Statistical tests for equal predictive ability across multiple forecasting methods

Daniel Borup and Martin Thyrsgaard

CREATES Research Paper 2017-19

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Statistical tests for equal predictive ability across multiple forecasting methods[*]

Daniel Borup[†]   Martin Thyrsgaard[**]

Abstract

We develop a multivariate generalization of the Giacomini-White tests for equal conditional predictive ability. The tests are applicable to a mixture of nested and non-nested models, incorporate estimation uncertainty explicitly, and allow for misspecification of the forecasting model as well as non-stationarity of the data. We introduce two finite-sample corrections, leading to good size and power properties. We also provide a two-step Model Confidence Set-type decision rule for ranking the forecasting methods into sets of indistinguishable conditional predictive ability, particularly suitable in dynamic forecast selection. In the empirical application we consider forecasting of the conditional variance of the S&P500 Index.

Keywords: forecast comparison; multivariate tests of equal predictive ability; Giacomini-White test; Diebold-Mariano test; conditional forecast combination

JEL Classification: C12, C52, C53

This version: May 16, 2017

[*] We thank Bent Jesper Christensen, Kim Christensen, and Anders B. Kock for insightful guidance. Furthermore, we would like to thank Christian M. Dahl, Cisil Sarisoy and Johan S. Jakobsen for their comments and suggestions. The authors acknowledge support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

[†] Corresponding author. CREATES, Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, Building 2622, 8210 Aarhus V, Denmark. Email: dborup@econ.au.dk, Tel.: +45 87165134.

[**] CREATES, Department of Economics and Business Economics, Aarhus University. Email: thyrsgaard@econ.au.dk.

I. Introduction

In many empirical applications two or more competing forecasting methods for predicting the same object of interest are available, and we may ask the question, whether they result in equal losses, potentially conditional on being in certain states. For instance, one may be interested in comparing forecasts generated from model-based methods (e.g. GARCH or component-based methods) with those generated from reduced-form methods (e.g. HAR or ARFIMA) to determine whether such families of forecasting methods possess equal predictive ability. Additionally, many empirical applications are concerned with the comparison of a set of models' forecasts to that of a baseline model. In macroeconomic applications researchers typically compare a first-order autoregressive model with models that include various predictors (see e.g. Stock and Watson (1999, 2003)), or in financial applications where efficient markets imply that excess returns form a martingale difference, leading to a null hypothesis where all predictive models nest the baseline model of zero expected excess returns (see e.g. Goyal and Welch (2003); Welch and Goyal (2008), and Phillips and Jin (2014)).

In this paper, we facilitate such analyses by developing statistical tests for equal conditional and unconditional predictive ability among two or more forecasting methods. The paper extends the tests in Giacomini and White (2006) (henceforth referenced as GW) to a multivariate setting where one may be interested in comparing the conditional predictive ability of multiple forecasting methods, and at the same time extends the (unconditional) multivariate Diebold-Mariano test statistic (Diebold and Mariano, 1995) in Mariano and Preve (2012) by allowing for non-stationarity in data (arising from e.g. model misspecification), a mixture of nested and non-nested models, and by explicitly accounting for estimation uncertainty in model parameters used in generating the forecasts. Whereas unconditional tests allow one to answer the question of whether a set of forecasting methods performed equally well on average in the past, the conditional tests allow one to investigate whether the set of forecasting methods performed equally well conditional on some information set containing e.g. macroeconomic or financial indicators. The latter reveals potential differences in predictive ability otherwise hidden in the unconditional test - what seems to be zero on average, may not be so when conditioning on additional information.

By developing multivariate versions of the GW tests, we enable testing of equal (un)conditional predictive ability among many forecast methods without having to employ multiple testing adjustments, which would otherwise be appropriate if one were to test the similar hypothesis via multiple pairwise tests using the GW tests. Such adjustments can be quite suboptimal in terms of power (see e.g. Romano, Shaikh, and Wolf (2010)), and Hubrich and West (2010) document that one may draw wrong conclusions on the basis of pairwise comparisons of the forecasting models. Since the proposed tests are natural extensions of GW, they inherit the main properties of the GW tests regardless of whether we take a conditional or unconditional perspective. First, they are applicable to multistep point, interval, probability, or density forecast evaluation for a general loss function. Secondly, they enable comparison of both nested and non-nested models and, thirdly, they incorporate non-vanishing estimation uncertainty of the parameters used in formulating the forecasts. That is, the tests incorporate differences in model complexities and estimation procedures, without explicitly requiring this to be done through the loss function. By formulating the tests in this manner, not only the model, but an additional number of choices made by the forecaster such as estimation method and window are included in the evaluation, making them tests for comparing forecast methods and not only forecasting models. Finally, the tests allow for non-stationarity in the data, arising from e.g. misspecification of the forecasting model and/or structural breaks in the data-generating process.

