instability in a regression context. In particular the attention has been focused on the recursive estimates test proposed by Ploberger *et al.* (1989). This choice is motivated especially in cases where no particular pattern of the deviation from the null hypothesis of constant parameters is assumed. Furthermore the proposal does not require the specification of the locations of the break points.
In order to handle parameter instability, some specific forecast combinations have been intro-

duced and discussed. They are based on different estimation windows with alternative weighting schemes. All of them do not incorporate explicitly the estimation of the break dates and, as shown by Pesaran and Timmermann (2007), they do not suffer of this estimation uncertainty. The forecasting performances of the proposed forecast combinations for the three different specifications of RV models have been compared in terms of two loss functions, the MSE and QLIKE. They are the loss functions most widely used to compare volatility forecasting performances and, according to Patton (2011), they provide robust ranking of the models.
In order to assess statistically if the differences in the forecasting performances of the considered forecast combinations are relevant, the model confidence set, proposed by Hansen *et al.* (2011) has been used. This procedure is able to construct a set of models from a specified collection which consists of the best models in term of a loss function with a given level of confidence and it does not require the specification of a benchmark. Moreover, the model confidence set procedure is a stepwise method based on a sequence of significance tests; each test is based on a null hypothesis that the two models under comparison have the same forecasting ability against the alternative that they are not equivalent. The test stops when the first hypothesis is not rejected and, therefore, the procedure does not accumulate type I error (Hansen *et al.*, 2005). The critical values of the test as well as the estimation of the variance useful to construct the test statistic, is determined by using the block bootstrap. This re-sampling technique preserves the dependence structure of the series and it works reasonably well under very weak conditions on the dependency structure of the data.
The empirical analysis has been conducted on four U.S. stock market indices: S&P 500, Dow Jones Industrial Average, Russell 2000 and Nasdaq 100. For all the series the 5-minute RV has been considered; it is one of the most used proxy of the volatility and, as shown in Liu *et al.* (2015), no other realized measure outperforms it in terms of estimation accuracy of asset price variation.

The structure of this paper is as follows. Section 2 introduces the empirical models for RV and briefly illustrates the problem of structural breaks. In Section 3 some of the most used procedures to test parameters instability in regression framework have been reviewed. The attention has been focused on the class of fluctuation tests and, in particular, the Recursive estimates test has been discussed. Section 4 introduces the problem of forecasting in presence of structural breaks and discusses some forecast combinations able to take into account parameters instability. In section 5 the problem of comparing forecasting models is briefly illustrated and the model confidence set is discussed. In section 6 the empirical results on the four U.S. stock market indices are reported and discussed. Some final remarks close the paper.

2 Realized volatility models

In the econometric literature many approaches have been developed to model and forecast realized volatility with the aim of reproducing the main empirical features of financial time series such as long memory, fat tails and self-similarity. In this paper the attention has been focused on the classic Heterogeneous Auto-Regressive model of Realized Volatility (HAR-RV) and on some of its extensions.

The HAR-RV model, proposed by Corsi (2004), has a very simple and parsimonious structure; moreover, empirical analysis (Corsi, 2009) shows remarkably good forecasting performance. In this model, lags of RV are used at daily, weekly and monthly aggregated periods.

More precisely, let $v_t = log(RV_t)$ where RV_t is the realized volatility at time $t = 1, 2, \dots, T$. The logarithmic version of the HAR-RV similar to that implemented by Andersen *et al.* (2007), is defined as:

$$v_t = \beta_0 + \beta_1 v_{t-1} + \beta_2 v_t^{(5)} + \beta_3 v_t^{(22)} + \epsilon_t \tag{1}$$

where $\epsilon_t \sim NID(0, \sigma^2)$ and $v_t^{(5)}$ and $v_t^{(22)}$ are defined, respectively, as:

$$v_t^{(5)} = \frac{v_{t-1} + v_{t-2} + \ldots + v_{t-5}}{5} \tag{2}$$

$$v_t^{(22)} = \frac{v_{t-1} + v_{t-2} + \ldots + v_{t-22}}{22} \tag{3}$$

The HAR-RV model is able to capture some well known features of financial returns such as long memory and fat tails (see Corsi, 2009).

The first extension of this model is the Leverage Heterogeneous Auto-Regressive model of Realized Volatility (LHAR-RV) proposed by Corsi and Renó (2010). This model is defined as:

$$\begin{aligned} v_t = \beta_0 + \beta_1 v_{t-1} + \beta_2 v_t^{(5)} + \beta_3 v_t^{(22)} + \beta_4 r_{t-1}^{-} + \beta_5 r_t^{(5)-} + \\ + \beta_6 r_t^{(22)-} + \beta_7 r_{t-1}^{+} + \beta_8 r_t^{(5)+} + \beta_9 r_t^{(22)+} + \epsilon_t \end{aligned} \tag{4}$$

where $\epsilon_t \sim NID(0, \sigma^2)$, r_t are the daily returns and:

$$r_t^{(5)-} = \frac{r_{t-1} + r_{t-2} + \ldots + r_{t-5}}{5} I_{\{(r_{t-1}+r_{t-2}+\ldots+r_{t-5})<0\}} \tag{5}$$

$$r_t^{(5)+} = \frac{r_{t-1} + r_{t-2} + \ldots + r_{t-5}}{5} I_{\{(r_{t-1}+r_{t-2}+\ldots+r_{t-5})>0\}} \tag{6}$$

$$r_t^{(22)-} = \frac{r_{t-1} + r_{t-2} + \ldots + r_{t-22}}{22} I_{\{(r_{t-1}+r_{t-2}+\ldots+r_{t-22})<0\}} \tag{7}$$

$$r_t^{(22)+} = \frac{r_{t-1} + r_{t-2} + \ldots + r_{t-22}}{22} I_{\{(r_{t-1}+r_{t-2}+\ldots+r_{t-22})>0\}} \tag{8}$$

where I is the indicator function. The LHAR-RV model approximates both long-range dependence and the leverage effect. Some authors (Asai *et al.*, 2012) suggest to include only the nagative part of heterogeneous returns since the estimates of the coefficients of the positive ones are usually not significant.

The second extension is the Asymmetric Heterogeneous Auto-Regressive model of Realized Volatility (AHAR-RV) which is a simplified version of the model proposed by Liu and Maheu (2008). It is defined as:

$$v_t = \beta_0 + \beta_1 v_{t-1} + \beta_2 v_t^{(5)} + \beta_3 v_t^{(22)} + \beta_4 \frac{|r_{t-1}|}{\sqrt{RV_{t-1}}} + \beta_5 \frac{|r_{t-1}|}{\sqrt{RV_{t-1}}} I_{\{r_{t-1}<0\}} + \epsilon_t \tag{9}$$

The last two terms allow for asymmetric effects from positive and negative returns in the spirit of the EGARCH model.

All the considered models can be rewritten in a standard regression framework:

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \epsilon_t \tag{10}$$

where $y_t = v_t$, $\mathbf{x}_t$ is the $p \times 1$ vector of the regressors at time t and $\boldsymbol{\beta}$ is the $p \times 1$ vector of the corresponding coefficients. Of course the number p and the specification of the vector $\mathbf{x}_t$ are different for each model.

Many studies (see, for example, Liu and Maheau, 2008) agree on the existence of structural breaks in RV. If structural breaks are present in the data generating process they could induce instability in the model parameters. Ignoring them in the specification of the model, could provide a wrong modelling and forecasting for the RV.

To dealing with structural breaks, the linear regression model (10) is assumed to have time varying coefficients and so it may be expressed as:

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_t + \epsilon_t \qquad t = 1, 2, \ldots T \tag{11}$$

In many applications it is reasonable to assume that there are m breakpoints at the date $\tau_1, \tau_2, \cdots, \tau_m$ in which the coefficients shift from one stable regression relationship to a different one. Thus, there are m + 1 segments in which the regression coefficients are constant. Model (11) can be rewritten as

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_{\tau_{j-1}+1:\tau_j} + \epsilon_t \qquad t = 1, 2, \ldots T \qquad j = 1, 2, \ldots m + 1 \tag{12}$$

and, by convention, $\tau_0 = 1$ and $\tau_{m+1} = T$.

3 Testing for structural changes

The presence of structural breaks can be tested through the null hypothesis that the regression coefficients remain constant over time, that is:

$$H_0 : \boldsymbol{\beta}_t = \boldsymbol{\beta} \qquad t = 1, 2, \dots T$$

against the alternative that at least one coefficient varies over time.
In the statistical and econometric literature, testing for parameters instability in a regression framework has been treated using different approaches.
The classical test for structural change is the well known Chow Test (Chow, 1960). This testing procedure splits the sample into two sub-samples, estimates the parameters for each sub-sample, and then, using a classic F statistic, a test on the equality of the two sets of parameters is performed. For a review, which includes also some extensions in different contexts, see Andrews and Fair (1988).The principal issue of the Chow test is the assumption that the break-date must be known a priori. Generally this procedure is used by fixing an arbitrary candidate break-date or by selecting it on the base of some known feature of the data. However, the results can be highly sensitive to these arbitrary choices; in the first case, the Chow test may be uninformative and, in the second case, it can be misleading (Hansen 2001).
More recently, the literature has focused on a more realistic problem in which the number of break points and their location are supposed to be unknown (see Perron, 2006, for a survey). In this context, one of the major contribution is the strategy proposed by Bai and Perron (1998, 2003a, 2003b) who developed an iterative procedure which allows consistent estimation of the number and the location of the breaks points together with the unknown regression coefficients in each regime. In their procedure the breaks are considered deterministic parameters and so it is not required the specification of their underlying generating process. The number of breaks can be determined sequentially by testing for $q+1$ against q or using a global approach of testing for q against no breaks. However the procedure needs the specification of some restrictions such as the minimum distance between breaks and their maximum number.
Another approach to change point testing is based on the generalized fluctuation tests (for a survey see Zeileis *et al.*, 2003). Such an approach has the advantage of not assuming a particular pattern of deviation from the null hypothesis. Moreover, although it is possible in principal to carry out the location of the break points, this method is commonly used only to verify their presence; with this aim the fluctuation tests will be used in this paper.