Our paper contributes to the large and active literature on forecast evaluation in several ways. First, we provide the first multivariate test for equal conditional predictive ability. Secondly, we facilitate easy testing of equal predictive ability (both conditional and unconditional), since all tests proposed in this paper are Wald statistics, hence have chi-squared limiting distributions in contrast to non-standard, context-specific distributions often found in the literature on forecast comparison tests (see e.g. Clark and McCracken (2001); McCracken (2007); Clark and McCracken (2012), and Gonçalves, McCracken, and Perron (2017)) for which the asymptotically valid critical values have to be obtained through burdensome simulation-based methods.[1] Moreover, we show that the tests are generally invariant to any reordering of the forecasting methods under comparison, ensuring that conclusions drawn from a single test is unaltered by any permutation of the ordering of the forecasting methods such that no multiple testing adjustments are required. Thirdly and in contrast to Hubrich and West (2010); Mariano and Preve (2012); Clark and McCracken (2012), the proposed tests are applicable to a mixture of nested and non-nested models, hold for a general loss function and allow for non-stationarity in data. Finally, we allow for comparison of a wider class of forecasting methods including linear, non-linear, Bayesian, and non-parameteric methods as opposed to the methods proposed in e.g. Clark and McCracken (2012); Granziera, Hubrich, and Moon (2014) and Gonçalves et al. (2017), which only apply in the case of linear models.

To improve upon the finite sample properties of the tests, we propose two adjustments. First, we introduce a threshold Wald statistic that employs a threshold estimator of the covariance matrix. Secondly, we introduce a power-enhancement component along the lines of Fan, Liao, and Yao (2015), potentially improving upon power, but with negligible size distortion under the null hypothesis. We examine the statistical properties of the tests in an elaborate Monte Carlo study, which indicates that they are well-sized and have good power. Moreover, the finite-sample adjustments succeed in improving both size and power noticeably.

Since rejection of the null hypothesis of equal conditional predictive ability suggests that one or more of the forecasting methods possess superior predictive ability, we develop a Model Confidence Set (Hansen, Lunde, and Nason, 2011) inspired rule for ranking of the forecasting methods into "method confidence sets", each containing sets of forecasting methods with indistinguishable conditional predictive ability. Via this rule, we can utilize that rejection of the null hypothesis implies that we can predict relative performances of the forecasting methods, leading to a decision rule for dynamic forecast selection.

In our empirical application, we consider forecasting the conditional variance

[1] Note, however, the recent paper by Hansen and Timmermann (2015), in which they show asymptotic equivalence of some of these tests with one based on simple Wald statistics.

of the S&P 500 Index' returns. Using the proposed theory, we investigate what drives (in)differences in forecasting performance over the 2009-2013 period between a large set of forecasting methods, including (G)ARCH, Realized GARCH (Hansen, Huang, and Shek, 2012), and Heterogeneous Autoregressive (HAR) specifications (Corsi, 2009). Examining the best set of forecasting methods, we document a number of interesting findings. First, we find that HAR specifications are preferred over the traditional (G)ARCH specifications, corroborating empirical findings in Andersen, Bollerslev, Diebold, and Labys (2003) and theoretical findings in Andersen, Bollerslev, and Meddahi (2004) and Sizova (2011). The inclusion of a realized measure of volatility in the GARCH dynamics as in the Realized GARCH model of Hansen et al. (2012) improves substantially on the performance of the GARCH framework, and makes it comparable to the best HAR type models. Secondly, we identify structural breaks in the composition of the best method confidence set. One of these events lines up with the Flash Crash of May 6, 2010. Specifically, the HAR of Corsi (2009) is consistently included during normal states of the markets in the period leading up to the Flash Crash, but drops out completely after this day. Thirdly, even though the forecasting methods of Patton and Sheppard (2015) are statistically indistinguishable based on their average past performances (using the unconditional test), our analysis indicates that the predictive gain relative to simpler models like HAR and HAR-J stems from very different states. For instance, the HAR-RS-I, HAR-RS-II, and HAR-SJ-I derive their gain almost exclusively during what we term as "leverage" and "jump" states, whereas HAR-SJ-II somewhat surprisingly is mostly excluded in leverage states, but performs especially well in normal market states. Finally, we show that exploiting the ranking rule based on these state-dependencies of the forecasting methods' predictive ability in a novel conditional forecast combination procedure leads to significant gains in predictive ability relative to individual forecasting methods and competing forecast combination methods.