3.1 Fluctuation tests

Although the generalized fluctuation test framework includes formal significance tests, it is commonly used in a graphic way. The general idea is to fit a regression model to the data and derive the empirical process that captures the fluctuation in the residuals or in the parameter estimates. Under the null hypothesis of constant regression coefficients, fluctuations are governed by functional central limit theorems (Kuan and Hornik 1995) and therefore boundaries can be found that are crossed by the corresponding limiting processes with fixed probability α. When the fluctuation of the empirical process increases, there is evidence of structural changes in the parameters. Moreover, its trajectory may also highlight the type of deviation from the null hypothesis as well as the dating of the structural breaks. As pointed out previously, the generalized fluctuation tests can be based on the residuals or on the parameter estimates of the regression model. The first class includes the classical CUSUM based on cumulative sums of recursive residuals (Brown *et al.*, 1975), the CUSUM test based on OLS residuals (Ploberger and Krämer, 1992) and the MOSUM tests (Chu *et al.*, 1995a). The second class includes the Recursive Estimates (RE) test (Ploberger *et al.*, 1989) and the Moving Estimates (ME) test (Chu *et al.*, 1995b). In both of them, the vector of unknown parameters is estimated recursively with a growing number of observations, in the RE test, or with a moving data window in the ME test and then compared to the estimates obtained by using the whole sample. Define:

$$\mathbf{y}'_{1:t} = (y_1, y_2, \ldots, y_t) \tag{13}$$

$$\mathbf{X}'_{1:t} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)' \tag{14}$$

and let

$$\widehat{\boldsymbol{\beta}}_{1:t} = (\mathbf{X}'_{1:t}\mathbf{X}_{1:t})^{-1}\mathbf{X}'_{1:t}\mathbf{y}_{1:t} \qquad t = p, p+1, \ldots, T \tag{15}$$

be the ordinary least squares (OLS) estimate of the regression coefficients based on the observations up to t.

The basic idea is to reject the null hypothesis of parameter constancy if these estimates fluctuate too much. Formally the test statistic is defined as:

$$S^{(T)} = \sup_{0 \le z \le 1} \left\| B^{(T)}(z) \right\|_\infty \tag{16}$$

with

$$B^{(T)}(z) = \frac{\phi(z)}{\widehat{\sigma} T}(\mathbf{X}'_{1:T}\mathbf{X}_{1:T})^{1/2}(\widehat{\boldsymbol{\beta}}_{1:\phi(z)} - \widehat{\boldsymbol{\beta}}_{1:T}) \tag{17}$$

where

$$\widehat{\sigma} = \left[\frac{1}{T-p} \sum_{t=1}^{T} (y_t - \mathbf{x}_t' \widehat{\boldsymbol{\beta}}_{1:T}) \right]^{1/2} \tag{18}$$

and $\phi(z)$ is the largest integer less than or equal to $p + z(T-p)$ and $\|.\|_\infty$ is the maximum norm. As proved in Ploberger *et al.* (1989), $B^{(T)}(z)$ is a p-dimensional stochastic process such that

$$B^{(T)}(z) \xrightarrow{\mathcal{D}} B(z) \tag{19}$$

where $B(z)$ is a Brownian Bridge.
The distribution of $\sup_{0 \leq z \leq 1} \|B(z)\|_\infty$ is given Billingsley (1999):

$$P\left(\sup_{0 \leq z \leq 1} \|B(z)\|_\infty \leq x \right) = \left[1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 x^2} \right]^p \qquad x \geq 0 \tag{20}$$

This result allows to construct the asymptotic boundaries crossing probabilities.

4 Forecasting Methods in presence of structural breaks

Once the parameter instability due to the presence of structural breaks has been detected, the problem is how to account for it when generating forecasts. Indeed parameter instability could cause forecast failures in macroeconomic and financal time series (for a survey see Clements and Hendry 2006).
When it is possible to identify the exact date of the last break, the standard solution is to use only observations over the post-break period. In practice, the dates of the break points are not known a priori and an estimation procedure has to be used. It could produce imprecise values which affect negatively the specification of the forecasting model and, as a consequence, poor performances of the forecasts. Furthermore, even if the last break date is correctly estimated, the forecasts generated by this scheme are likely to be unbiased and may not minimize the mean square forecast error (Pesaran and Timmermann 2007). Moreover, if the last break is detected close to the boundaries of the data sample, the parameters of the forecasting model are estimated with relatively short sample and the estimation uncertainty may be large.
However, as pointed out by Pesaran and Timmermann (2007), the pre-break observations are informative for forecasting even after the break. More specifically, it is appropriate to choose a high fraction of the pre-break observations especially when the break is small, the variance parameter increases at the break point and the number of post break observations is small.

Furthermore the forecasting performance is sensitive to the choice of the observation window. A relative long estimation window reduces the forecast error variance but increases its bias; on the other hand, a short estimation window produces an increase in the forecast error variance although the bias decreases. Therefore, an optimal window size should balance the trade off between an accurate estimate of the parameters and the possibility that the data come from different regimes.

Unfortunately, the selection of this single best estimation window, is not an easy task and, in many empirical studies, it is arbitrarily determined. However, Pesaran and Timmermann (2007) propose an optimal estimation window that leads to the lowest average out-of-sample loss over the cross-validation sample. More precisely, the cross-validation approach reserves some of the last observations for an out-of-sample estimation exercise and the estimation window is chosen in order to minimize the mean square forecast error on this sample.

This approach performs quite well if the size of the break is large. Alternatively, for relatively small breaks it is possible to combine forecasts generated from the same model but over different estimation windows. This approach is in the spirit of forecast combination obtained by estimating a number of alternative models over the same sample period (for a review see Timmermann, 2006). However, it can be used also to deal with other types of model uncertainty, such as the uncertainty over the size of the estimation window (Pesaran and Timmermann, 2007). In particular, Pesaran and Pick (2011) show that also a scheme based on forecast averaging over estimation window leads to a lower bias and root mean square forecast error than forecasts based on a single estimation window. Their results has been confirmed by an application to weekly returns on 20 equity index futures. This idea has been fruitfully applied in macroeconomic forecasting (see Assenmacher-Wesche and Pesaran, 2008 and Pesaran, *et al.*, 2009).

In view of the above considerations, in this paper the attention has been focused on the forecasts schemes generated from the same model but over different estimation windows. In particular for each of the considered realized volatility model, different forecast combinations have been considered focusing on those that do not incorporate explicitly the estimation of the break dates. Moreover, in the analysis one-step ahead forecasts have been considered and so it is assumed that no structural breaks occur in the forecast period (for forecasting with structural breaks over the forecast period see Pesaran *et al.*, 2006 and Maheu and Gordon, 2008).

Forecast combination with equal weights.

As pointed out previously, the forecast combination with equal weights is the simplest combination but it is robust to structural breaks of unknown break dates and sizes. Moreover it performs quite well especially when the break is of moderate magnitude and it is located close to the boundaries of the data sample (Pesaran and Pick, 2011).

Let ω be the minimum acceptable estimation window size. The forecast combination with equal weights is defined by:

$$\widehat{y}_{T+1} = \frac{1}{T-\omega}\sum_{\tau=1}^{T-\omega}\left(\mathbf{x}'_{T+1}\widehat{\boldsymbol{\beta}}_{\tau+1:T}\right) \tag{21}$$

Many researches (see for example Pesaran and Timmermann, 2007) have highlighted the advantages of this scheme; it has good performances also when there is uncertainty about the presence in the data of structural breaks. This approach avoids also any estimation procedure for the weights.

Forecast combination with location weights.

By looking at equation (21) it is evident that the weights in the equally weighted combination can be converted into weights on the sample observations $\mathbf{x}_t$. As discussed in Tian and Anderson (2014), the ω most recent observations are used in all of the forecasts, whereas the older observations are used less. Furthermore the influence of each observation is inversely proportional to its distance from the forecasting origin: the most recent data are usually more relevant especially if the regression parameters have significant changes close to the end of the sample.

A way to place heavier weights on the forecasts which are based on more recent data much more than under the equally weighted forecast combination is to use constant weights proportional to the location of τ in the sample.

More precisely, this combination, known as the forecast combination with location weights, is defined by:

$$\widehat{y}_{T+1} = \frac{1}{\sum_{\tau=1}^{T-\omega}\tau}\sum_{\tau=1}^{T-\omega}\tau\left(\mathbf{x}'_{T+1}\widehat{\boldsymbol{\beta}}_{\tau+1:T}\right) \tag{22}$$

Also in this case, no estimation of the weights is needed.

Forecast combination with MSFE weights

This approach, proposed by Pesaran and Timmerman (2007), is based on the idea that the weights of the forecasters obtained with different estimation windows should be proportional to the inverse of the associated out-of-sample mean square forecast error (MSFE) values. To this aim, a cross validation approach is used.

Common to all the considered methods is that the estimation window should not be smaller than a minimum length ω that can be set equal to the number of regressors plus one. However, as pointed out by Pesaran and Timmerman (2007), to account for the very large effect of parameter estimation error, ω should be at least 3 times the number of unknown parameters.

To better understand, let m be the generic start point of the estimation window and assume that $\tilde{\omega}$ is the number of the observations used in the cross validation set, that is the observations used to measure pseudo out-of-sample forecasting performance. The recursive pseudo out-of-sample MSFE value is computed as:

$$MSFE(m|T,\tilde{\omega}) = \tilde{\omega}^{-1} \sum_{\tau=T-\tilde{\omega}}^{T-1} \left(y_{\tau+1} - \mathbf{x}'_{\tau} \widehat{\boldsymbol{\beta}}_{m:\tau} \right) \tag{23}$$

The forecast combination with MSFE weights is than defined as:

$$\widehat{y}_{T+1} = \frac{\sum_{m=1}^{T-\omega-\tilde{\omega}} \left(\mathbf{x}'_{T} \widehat{\boldsymbol{\beta}}_{m:T} \right) (MSFE(m|T,\tilde{\omega}))^{-1}}{\sum_{m=1}^{T-\omega-\tilde{\omega}} (MSFE(m|T,\tilde{\omega}))^{-1}} \tag{24}$$

Together with the parameter ω, the length of the minimal estimation window, this method requires also the choice of the parameter $\tilde{\omega}$, the length of the evaluation window. If this parameter is set too large, too much smoothing may result and, as a consequence, in the combination, the forecasting based on older data will be preferred. On the other hand, if $\tilde{\omega}$ is set too short, although a more precise estimation of the MSFE can be obtained, the ranking of the forecasting methods is more affected by noise. Of course the selection of this parameter depends on the problem at hand and on the length of the series.

Forecast combination with ROC weights

This approach has been proposed by Tian and Anderson (2014). It is based on by-products of the Reverse Ordered CUSUM (ROC) structural break test considered by Pesaran and Timmermann (2002).