The remainder of the paper is organized as follows: Section II introduces multivariate statistical tests for equal conditional and unconditional predictive ability for one-step and multistep forecast horizons including their asymptotic properties. Section III provides finite-sample corrections for the statistical tests, whereas Section IV reports size and power properties of the proposed tests in two Monte

Carlo studies. In Section V, we introduce a Model Confidence Set-type decision rule suitable for dynamic forecast selection, and provide an empirical analysis of forecasting the conditional variance of the S&P 500 Index' returns in Section VI. Finally, Section VII concludes. All proofs are in the Appendix.

II. Multivariate tests for equal predictive ability

This section builds upon the work of Giacomini and White (2006), hence our usage of notation will be similar. We consider an observed vector $\boldsymbol{W}_t \equiv (Y_t, \boldsymbol{X}_t)'$ defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where Y_t is the object of interest and $\boldsymbol{X}_t$ is a vector of predictors.[2] The filtration $\mathscr{F}_t$ is defined as the σ-field generated by past and current values of $\boldsymbol{W}_t$, $\mathscr{F}_t = \sigma(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_t)$. We consider a setting where $k+1$, $k \geq 1$, methods are available for forecasting τ periods into the future. We denote the time t forecast of $Y_{t+\tau}$ by $\hat{f}^i_{t,\tau,m^i} = f^i(\boldsymbol{W}_t, \boldsymbol{W}_{t-1}, \ldots, \boldsymbol{W}_{t-m^i+1}; \hat{\boldsymbol{\theta}}^i_{t,m^i})$ for $i = 1, \ldots, k+1$, where f^i is a measurable function. Subscript m^i on $\hat{f}$ indicates that the forecast is generated using m^i observations prior to time t. Moreover, $\hat{\boldsymbol{\theta}}^i_{t,m^i}$ denotes the parameter estimates (parametric, semi-parametric, or non-parametric) used in constructing the forecast for the i'th forecasting method. Let $m = \max\{m^1, \ldots, m^{k+1}\}$. For ease of exposition and along the lines of Giacomini and White (2006), we require that $m < \infty$, thus ruling out an expanding window, but allowing for e.g. a rolling window estimator, where the window is allowed to change size over time as well. Consequently, let $m_t = \max\{m^1_t, \ldots, m^{k+1}_t\}$, such that the first forecasts are formulated at time m_1 and $m = \max\{m_1, m_2, \ldots\}$. The requirement of finiteness of m also allows for a fixed estimation sample scheme, where the model parameters are estimated once using the first m_1 observations and then used to generate all T forecasts. In any case, the number of out-of-sample forecasts is $T = N - (m_1 + \tau - 1)$ with a total sample size of N (time series) observations.

In order to assess the forecasting ability of each forecasting method, we introduce the real-valued loss function $L_{t+\tau}\left(Y_{t+\tau}, \hat{f}^i_{t,\tau,m^i}\right)$. Important examples of L include economic measures such as utility or profits, or statistical measures such as the

[2] To keep the notation simple, we will focus on the case where Y_t is a scalar. The theory presented below applies in the general case as well.

square or absolute value of the forecast errors, where forecast errors are given by $e^i_{t+\tau} = \hat{f}^i_{t,\tau,m^i} - Y_{t+\tau}$. For additional examples of loss functions, see e.g. Granger and Machina (2006) for economic measures and West (2006) or Patton (2011) for statistical measures. To ease on notation, we suppress in the following the arguments of $L_{t+\tau}\left(Y_{t+\tau}, \hat{f}^i_{t,\tau,m^i}\right)$ and write the i'th loss function as $L^i_{m,t+\tau}$.

A. The hypothesis of equal conditional predictive ability

For a given loss function, we are interested in determining whether a set of $k+1$ forecasting methods perform equally well conditional on some σ-field, $\mathscr{G}_t$. That is, we want to test the hypothesis that

$$H_0 : \mathbb{E}[L^i_{m,t+\tau}|\mathscr{G}_t] = \mathbb{E}[L^{i+1}_{m,t+\tau}|\mathscr{G}_t], \quad i = 1,\ldots,k, \tag{1}$$

or equivalently that

$$H_0 : \mathbb{E}[\Delta \boldsymbol{L}_{m,t+\tau}|\mathscr{G}_t] = \mathbf{0}, \tag{2}$$

where $\Delta \boldsymbol{L}_{m,t+\tau} = (\Delta L^1_{m,t+\tau},\ldots,\Delta L^k_{m,t+\tau})'$ and $\Delta L^j_{m,t+\tau} = L^j_{m,t+\tau} - L^{j+1}_{m,t+\tau}$ for $j = 1,\ldots,k$. The null hypothesis implies that one cannot predict, conditional on the information contained in $\mathscr{G}_t$, whether one or more forecasting methods will be more accurate for forecasting the object of interest τ periods into the future.

We make two remarks on the formulation of the null hypothesis. First, the null hypothesis is expressed in terms of a conditional expectation, where the choice of conditioning information is made by the researcher. If $\mathscr{G}_t$ is set to the trivial σ-field, $\mathscr{G}_t = \{\emptyset, \Omega\}$, the null hypothesis is comparable to the one considered in Mariano and Preve (2012). In this case, the hypothesis test provides information about the average predictive ability of the forecasting methods in the past - the idea of Diebold and Mariano (1995) and West (1996) among others. In contrast, conditioning information enables the researcher to investigate whether additional information can assist in predicting performance differences between the forecasting methods. A leading example of conditioning information is $\mathscr{G}_t = \mathscr{F}_t$, which enables the test to capture any persistence in forecasting ability arising from e.g. misspecification of the forecasting models. Moreover, it is plausible that some forecasting methods' predictive ability varies according to the state of the economic environment, such that conditioning on macroeconomic or financial indicators would be appropriate.

Secondly, the loss functions depend explicitly on the parameter estimates and not on their probability limits, leading to a test statistic that takes into account estimation uncertainty. Importantly, by allowing for asymptotically non-vanishing estimation uncertainty, the test can accommodate the inclusion of nested models in the set of forecasting methods - a feature that the (unconditional) multivariate test in Mariano and Preve (2012) cannot handle.[3]

A.1. One-step multivariate conditional predictive ability test

In this section, we present the test statistic and its asymptotic properties. The null hypothesis in (2) is equivalent to stating that

$$H_0 : \mathbb{E}[\tilde{h}_t \Delta \boldsymbol{L}_{m,t+\tau}] = \boldsymbol{0} \tag{3}$$

for all $\mathscr{G}_t$-measurable functions $\tilde{h}_t$. We restrict attention to a subset of these functions, which we gather in the q-dimensional vector $\boldsymbol{h}_t = (\tilde{h}_t^{(1)}, \ldots, \tilde{h}_t^{(q)})'$, referred to as the test function. For some choice of test function, we construct a multivariate test for equal conditional predictive ability by

$$H_{0,h} : \mathbb{E}[\boldsymbol{h}_t \otimes \Delta \boldsymbol{L}_{m,t+\tau}] = \boldsymbol{0}, \tag{4}$$

where subscript h indicates the dependence on the test function. The specification in (4) is a natural multivariate extension of the test in Giacomini and White (2006), whose test is a special case obtained when $k = 1$.

We now consider the leading case with one-step ahead forecasting, $\tau = 1$ and $\mathscr{F}_t \subseteq \mathscr{G}_t$. For that purpose, we let $\boldsymbol{d}_{m,t+\tau} = \boldsymbol{h}_t \otimes \Delta \boldsymbol{L}_{m,t+\tau}$ and impose three assumptions similar to those of Giacomini and White (2006).

[3] Technically, with $\mathscr{G}_t = \{\emptyset, \Omega\}$ and asymptotically vanishing estimation uncertainty the standard errors of differences in forecast performance between a set of nested models will equal zero, leading to non-standard limiting distributions of the test statistics.