It is a two stages forecasting strategy. In the first step a sequence of ROC test statistics, starting from the most recent observations and going backwards in time, is calculated. Each point in the sample is considered as a possible most recent break point.

This test is related to the classical CUSUM test but, in this case, the test sequence is made in reverse chronological order. In particular, the time series observations are placed in reverse order and the standard CUSUM test is performed on the rearranged data set.

In the paper by Pesaran and Timmermann (2002), the test statistics are used to perform a formal structural break test and to estimate the last breakpoint in the sample.

In the second step, the ROC statistics are used to weight the associated post break forecast, developing a forecast combination. Moreover, the weights do not depend on finding and dating a structural break but they are constructed in order to give more weights to observations subsequent to a potential structural break.

In the first step of the procedure, for $\tau = T-\omega+1, T-\omega, \ldots, 2, 1$, let

$$\mathbf{y}'_{T:\tau} = (y_T, y_{T-1}, \ldots, y_{\tau+1}, y_\tau) \tag{25}$$

$$\mathbf{X}'_{T:\tau} = (\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_{\tau+1}, \mathbf{x}_\tau) \tag{26}$$

be the observation matrices, and let

$$\widehat{\boldsymbol{\beta}}^{(R)}_{T:\tau} = (\mathbf{X}'_{T:\tau}\mathbf{X}_{T:\tau})^{-1}\mathbf{X}'_{T:\tau}\mathbf{y}_{T:\tau} \tag{27}$$

be a sequence of least squares estimates of $\boldsymbol{\beta}$ associated with the reverse-ordered data sets. The ROC test statistics s_t are defined as:

$$s_\tau = \frac{\sum_{t=\tau}^{T-\omega} \xi_t^2}{\sum_{t=1}^{T-\omega} \xi_t^2} \qquad \text{for } \tau = T-\omega, T-\omega-1, \ldots, 2, 1 \tag{28}$$

where ξ_t are the standardized one-step-ahead recursive residuals defined as:

$$\xi_t = \frac{y_t - \mathbf{x}'_t \widehat{\boldsymbol{\beta}}^{(R)}_{T:t+1}}{(1+\mathbf{x}'_t(\mathbf{X}'_{T:t+1}\mathbf{X}_{T:t+1})^{-1}\mathbf{x}_t)^{1/2}} \tag{29}$$

In the second step of the procedure, all dates τ are considered as possible choices for the last breakpoint. The combination weight on each τ is constructed as:

$$cw_\tau = \frac{\left|s_\tau - \left(\frac{T-\omega-\tau+1}{T-\omega}\right)\right|}{\sum_{\tau=1}^{T-\omega}\left|s_t - \left(\frac{T-\omega-\tau+1}{T-\omega}\right)\right|} \qquad \tau = 1, 2, \ldots, T-\omega \tag{30}$$

Since, under the null hypothesis of no structural break in τ, it is:

$$E(s_\tau) = \frac{T-\omega-\tau+1}{T-\omega} \tag{31}$$

the combination weights vary according to the absolute distances between s_τ and its expected value. As a consequence, cw_τ is larger if this distance is large, that is if the evidence of a structural break is stronger. On the contrary, if in τ there is no evidence of substantial breakpoint, the associated weight is small.

Moreover, the weights do not depend on finding and dating a structural break. However, if the absolute values of the difference between the ROC statistics and their expectation, under the null, start to grow (giving evidence of a potential structural break) the weights cw_τ increase giving more weights to the observations on data subsequent to τ.

The one-step ahead forecast based on ROC statistics is defined as:

$$\widehat{y}_{T+1} = \sum_{\tau=1}^{T-\omega} \left(cw_\tau (\mathbf{x}'_{T+1} \widehat{\boldsymbol{\beta}}_{\tau+1:T}) \right) \tag{32}$$

Forecast combination with ROC location weights

In order to take into account a prior belief on the probability that a time τ could be the most recent break point, it is possible, in the definition of ROC weights, to incorporate an additional weight function l_τ.

The new weights are defined as:

$$cw_\tau = \frac{\left| s_\tau - \left(\frac{T-\omega-\tau+1}{T-\omega}\right)\right| l_\tau}{\sum_{\tau=1}^{T-\omega} \left| s_t - \left(\frac{T-\omega-\tau+1}{T-\omega}\right)\right| l_\tau} \qquad \tau = 1, 2, \ldots, T-\omega \tag{33}$$

For example, if a single break point seems to be equally likely at each time point, the natural choice is $l_\tau = 1$ for $\tau = 1, 2, \ldots, T-\omega$. In this case the weights depend only on the magnitude of the ROC statistics and the combination defined in (32) is obtained.

However, in a forecasting context, where the identification of the most recent break is essential, the prior weight l_τ could be chosen as an increasing function of the location of time τ in the full sample. In the spirit of forecast combination with location weights, the most natural choice is $l_\tau = \tau$. Of course different specifications are also allowed.

5 Comparing forecasting performances

In order to evaluate the advantages of the defined forecast combinations, for each RV-models, it is necessary to compare them in terms of predictive performances.

The natural way to compare the forecasting abilities of alternative methods is to perform out-of-sample validation in which the sample is divided in such a way that the first R observations are used for model estimation and the last P for forecasting evaluation.

The comparison is made by using a loss function. For univariate forecast, at time t, the loss function can be specified as:

$$L_{i,t} = L(y_t, \widehat{y}_t^{(i)}) \tag{34}$$

where y_t is the actual value at time t and $\widehat{y}_t^{(i)}$ is a forecast obtained using a particular method (i).

Loss functions are a crucial ingredient for assessing the accuracy of forecasts and for discriminating across competing forecasting methods. A generic loss function should satisfy these three proprieties: a) it is a non negative function; b) it is 0 if and only if the forecast equals the actual (no error and no loss); c) an increasing penalty should be applied to increasing forecast errors. Patton (2011) details the most used loss functions in a volatility context and gives evidence on the ability of the loss functions to identify superior forecasts.

The loss functions can be used to rank competing methods in term of forecasting accuracy. To this aim the average loss function for each method is usually calculated and the method with the lowest value is considered to be the "best".

Ranking the loss measures is the basic tool to discriminate among alternative methods. However, this procedure does not indicate whether the differences among the performances of the comparing methods are statistically significant.

The problem of assessing statistically if the differences in the performances are relevant, has received a growing interest in the econometric literature and, to this aim, a number of statistical tests has been proposed. Diebold and Mariano (1995) has reviewed some of these tests such as the F-test, the Morgan-Granger Newbold test and the Meese-Rogoff test. They have found that these tests can be negatively influenced by non-Gaussianity, contemporaneous and/or serial correlation of the forecast errors and small sample size. To solve these problems, the tests of equal predictive ability by Diebold and Mariano (1995) and by West (1996) have been proposed. Both of them test the hypothesis that two models have equal predictive ability; the difference between them is that the Diebold and Mariano test treats the forecast as given while the West test accounts for the use of estimated parameters in the forecasting model. The advantages of these tests are that they do not require specific assumptions on the forecast errors and they are flexible about the specification of the loss function, in the sense that the loss function needs not be quadratic and symmetric.

However some problems arise when there are a large number of competing forecast methods. White (2000) has highlighted that in this case data snooping can arise, a problem which occurs when a given set of data is used more than once for inference or model selection. He has proposed a straightforward approach, the reality check procedure, for testing the null hypothesis that the best model has no predictive superiority over a given benchmark model. This permits data snooping to be undertaken with some degree of confidence.

Even if the problem of data snooping is solved, Hansen (2005) has pointed out that this test is affected by the inclusion of poor performing models. For this reason he has introduced the superior predictive ability test; this test improves the reality check test by using a studentized test statistic that reduces the influence of erratic forecasts and invoke a sample-dependent null

Hansen (2005) has shown that the Diebold and Mariano test and the West test coincide if the forecasts are treated as given.

distribution.
However the superior predictive ability test only provides evidence against a benchmark which, in many cases, could be difficult to specify. A solution to the limitation of the superior predictive ability test is the model confidence set approach, proposed by Hansen *et al.* (2003; 2011).

5.1 The Model Confidence Set

The goal of this procedure, which does not require the specification of a benchmark explicitly, is to determine which models, from an initial set M^0 of models indexed by $i = 1, \ldots, M^0$, exhibit the same predictive ability in term of a loss function. Following Hansen *et al.* (2005), the term *model* is used loosely. It can refer to a model, a method, or a rule. In the following it is used to indicate a forecasting method associated to a RV-model.
The model confidence set is a stepwise procedure which starts by setting $M = M^0$. The algorithm is based on an equivalence test δ_M and an elimination rule e_M . The equivalence test assumes the value 0 if a properly defined null hypothesis $H_{0,M}$ in not rejected at significance level α and 1 otherwise. In the latter case the model that is to be removed from M is identified by the elimination rule e_M and the set M is updated. The procedure is repeated until the null hypothesis $H_{0,M}$ is not rejected. The collection of the optimal models at level $1-\alpha$, denoted with $\widehat{M}^*_{1-\alpha}$, contains the surviving models.
Despite its sequential nature, the model confidence set procedure does not accumulate type I error. This is due to the fact that the test stops when the first hypothesis is not rejected.
To compare the relative performances of the models under comparison, for every $i = 1, \ldots, M^0$ we specify $L_{i,t}$, a loss function associated with model i and time t, where $t = R+1, R+2, \ldots, T$ while R is the length of the in-sample period and it will further discussed in section 6.2.
Let

$$d_{ij,t} = L_{i,t} - L_{j,t} \quad \forall i, j \in M \tag{35}$$

be the differential loss between model i and j. The hypothesis compared in each significance test are:

$$\mathrm{H}_{0,M} : \mathrm{E}(d_{ij,t}) = 0 \quad \text{for all } i, j \in M \tag{36}$$

$$\mathrm{H}_{A,M} : \mathrm{E}(d_{ij,t}) \neq 0 \quad \text{for some } i, j \in M$$