Assumption 1. *$\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{W}_t\}$ are ϕ−mixing with $\phi(t) = O\left(t^{-r/(2r-1)-\iota}\right)$, $r \geq 1$, or α−mixing with $\alpha(t) = O\left(t^{-\frac{r}{r-1}-\iota}\right)$, $r > 1$, for some $\iota > 0$.*

Assumption 1 imposes relatively mild restrictions on the dependence structure and heterogeneity of data. We do not impose the common (covariance) stationarity assumption as used in for instance Diebold and Mariano (1995) and Mariano and Preve (2012). Specifically, data may exhibit arbitrary structural changes, which is a common feature found in many empirical studies within e.g. macroeconomic prediction (see e.g. Stock and Watson (2003) and Schrimpf and Wang (2010)), stock return prediction (see e.g. Fama and French (1997) and Paye and Timmermann (2006)), and exchange rate prediction (see e.g. Giacomini and Rossi (2010)) to name a few. We also document such a case in the empirical section below.

Assumption 2. *$\mathbb{E}[|\boldsymbol{d}_{m,t+1,i}|^{2(r+\delta)}] < \infty$ for some $\delta > 0$, $i = 1, \ldots, qk$, and for all t, where subscript i indicate the i'th element of $\boldsymbol{d}_{m,t+1}$.*

Assumption 3. *$\boldsymbol{\Sigma}_T \equiv T^{-1}\sum_{t=1}^{T}\mathbb{E}[\boldsymbol{d}_{m,t+1}\boldsymbol{d}'_{m,t+1}]$ is uniformly positive definite.*

Assumptions 2-3 are mainly technical assumptions ensuring (uniformly) bounded moments of data and positive definiteness of the asymptotic variance. Both of these assumptions are common in the forecast evaluation literature. We then consider the following Wald statistic

$$S_{m,h} = T\bar{\boldsymbol{d}}'_m\hat{\boldsymbol{\Sigma}}_T^{-1}\bar{\boldsymbol{d}}_m, \tag{5}$$

where $\bar{\boldsymbol{d}}_m \equiv T^{-1}\sum_{t=1}^{T}\boldsymbol{d}_{m,t+1}$, and $\hat{\boldsymbol{\Sigma}}_T \equiv T^{-1}\sum_{t=1}^{T}\boldsymbol{d}_{m,t+1}\boldsymbol{d}'_{m,t+1}$ is a $(qk \times qk)$ sample covariance matrix that consistently estimates the variance of $\boldsymbol{d}_{m,t+1}$. We note that for large values of q and/or k, the dimension of $\boldsymbol{\Sigma}_T$ and $\bar{\boldsymbol{d}}_m$ may become large, potentially leading to issues with statistical inferences in finite samples. We propose remedies in Section III, but for now we restrict our attention to the properties of $S_{m,h}$ in (5). The asymptotic properties of the test statistic is summarized in Theorem 1.

Theorem 1 (One-step multivariate conditional predictive ability test)**.** *Suppose Assumptions 1-3 hold. For forecast horizon $\tau = 1$, test function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$ and under H_0 in (2),*

$$S_{m,h} \xrightarrow{d} \chi^2(qk), \quad as\ T \to \infty. \tag{6}$$

Therefore, by Theorem 1 a multivariate test for equal conditional predictive ability can be conducted by rejecting the null hypothesis whenever $S_{m,h} > z_{1-\alpha,qk}$, where $z_{1-\alpha,qk}$ is the $(1-\alpha)$ quantile of the chi-squared distribution with qk degrees of freedom.

Since any reordering of the forecasting methods alters the dynamics of $\boldsymbol{d}_{m,t+1}$, it motivates the following result, which shows that for each permutation (reordering) of the forecasting methods, regardless of whether the null is true or not, we get the same value of the test statistic and the same limiting distribution under the null hypothesis.

Proposition 2 (Permutation invariance)**.** *Let $\boldsymbol{L}^*_{t+1}$ be an arbitrary permutation of the forecast losses, and define $\Delta\boldsymbol{L}^*_{m,t+1} = \boldsymbol{D}\boldsymbol{L}^*_{t+1}$, where*

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix} \tag{7}$$

is a $k \times (k+1)$ matrix. Let $\bar{\boldsymbol{d}}^*_m = T^{-1}\sum_{t=1}^{T} \boldsymbol{d}^*_{m,t+1}$ with $\boldsymbol{d}^*_{m,t+1} = \boldsymbol{h}_t \otimes \Delta\boldsymbol{L}^*_{m,t+1}$ and $\hat{\boldsymbol{\Sigma}}^*_T \equiv \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{d}^*_{m,t+1}\boldsymbol{d}^{*'}_{m,t+1}$. Then,

$$S^*_{m,h} \equiv T\bar{\boldsymbol{d}}^{*'}_m(\hat{\boldsymbol{\Sigma}}^*_T)^{-1}\bar{\boldsymbol{d}}^*_m = S_{m,h} \tag{8}$$

for all T.