In order to test the hypothesis (36), let $\bar{L}$ be the loss of model i averaged over the forecasting period:

$$\bar{L}_i = \frac{1}{T-R} \sum_{t=R+1}^{T} L_{i,t} \tag{37}$$

$\overline{d}_{ij}$ the average loss of model i relative to model j:

$$\overline{d}_{ij} = \frac{1}{T-R} \sum_{t=R+1}^{T} d_{ij,t} = \bar{L}_i - \bar{L}_j \tag{38}$$

and $\overline{d}_{i.}$ the average loss of model i relative to all models:

$$\overline{d}_{i.} = \frac{1}{m} \sum_{j \in M} \overline{d}_{ij} = \bar{L}_i - \frac{1}{m} \sum_{j \in M} \bar{L}_j \tag{39}$$

where m is the total number of models in M.
The hypothesis (36) maps naturally into the test statistic:

$$T_{max,M} = \max_{i \in M} t_{i.} \tag{40}$$

where

$$t_{i.} = \frac{\overline{d}_{i.}}{\sqrt{\widehat{var}(\overline{d}_{i.})}} \tag{41}$$

and $\widehat{var}(\overline{d}_{i.})$ denotes the estimate of the variance of $\overline{d}_{i.}$.
When using this test statistic, it is necessary to define an elimination rule. The most reasonable choice associated to the $T_{max,M}$ statistic is:

$$e_{max,M} = \underset{i \in M}{\operatorname{argmax}}\, t_i \tag{42}$$

Since the distribution, under the null hypothesis, of the test statistic is not known, the critical values of the test can not be determined analytically and re-sampling techniques have to be used. In the context of model confidence set, the block bootstrap is generally used. It also allows to obtain the estimate of the variances in (41) and the estimation of the p-value associated to the null hypothesis $H_{0,M}$.
The basic idea of the block bootstrap procedure is to resample randomly with replacement blocks of consecutive observations so that the original time series structure is preserved in each block. The blocks are assembled together in random order to obtain a simulated version of the original series.
The block bootstrap procedure works reasonably well under very weak conditions on the de-

pendency structure of the data. Moreover it does not require specific assumptions on the data generating process. As a consequence, it has been applied to a very broad range of applications especially in the context of time series analysis.
According to Hansen *et al.* (2011), the block bootstrap procedure can be adapted in the context of model confidence set as shown in algorithm 1 reported in the appendix.

6 Empirical application

The data is obtained from the Oxford-Man Institute's realised library. It consists of 5 minute realized volatility and daily returns of four U.S. stock market indices: S&P 500, Dow Jones Industrial Average, Russell 2000 and Nasdaq 100.
The 5 minute realized volatility is calculated by the sum of squared 5-minute returns within a day;

$$RV^{(5)} = \sum_{i=1}^{m} r^2_{t-1+5i,t-1+5(i-1)} \tag{43}$$

where

$$r^2_{t-1+5i,t-1+5(i-1)} = p_{t-1+5i} - p_{t-1+5(i-1)} \tag{44}$$

is the five-minute return computed as the log-difference of two consecutive index values over five-minutes, multiplied by 100; p_t denotes the logarithm of the index value; m is the number of five-minute returns per day calculated as $m = 390/5$ being 390 the total minutes in a day.
The choice of 5-minute data is justified on the grounds of past empirical findings that show that at this frequency there is no evidence of micro-structure noise (Andreou *et al.*, 2012). Moreover, as shown in Liu *et al.* (2015), no other more sophisticated measures outperform 5-minute RV in terms of estimation accuracy.
The sample covers the period from January 1, 2012 to February 4, 2016; the plot of the RV series and the log-RV series are reported in Figure 1.

6.1 Evidences of structural breaks: in-sample analysis

In order to investigate the constancy of the regression coefficients in all the considered models (HAR-RV, LHAR-RV and AHAR-RV), an analysis based on the recursive estimates test has been employed for all the four series. The results are reported in Table 1. It is evident the rejection of the null hypothesis of constancy of the regression parameters for all the models and for all the series.

S&P 500			**DJIA**		
Model	statistic	p-value	Model	statistic	p-value
HAR-RV	2.369	0.0001	HAR-RV	2.167	0.0006
LHAR-RV	2.198	0.0010	LHAR-RV	1.943	0.0084
AHAR-RV	2.156	0.0011	AHAR-RV	2.089	0.0019
RUSSELL 2000			**NASDAQ 100**		
Model	statistic	p-value	Model	statistic	p-value
HAR-RV	2.216	0.0004	HAR-RV	1.848	0.0086
LHAR-RV	1.997	0.0048	LHAR-RV	1.659	0.0104
AHAR-RV	2.044	0.0028	AHAR-RV	1.782	0.0093

Table 1: Recursive Estimates Test

Moreover, in Figure 2, 3 and 4 the fluctuation process, defined in equation (16), is reported for each model specification and for each series, along with the boundaries obtained by its limiting process at level $\alpha = 0.05$. The paths of the empirical fluctuation process confirm the parameters instability: the boundaries are crossed for all the series and for all the models. Moreover they give also a graphical indication on the location and the size of the breaks.
The above analysis seems to confirm the effectiveness of parameter instability in all the three RV-models for all the considered indices. Moreover, it suggests the use of forecasting methods which take into account the presence of structural breaks.

6.2 Forecasting evaluation: out-of-sample analysis

In order to evaluate and to compare the forecasting performances of the proposed forecast combinations for each class of model, an initial sub-sample, composed by the data from $t = 1$ to $t = R$, is used to estimate the model and the 1-step ahead out-of-sample forecast is produced. The sample is than increased by one, the model is re-estimated using data from $t = 1$ to $t = R + 1$ and 1-step ahead forecast is produced. The procedure continues until the end of the available out-of-sample period. In the following examples R has been fixed such that the number of out of sample observations is 300. To generate the 1-step ahead forecasts the five competing forecast combinations, defined in section 4, have been considered together with a a benchmark method.
As a natural benchmark we refer to the Expanding window method, that ignores the presence of structural breaks. In fact it uses all the available observations. As pointed out in Pesaran and Timmermann (2007), this choice is optimal in situations with no breaks and it is appropriate for forecasting when the data is generated by a stable model. For each class of models, this

method produces out-of-sample forecasts using a recursive expanding estimation window.
Common to all the considered forecast combinations is the specification of the minimum acceptable estimation window size. As highlighted in section 4 it should be at least three times the number of unknown parameters; for simplicity this parameter has been held fixed at 40 for all the RV-model specifications.
Moreover, for the MSFE weighted average combination, the length of the evaluation window has been held fixed at 100. This value allows a good estimation of the MSFE and, at the same time, a not excessive loss of data at the end of the sample where a change point could be very influential for the forecasting.
In order to evaluate the quality of the volatility forecasts the MSE and the QLIKE loss functions have been considered. Their are defined as:

$$\mathrm{MSE}_t = (y_t - \widehat{y}_t)^2 \tag{45}$$

$$\mathrm{QLIKE}_t = \frac{y_t}{\widehat{y}_t} - \log\left(\frac{y_t}{\widehat{y}_t}\right) - 1 \tag{46}$$

where y_t is the actual value of the 5-minute RV at time t and $\widehat{y}_t$ is the corresponding RV forecast. They are the most widely used loss functions and they provide robust ranking of the models in the context of volatility forecasts (Patton, 2011).
The QLIKE loss is a a simple modification of the Gaussian log-likelihood in such a way that the minimum distance of zero is obtained when $y_t = \widehat{y}_{t+1}$. Moreover, according to Patton (2011), it is able to better discriminate among models and it is less affected by the most extreme observations in the sample .
These loss functions have been used to rank the six competing forecasting methods for each RV-model specification. To this aim, for every methods, the average loss (37) and the ratio between its value and the average loss of the benchmark method have been computed. Obviously, a value of the ratio below unit indicates that the forecasting method beats the benchmark according to the loss function metric.
Moreover, to assess statistically if the differences in the performances are relevant, the model confidence set procedure, introduced in the previous section, has been implemented in R-code using the package MCS (Catania and Bernardi, 2015) with $\alpha = 0.10$. As pointed out, the model confidence set procedure uses extensively the block bootstrap to obtain the critical values of the test as well as the estimation of the variance useful to construct the test statistic. In this context, the block bootstrap length, according to Hansen *et al.* (2011), has been determined by the maximum number of significant parameters resulted after fitting an AR(p) process on all the loss differences with a minimum length of the the blocks equal to 3. The number of

bootstrapped samples has been fixed to 5000.
In the following the results obtained for the three different model specifications for the Realized Volatility of the four considered series are reported and discussed.
Table 2, 3 and 4 provide the results of the analysis for the HAR-RV, LHAR-RV and AHAR-RV model specification respectively for the four considered series. In particular they report the average MSE and the average QLIKE for each forecasting method, as well as their ratio for an individual forecasting method to the benchmark expanding window method and the ranking of the considered method with respect to each loss function. Moreover it is also reported the value of the test statistic T_{MAX} of the model confidence set approach and the associate p-value.

6.2.1 Results for the HAR model

For all the considered series and for both the loss functions there is a significant evidence that all the forecast combinations have better forecasting ability with respect the expanding window procedure; the ratio values are all less than one. Moreover, for both the loss functions, the best method is the forecast combination with ROC location weights for S&P500 and Dow Jones Industrial Average and with location weights for Russell 2000 and Nasdaq 100; this result confirms the importance of placing heavier weights of the forecast based on more recent data. The model confidence set has the same structure for all the four series and for both the loss functions; it excludes only the forecast generated by the expanding window procedure and includes all the forecast combinations.

6.2.2 Results for the LHAR model

Focusing on the result of S&P500 and Dow Jones Industrial Average, which have very similar behaviour, the forecast combination with MSFE weights offers the best improvement in forecasting accuracy according to the MSE metric while the forecast combination with ROC weights according to the QLIKE metric. For Russell 2000 the forecast combination with ROC location weights beats all the competing models according to MSE loss function while, under the QLIKE, the best method is the forecast combination with location weights. This last method has better performance with respect to both of the loss function metrics for Nasdaq 100.
By looking at the MSE ratios for S&P500 and Dow Jones Industrial Average Index it is evident that the forecast combination with location weights is unable to beat the expanding window procedure; for all the others, the combinations are able to consistently outperform it.
In the model confidence set, when the MSE loss function is used, the expanding window is eliminated from the model confidence set for all the series. However it is also eliminated the

forecast combination with location weights for S&P500 and Dow Jones Industrial Average and that with MSFE weights for Russell 2000. A quite different situation arises when the loss function QLIKE is considered. In this case the only surviving models in the model confidence set are the two combinations based on ROC statistics for S&P500 and Dow Jones Industrial Average and those based on ROC location weights and on location weights for Russell 2000. For Nasdaq 100 all the combinations have the same forecasting accuracy. Excluding the last case, the QLIKE loss function, as pointed out previously, seems to better discriminate among forecasting methods.

6.2.3 Results for the AHAR model

In line with the previous results, the expanding window appears to offer the worst forecasting performance overall.
Moreover, for both MSE and QLIKE loss functions, the method which offers the major improvement in forecasting accuracy is the forecast combination with ROC location weights for S&P500 and Dow Jones Industrial Average Index and the the forecast combination with location weights for Russell 2000 and Nasdaq 100.
Focusing on the model confidence set, for the MSE loss function the expanding window is always eliminated for all the series together with the the forecast combination with MSFE weights for Russell 2000. With respect to the QLIKE loss function, for Dow Jones Industrial Average Index and Nasdaq 100 the only excluded method is the expanding window procedure. For the other series, the results confirms the better discriminate property of the QLIKE metric; the only surviving methods are forecast combination with ROC location weights and with location weights.

In conclusion, even if it is not clear which combination has the best forecasting performance, the forecast combination with ROC location weights and that with location weights seems to be always among the best methods. However the forecast combination with ROC location weights always outperform the expanding window method, it is always in the top position with respect the loss function ratio and it is never excluded by the model confidence set.

7 Concluding remarks

This paper has explored the relevance of taking into account the presence of structural breaks in forecasting realized volatility. The analysis has been based on 5-minute realized volatility of four U.S. stock market indices: S&P 500, Dow Jones Industrial Average, Russell 2000 and Nasdaq 100. Three different model specifications of the log-realized volatility have been considered. For all the considered market indices, the instability in the parameters of the RV-models has been verified through the recursive estimates test. In order to handle this problem, five forecast combinations, based on different estimation windows with alternative weighting schemes, have been introduced and compared with the expanding window method, a natural choice when the data are generated by a stable model. The forecasting performances have been evaluated, for each RV-model specification, through two of the most relevant loss functions, the MSE and the QLIKE. Moreover, to this aim the average loss function has been calculate and, to assess statistically if the differences in the performances are relevant, the model confidence set approach has been considered.

The analysis, repeated for each class of RV-models separately, has highlighted the importance of taking into account structural breaks; in fact the expanding window appears to offer the worst forecasting performances overall. In particular in almost all the considered cases the two combinations which make adjustments for accounting the possible most recent break point (the forecast combination with location weights and with ROC location weights) are placed in first position and, as a consequence, have better forecasting performances. Nevertheless the forecast combination with ROC location weights always outperform the expanding window method, it is always in the top position with respect the loss function ratio and it is never excluded by the model confidence set.

References

[1] Andersen TG, Bollerslev T, Diebold FX. 2007. Roughing It Up: Including Jump Components in the Measurement, Modeling anf Forecasting of Returns Volatility. *The Review of Economics and Statistics* **89(4)**: 701-720.

[2] Andreou E, Ghysels E, Kourouyiannis C. 2012. *Robust volatility forecasts in the presence of structural breaks* (No. 08-2012). University of Cyprus Department of Economics.

[3] Andrews DWK, Fair RC. 1988. Inference in Nonlinear Econometric Models with Structural Change. *The Review of Economic Studies* **55(4)**: 615-639

[4] Asai M, McAleer M, Medeiros MC. 2012. Asymmetry and long memory in volatility modeling. *Journal of Financial Econometrics* **10(3)**: 495-512.

[5] Assenmacher-Wesche K, Pesaran MH. 2008. Forecasting the Swiss economy using VECX models: An exercise in forecast combination across models and observation windows. *National Institute Economic Review* **203(1)**: 91-108.

[6] Bai J, Perron P. 1998 . Estimating and testing linear models with multiple structural changes. *Econometrica* **66(1)**: 47-78.

[7] Bai J, Perron P. 2003a. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* **18(1)**: 1-22.

[8] Bai J, Perron P. 2003b. Critical values for multiple structural change tests. *Econometrics Journal* **6(1)**: 72-78.

[9] Billingsley P. 1999. *Convergence of Probability Measures second edition* John Wiley & Sons, NewYork

[10] Brown RL, Durbin J, Evans JM. 1975. Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society* **37(2)**: 149-163.

[11] Catania L, Bernardi M. 2015. MCS: Model Confidence Set Procedure. R package. http://CRAN.R-project.org/package=MCS

[12] Chow GC. 1960. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica* **28(3)**: 591-605.

[13] Chu CSJ, Hornik K, Kuan CM. 1995a. MOSUM Tests for Parameter Constancy. *Biometrika* **82(3)**: 603-617.

[14] Chu CSJ, Hornik K, Kuan CM. 1995b. The Moving-Estimates Test for Parameter Stability. *Econometric Theory* **11(4)**: 669-720.

[15] Clements MP, Hendry DF. 2006. Forecasting with breaks. *In Handbook of Economic Forecasting*, ed. G. Elliot, C.W.J. Granger, and A. Timmermann (Elsevier) 605-657.

[16] Corsi F. 2004. A Simple Long Memory Model of Realized Volatility. *Available at SSRN 626064.*

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

[18] Corsi F, Renó R. 2010. HAR Volatility Modelling with Heterogeneous Leverage and Jumps. *Available at SSRN 1316953*.

[19] Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13(3)**: 253-263.

[20] Hansen BE. 2001. The new econometrics of structural change: Dating breaks in US labor productivity. *The Journal of Economic Perspectives* **15(4)**: 117-128.

[21] Hansen PR. 2005. A test for superior predictive ability. *Journal of Business & Economic Statistics* **23(4)**: 365-380.

[22] Hansen PR, Lunde A, Nason JM. 2003. Choosing the best volatility models: The model confidence set approach. *Oxford Bulletin of Economics and Statistics* **65(1)**: 839-861

[23] Hansen PR, Lunde A, Nason JM. 2005. Model confidence sets for forecasting models. *Working Paper, Federal Reserve Bank of Atlanta.*

[24] Hansen PR, Lunde A, Nason JM. 2011. The model confidence set. *Econometrica* **79(2)**: 453-497.

[25] Kuan CM, Hornik K. 1995. The generalized fluctuation test: A unifying view. *Econometric Reviews* **14(2)**: 135-161.

[26] Liu C, Maheu JM. 2008. Are there structural breaks in realized volatility?. *Journal of Financial Econometrics* **6(3)**: 326-360.

[27] Liu LY, Patton AJ, Sheppard K. 2015. Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics* **187(1)**: 293-311.

[28] Maheu JM, Gordon S. 2008. Learning, forecasting and structural breaks. *Journal of Applied Econometrics* **23(5)**: 553-583.

[29] Patton AJ. 2011. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* **160(1)**: 246-256.

[30] Perron P. 2006. Dealing with structural breaks. *Palgrave handbook of econometrics vol I*: 278-352.

[31] Pesaran MH, Pettenuzzo D, Timmermann A. 2006. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies* **73(4)**: 1057-1084.

[32] Pesaran MH, Pick A. 2011. Forecast combination across estimation windows. *Journal of Business and Economic Statistics* **29(2)**: 307-318.

[33] Pesaran MH, Schuermann T, Smith LV. 2009. Forecasting economic and financial variables with global VARs. *International Journal of Forecasting* **25(4)**: 642-675.

[34] Pesaran MH, Timmermann A. 2002. Market Timing and Return Prediction under Model Instability. *Journal of Empirical Finance* **9(5)**: 495-510.

[35] Pesaran MH, Timmermann A. 2007. Selection of estimation windows in the presence of breaks. *Journal of Econometrics* **137(1)**: 134-161.

[36] Ploberger W, Krämer W. 1992. The CUSUM Test With OLS Residuals. *Econometrica* **60(2)**: 271-285.

[37] Ploberger W, Krämer W, Kontrus K. 1989. A new test for structural stability in the linear regression model. *Journal of Econometrics* **40(2)**: 307-318.

[38] Tian J, Anderson HM. 2014. Forecast Combinations under Structural Break Uncertainty. *International Journal of Forecasting* **30(1)**: 161-175.

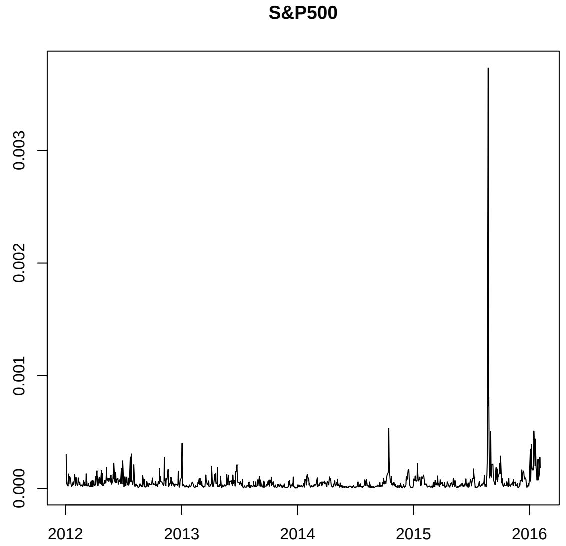
[39] Timmermann A. 2006. Forecast combinations. *In Handbook of Economic Forecasting*, ed. G. Elliot, C.W.J. Granger, and A. Timmermann (Elsevier) 135-196.

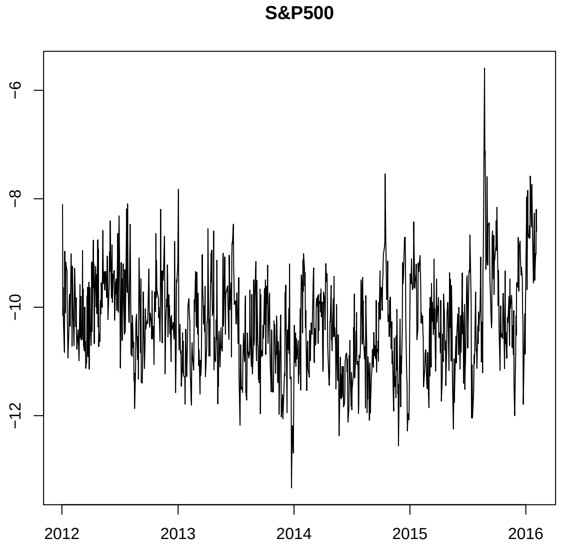
[40] West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64(5)**: 1067-1084.

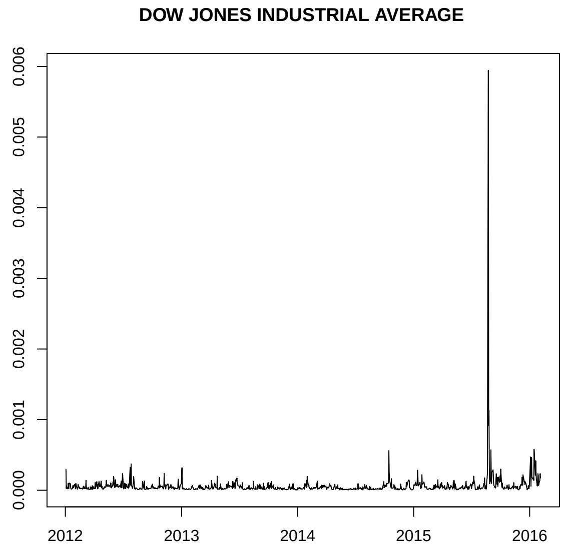
[41] White H. 2000. A reality check for data snooping. *Econometrica* **68(5)**: 1097-1126.

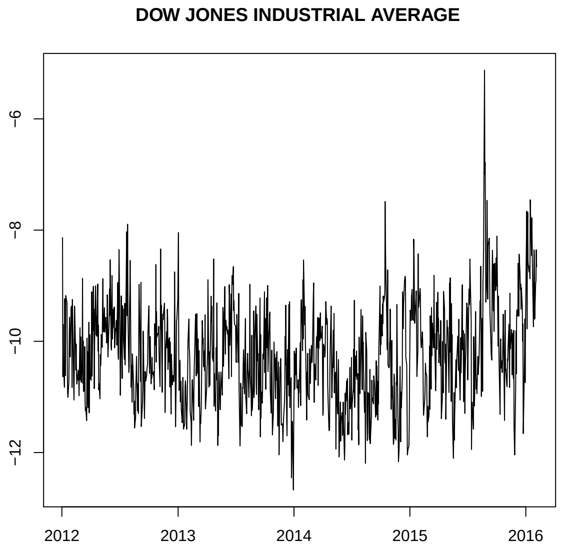
[42] Zeileis A, Kleiber C, Kramer W, Hornik K. 2003. Testing and dating of structural changes in practice. *Computational Statistics and Data Analysis* **44(1)**: 109-123.

Figures and Tables









continued on next page

continued from previous page

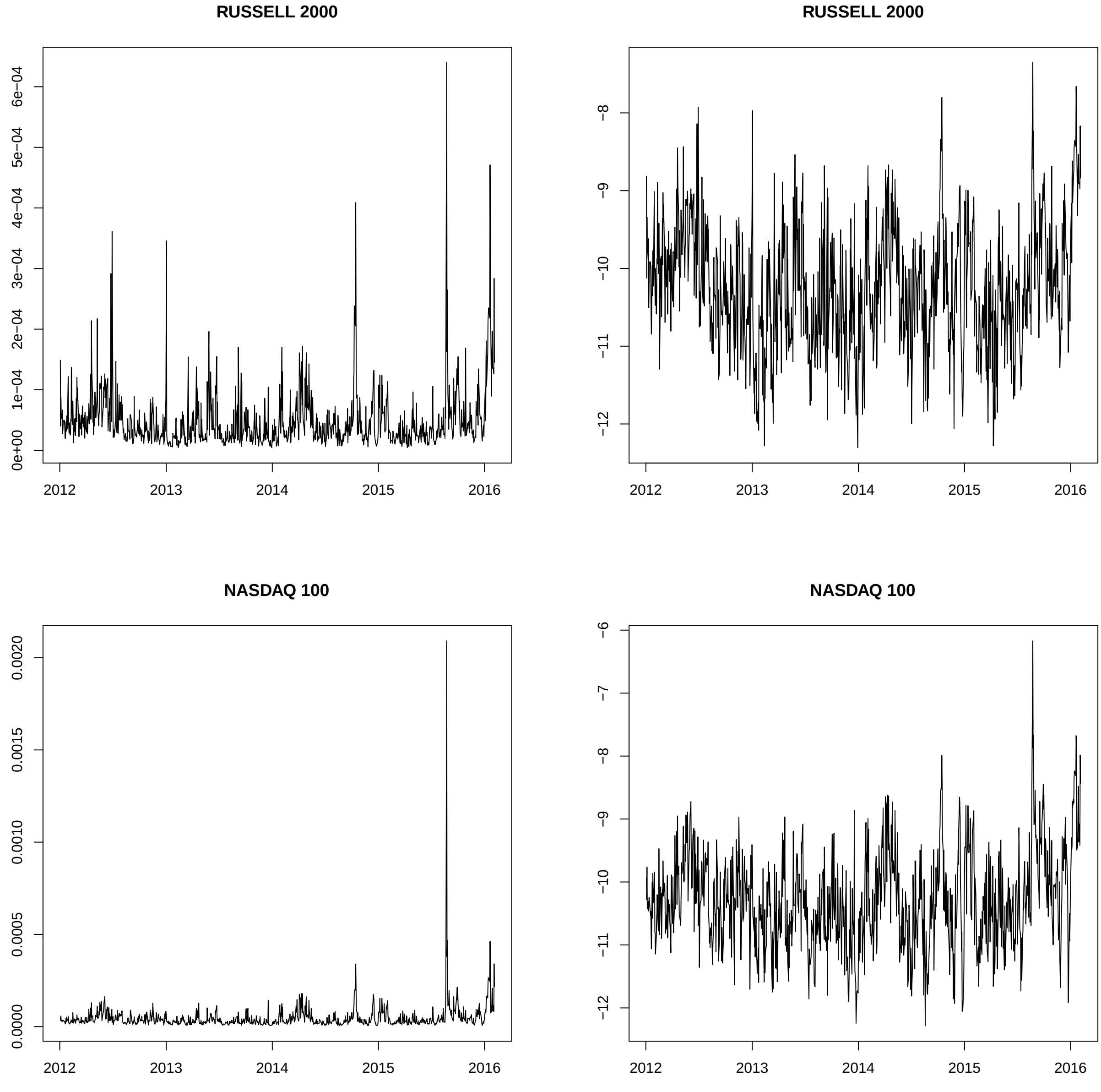


Figure 1: 5 minutes RV (on the left) and log RV (on the right) for S&P500, Dow Jones Industrial Average, Russell 2000 and Nasdaq 100

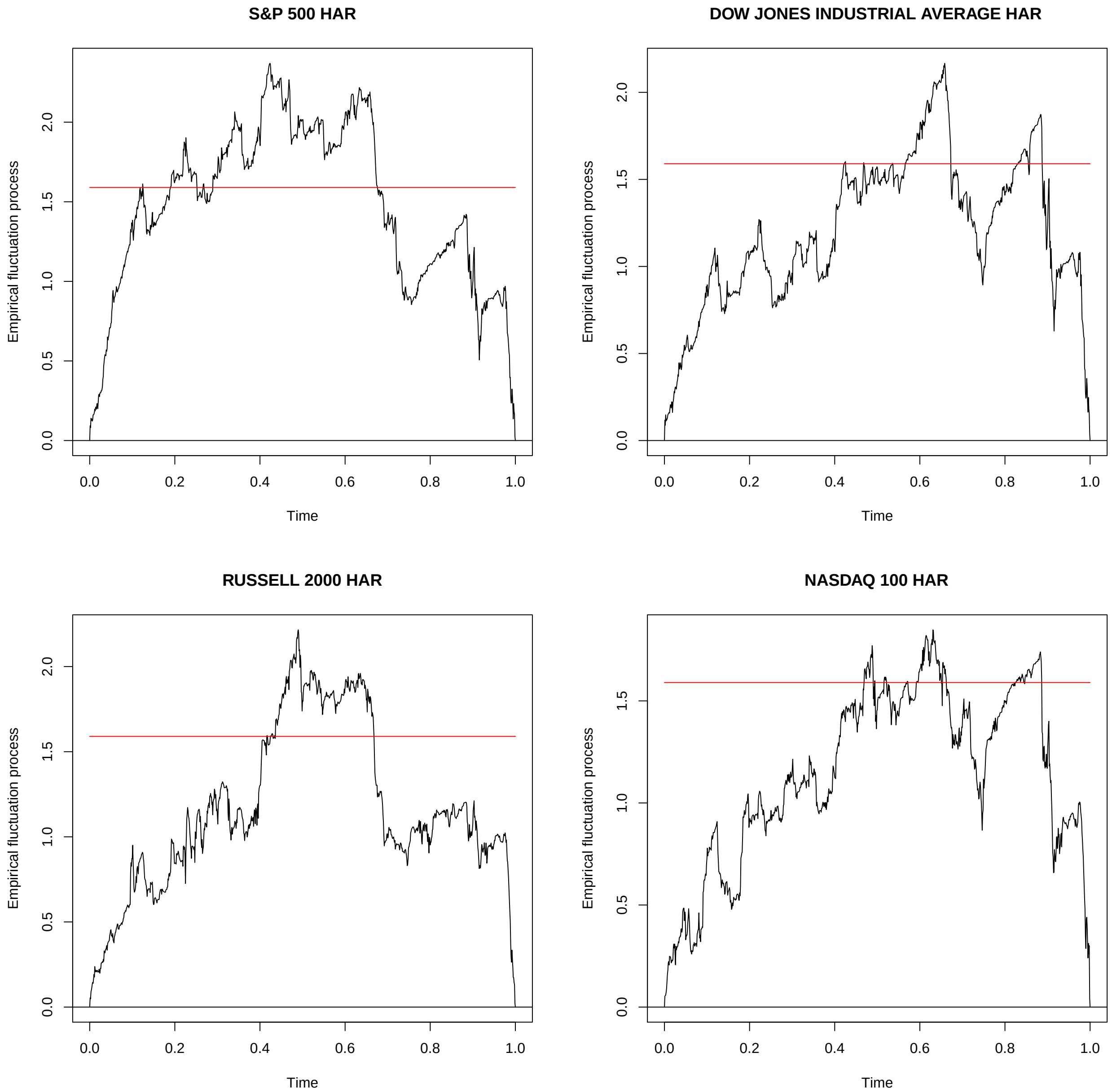


Figure 2: Empirical fluctuation process of the HAR-RV model parameters for S&P500, Dow Jones Industrial Average, Russell 2000 and Nasdaq 100. The red line refers to the boundary with $\alpha = 0.05$.

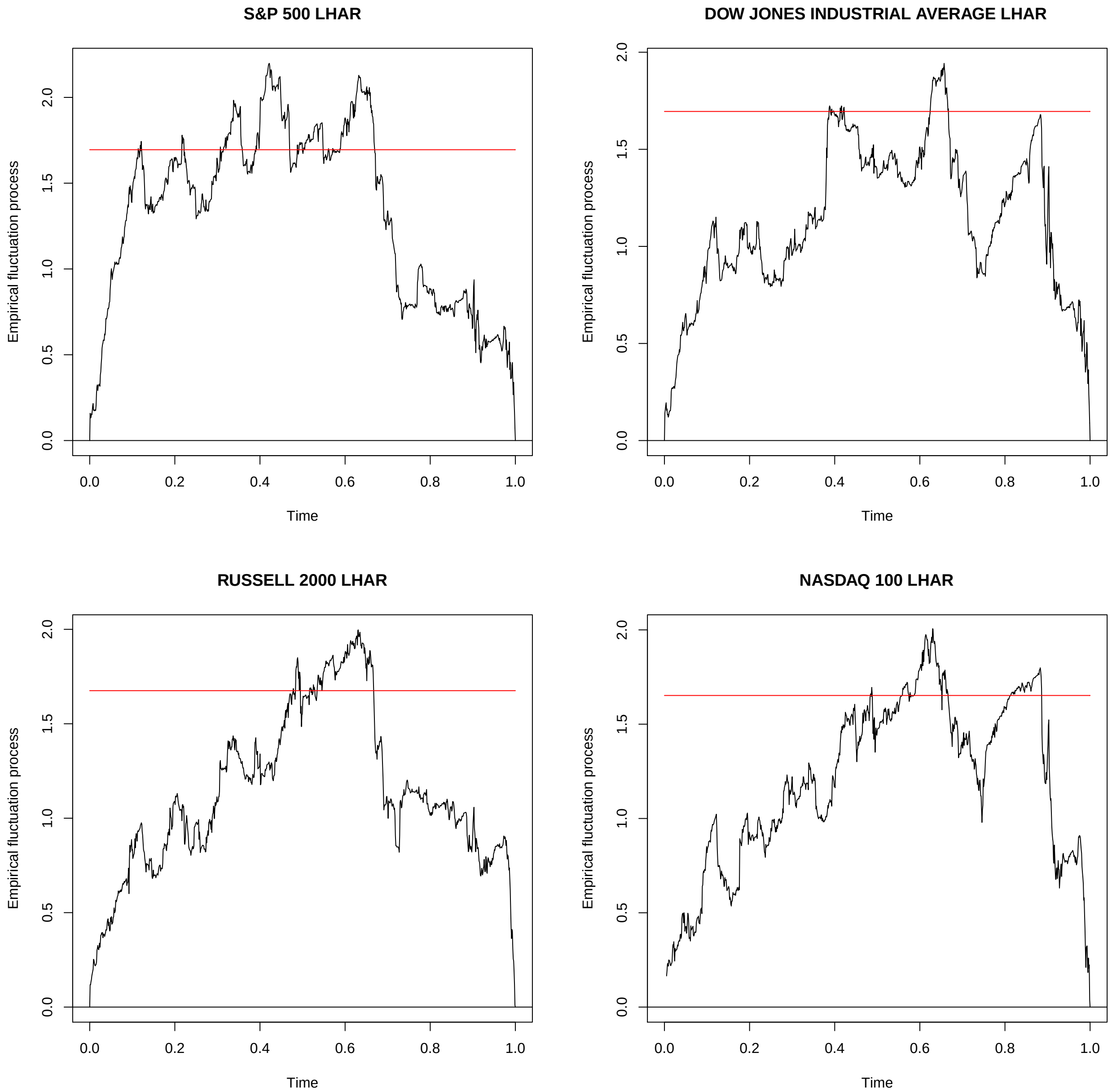


Figure 3: Empirical fluctuation process of the LHAR-RV model parameters for S&P500, Dow Jones Industrial Average, Russell 2000 and Nasdaq 100. The red line refers to the boundary with $\alpha = 0.05$.

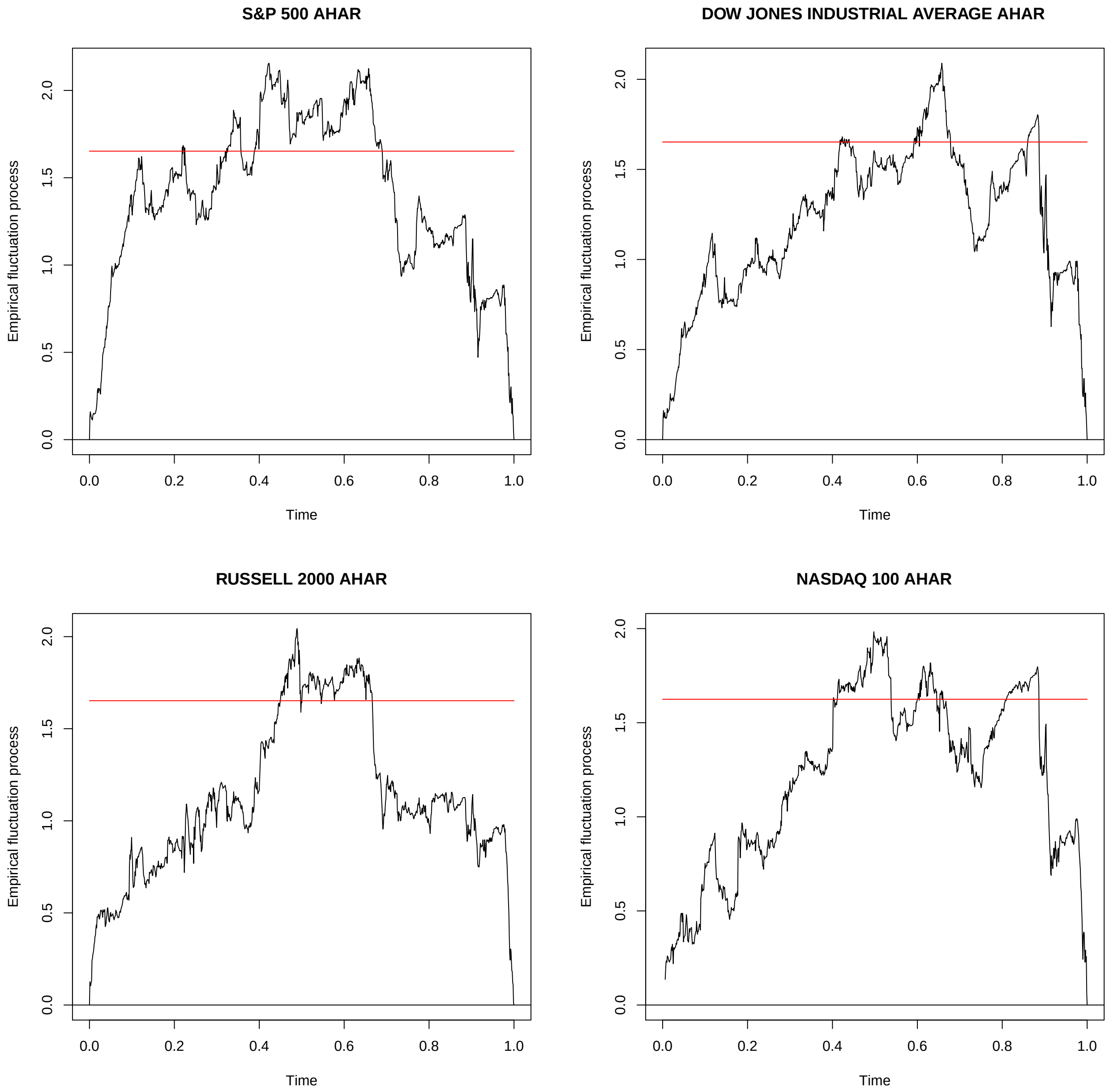


Figure 4: Empirical fluctuation process of the AHAR-RV model parameters for S&P500, Dow Jones Industrial Average, Russell 2000 and Nasdaq 100. The red line refers to the boundary with $\alpha = 0.05$.

S&P 500

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.5166	1.0000	6	2.154 (0.00)	0.4082	1.0000	6	2.035 (0.00)*
MSFE weights	0.5016	0.9710	5	1.010 (1.00)	0.3879	0.9500	4	0.610 (1.00)
ROC weights	0.4986	0.9653	2	-0.956 (1.00)	0.3825	0.9370	2	-0.753 (1.00)
ROC location weigh.	0.4980	**0.9639**	1	-1.410 (1.00)	0.3794	**0.9294**	1	-1.534 (1.00)
Equal Weights	0.5015	0.9708	4	0.917 (1.00)	0.3901	0.9557	5	1.191 (0.99)
Location weights	0.5007	0.9694	3	0.437 (1.00)	0.3874	0.9489	3	0.492 (1.00)

DJIA

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.6182	1.0000	6	2.112 (0.00)	0.5610	1.0000	6	2.007 (0.00)
MSFE weights	0.6088	0.9849	5	1.344 (0.97)	0.5410	0.9643	4	0.441 (1.00)
ROC weights	0.6066	0.9813	3	-0.213 (1.00)	0.5388	0.9603	2	-0.102 (1.00)
ROC location weigh.	0.6046	**0.9781**	1	-1.599 (1.00)	0.5319	**0.9480**	1	-1.789 (1.00)
Equal Weights	0.6079	0.9834	4	0.690 (1.00)	0.5441	0.9699	5	1.208 (1.00)
Location weights	0.6066	0.9813	2	-0.220 (1.00)	0.5402	0.9629	3	0.252 (1.00)

RUSSELL 2000

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.3491	1.0000	6	1.904 (0.00)	0.2066	1.0000	6	1.649 (0.00)
MSFE weights	0.3450	0.9881	5	1.235 (0,98)	0.2039	0.9868	4	1.018 (1.00)
ROC weights	0.3445	0.9868	4	0.925 (1.00)	0.2040	0.9872	5	1.071 (0.65)
ROC location weigh.	0.3424	0.9808	2	-0.548 (1.00)	0.2011	0.9733	2	-0.702 (1.00)
Equal Weights	0.3430	0.9826	3	-0.106 (1.00)	0.2025	0.9798	3	0.123 (1.00)
Location weights	0.3410	**0.9768**	1	-1.500 (1.00)	0.1998	**0.9670**	1	-1.507 (1.00)

NASDAQ 100

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.3509	1.0000	6	2.061 (0.00)	0.2397	1.0000	6	1.897 (0.00)
MSFE weights	0.3448	0.9825	4	1.007 (1.00)	0.2307	0.9697	5	1.106 (1.00)
ROC weights	0.3449	0.9829	5	1.130 (1.00)	0.2319	0.9676	4	0.907 (1.00)
ROC location weigh.	0.3433	0.9784	3	-0.229 (1.00)	0.2285	0.9535	2	-0.484 (1.00)
Equal Weights	0.3432	0.9781	2	-0.319 (1.00)	0.2300	0.9595	3	0.108 (1.00)
Location weights	0.3417	**0.9738**	1	-1.578(1.00)	0.2257	**0.9417**	1	-1.631 (1.00)

Table 2: **Out of sample forecasting result for the HAR-RV model**. For both the loss functions (MSE and QLIKE) the entries are: the values of the average loss; the ratio of the average loss to that of the expanding window method; the rank (rk) according to the average loss function; the T_{MAX} statistic and the p-value of the Model Confidence Set (MCS) procedure with $\alpha = 0.10$. A bold entry denotes the value of the smallest average loss

S&P 500

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.4247	1.0000	5	1.776 (0.00)	0.2858	1.0000	6	2.143 (0.00)
MSFE weights	0.4176	**0.9834**	1	-1.258 (1.00)	0.2666	0.9328	3	1.316 (0.01)
ROC weights	0.4183	0.9851	2	-0.617 (1.00)	0.2657	**0.9298**	1	-0.632 (1.00)
ROC location weigh.	0.4204	0.9899	4	1.275 (0.13)	0.2658	0.9301	2	0.618 (0.55)
Equal Weights	0.4197	0.9882	3	0.599 (1.00)	0.2678	0.9370	4	1.557 (0.00)
Location weights	0.4260	1.0031	6	1.536 (0.00)	0.2695	0.9430	5	1.673 (0.00)

DJIA

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.5291	1.0000	5	1.191 (0.09)	0.4236	1.0000	6	1.998 (0.00)
MSFE weights	0.5255	**0.9932**	1	-1.181 (1.00)	0.4012	0.9471	3	1.368 (0.00)
ROC weights	0.5260	0.9942	2	-0.844 (1.00)	0.3996	**0.9433**	1	-0.913 (1.00)
ROC location weigh.	0.5290	0.9998	4	1.092 (0.57)	0.3999	0.9440	2	0.912 (0.44)
Equal Weights	0.5268	0.9957	3	-0.311 (1.00)	0.4063	0.9592	4	1.633 (0.00)
Location weights	0.5325	1.0063	6	1.783 (0.00)	0.4091	0.9658	5	1.522 (0.00)

RUSSELL 2000

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.3096	1.0000	6	2.082 (0.00)	0.1828	1.0000	6	1.564 (0.00)
MSFE weights	0.3058	0.9877	5	1.783 (0.00)	0.1821	0.9964	5	1.464 (0.00)
ROC weights	0.3044	0.9833	4	0.777 (1.00)	0.1809	0.9895	4	1.334 (0.00)
ROC location weigh.	0.3035	**0.9805**	1	-1.676 (1.00)	0.1794	0.9812	2	-0.989 (1.00)
Equal Weights	0.3044	0.9832	3	0.686 (1.00)	0.1804	0.9867	3	1,268 (0.00)
Location weights	0.3042	0.9827	2	0.210 (1.00)	0.1789	**0.9786**	1	-0.993 (1.00)

NASDAQ 100

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.3180	1.0000	6	2.083 (0.00)	0.2070	1.0000	6	1.984 (0.00)
MSFE weights	0.3112	0.9787	4	1.058 (0.99)	0.1979	0.9562	5	1.066 (1.00)
ROC weights	0.3113	0.9788	5	1.086 (0.99)	0.1977	0.9551	4	0.970 (1.00)
ROC location weigh.	0.3099	0.9747	3	-0.034 (1.00)	0.1947	0.9406	2	-0.315 (1.00)
Equal Weights	0.3093	0.9726	2	-0.586 (1.00)	0.1953	0.9436	3	-0.047 (1.00)
Location weights	0.3082	**0.9691**	1	-1.516 (1.00)	0.1915	**0.9252**	1	-1.665 (1.00)

Table 3: **Out of sample forecasting result for the LHAR-RV model**. For both the loss functions (MSE and QLIKE) the entries are: the values of the average loss; the ratio of the average loss to that of the expanding window method; the rank (rk) according to the average loss function; the T_{MAX} statistic and the p-value of the Model Confidence Set (MCS) procedure with $\alpha = 0.10$. A bold entry denotes the value of the smallest average loss

S&P 500

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.4576	1.0000	6	2.186 (0.00)	0.3315	1.0000	6	2.151 (0.00)
MSFE weights	0.4397	0.9608	5	1.230 (0.94)	0.3072	0.9267	5	1.268 (0.00)
ROC weights	0.4371	0.9552	2	-0.860 (1.00)	0.3043	0.9179	3	1.393 (0.00)
ROC location weigh.	0.4364	**0.9537**	1	-1.416 (1.00)	0.3012	**0.9085**	1	-0.973 (1.00)
Equal Weights	0.4391	0.9596	4	0.781 (1.00)	0.3063	0.9240	4	1.413 (0.00)
Location weights	0.4385	0.9582	3	0.252 (1.00)	0.3016	0.9097	2	-0.977 (0.38)

DJIA

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.5669	1.0000	6	2.189 (0.00)	0.4718	1.0000	6	2.091 (0.00)
MSFE weights	0.5534	0.9762	5	1.015 (0.99)	0.4418	0.9365	5	0.986 (1.00)
ROC weights	0.5529	0.9754	3	0.291 (1.00)	0.4404	0.9335	3	0.688 (1.00)
ROC location weigh.	0.5516	**0.9731**	1	-1.744 (1.00)	0.4308	**0.9132**	1	-1.337 (1.00)
Equal Weights	0.5532	0.9758	4	0.679 (0.99)	0.4407	0.9341	4	0.749 (1.00)
Location weights	0.5526	0.9748	2	-0.219 (1.00)	0.4320	0.9158	2	-1.083 (1.00)

RUSSELL 2000

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.3193	1.0000	6	1.964 (0.00)	0.1898	1.0000	6	1.706 (0.00)
MSFE weights	0.3132	0.9811	5	1.498 (0.00)	0.1868	0.9842	5	1.348 (0.00)
ROC weights	0.3113	0.9750	4	1.183 (0.19)	0.1854	0.9769	4	1.385 (0.00)
ROC location weigh.	0.3086	0.9665	2	-0.735 (1.00)	0.1824	0.9610	2	0.995 (0.20)
Equal Weights	0.3107	0.9731	4	0.760 (1.00)	0.1849	0.9741	3	1.385 (0.00)
Location weights	0.3079	**0.9644**	1	-1.204 (1.00)	0.1818	**0.9579**	1	-0.995 (1.00)

NASDAQ 100

	MSE	Ratio	rk	MCS	**QLIKE**	Ratio	rk	MCS
Expanding Window	0.3299	1.0000	6	2.078 (0.00)	0.2223	1.0000	6	1.992 (0.00)
MSFE weights	0.3207	0.9719	4	0.825 (1.00)	0.2122	0.9545	4	0.936 (1.00)
ROC weights	0.3213	0.9738	5	1.212 (1.00)	0.2125	0.9560	5	1.073 (1.00)
ROC location weigh.	0.3192	0.9675	3	-0.092 (1.00)	0.2088	0.9394	2	-0.429 (1.00)
Equal Weights	0.3189	0.9665	2	-0.298 (1.00)	0.2101	0.9451	3	0.085 (1.00)
Location weights	0.3167	**0.9599**	1	-1.643 (1.00)	0.2058	**0.9257**	1	-1.651 (1.00)

Table 4: **Out of sample forecasting result for the AHAR-RV model**. For both the loss functions (MSE and QLIKE) the entries are: the values of the average loss; the ratio of the average loss to that of the expanding window method; the rank (rk) according to the average loss function; the T_{MAX} statistic and the p-value of the Model Confidence Set (MCS) procedure with $\alpha = 0.10$. A bold entry denotes the value of the smallest average loss

Appendix. Block bootstrap procedure for Model Confidence Set

Algorithm 1 Block bootstrap procedure for model confidence set

Choose the block-length bootstrap parameter l as the max number of significant parameters obtained by fitting an AR(p) model on all the d_{ij} terms.

Let $g = [T/l]$ where $[x]$ denote the smaller integer greater or equal to x.

Generate B bootstrap resamples of $\{1, \ldots, T\}$

- Let $(\psi_1^b, \ldots, \psi_g^b)$ *i.i.d.* uniform on $\{1, \ldots, T\}$ for all $b = \{1, \ldots, B\}$
- Define $\tau_k^b = (\psi_k^b, \psi_k^b + 1, \ldots, \psi_k^b + l - 1)$ for all $k = \{1, \ldots, g\}$. For convention $T + i = T$ for $i \geq 1$ and, if T is not a multiple of l, only $T + l - gl$ observations from the last block are used.
- The b bootstrap rasample is $\tau^b = (\tau_1^b, \ldots, \tau_g^b)$.

Compute $L^*_{b,i,t} = L_{i,\tau^b,t}$ and $\bar{L}^*_{b,i} = T^{-1} \sum_{t=1}^{T} L^*_{b,i,t}$, the corresponding bootstrap of $L_{i,t}$ and $\bar{L}_i$ respectively, for each model i=(1, ...,m) and for each t=(1, ...,T)

Calculate the variables $\xi^*_{b,i} = \bar{L}^*_{b,i} - \bar{L}_i$

Initialize by setting $M = M_0$

repeat

  Let $\xi^*_{b,.} = m^{-1} \sum_{i=1}^{m} \xi^*_{b,i}$

  Determine $t^*_{b,i} = (\xi^*_{b,i} - \xi^*_{b,.}) / \sqrt{var^*(\bar{d}_{i,.})}$
  where $var^*(\bar{d}_{i,.}) = B^{-1} \sum_{b=1}^{B} (\xi^*_{b,i} - \xi^*_{b,.})^2$ is the bootstrap variance.

  Define $T^*_{b,max} = \max_i t^*_{b,i}$ for $b = \{1, \ldots, B\}$.

  Calculate the p-value of $H_{0,M}$ given by:
  $P_{H_{0,M}} = B^{-1} \sum_{b=1}^{B} I_{(T_{max} > T^*_{b,max})}$
  where I is the indicator function

  if $P_{H_{0,M}} < \alpha$ **then**

    $H_{0,M}$ is rejected and the model with $e_M = \text{argmax}_i\, t_i$ is eliminated from M

  end if

until $P_{H_{0,M}} \geq \alpha$ for the first time.

The model confidence set is the set $\widehat{M}^*_{1-\alpha}$