Proposition 2 ensures that conclusions drawn from the hypothesis testing are unaltered for any permutation of the ordering of the forecasting methods. This allows the researcher to perform just a single test.

A.2. Alternative hypothesis

When formulating an alternative hypothesis, one must take into account the fact that data may exhibit non-stationarity. For some $c > 0$, we formulate the alternative in line with Giacomini and White (2006) as

$$H_{A,h} : \mathbb{E}[\bar{\boldsymbol{d}}'_m]\mathbb{E}[\bar{\boldsymbol{d}}_m] \geq c, \tag{9}$$

for all T sufficiently large. Under stationarity the null and alternative hypothesis are exhaustive. Under non-stationarity this may not necessarily be the case. If an important $\mathscr{G}_t$-measurable variable is omitted in the test function, it may happen that $\mathbb{E}[\bar{\boldsymbol{d}}_m']\mathbb{E}[\bar{\boldsymbol{d}}_m] = 0$ for a particular sample size due to for instance shifting means without the null hypothesis being true - for example a situation where one method outperforms (some of) the other methods in certain periods/states, while it performs worse than the same methods in other periods/states. We consider this situation in the simulation study of Section IV and document its relevance in the empirical section below. Therefore, the test has little power against alternatives where the loss differentials are correlated with $\mathscr{G}_t$-measurable random variables not included in the test function. While this concern is important, it also highlights the flexibility of the test statistic. As mentioned above, the choice of test function is made by the researcher to include relevant variables supposed to assist in disentangling the forecasting abilities of the set of forecast methods. As a result, the test changes depending on the choice of test function. The result in Theorem 3 summarizes the power properties of the test statistic under the alternative hypothesis in (9).

Theorem 3. *Suppose Assumptions 1-3 hold. For any $c \in \mathbb{R}_+$ and under $H_{A,h}$ in (9),*

$$\mathbb{P}[S_{m,h} > c] \to 1, \quad \textit{as } T \to \infty. \tag{10}$$

In particular, regardless of the critical value chosen for the test, the probability of rejecting the null hypothesis when the alternative hypothesis is true tends to unity for $T \to \infty$.

A.3. Multistep multivariate conditional predictive ability test

For a multistep forecast horizon, $\tau > 1$, and $\mathscr{F}_t \subseteq \mathscr{G}_t$ we note that the sequence $\{\boldsymbol{h}_t \otimes \Delta\boldsymbol{L}_{m,t+\tau}\}$ may be serially autocorrelated up to the order of $\tau - 1$, since the null hypothesis in (4) implies that $\mathrm{Cov}[\boldsymbol{h}_t \otimes \Delta\boldsymbol{L}_{m,t+\tau}, \boldsymbol{h}_{t-j} \otimes \Delta\boldsymbol{L}_{m,t+\tau-j}] = 0$ for all $j \geq \tau$. That is, $\{\boldsymbol{h}_t \otimes \Delta\boldsymbol{L}_{m,t+\tau}\}$ may be serially correlated in the forecasting window. Consequently, we can no longer rely on the sample variance under the null for estimating the covariance matrix as was the case in the one-step formulation. Instead, we consider a HAC-type estimator (see e.g. Newey and West (1987) and

Andrews (1991)) with a bandwidth choice guided by the implications of the null hypothesis. The estimator is given by

$$\tilde{\boldsymbol{\Sigma}}_T = \frac{1}{T}\left[\sum_{t=1}^{T} \boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau} + \sum_{j=1}^{\tau-1}\kappa(j,\tau)\sum_{t=1+j}^{T}\left(\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-j} + \boldsymbol{d}_{m,t+\tau-j}\boldsymbol{d}'_{m,t+\tau}\right)\right], \tag{11}$$

where $\kappa(\cdot,\cdot)$ is a real-valued kernel weight function such that $\kappa(j,\tau)\to 1$ as $T\to\infty$ for each $j = 1,\ldots,\tau-1$ (see Andrews (1991)), and where we put weight only on the relevant $\tau-1$ lags of the sequence. The estimator in (11) is known as the truncated HAC estimator. In the parsimonious choice of equal weighting, one obtains the HAC estimator in Hansen (1982) with a rectangular kernel. For a discussion and investigation of the choice of kernel, we refer the reader to West (2008) and Clark and McCracken (2013).

For the conditional multistep hypothesis testing, we impose three assumptions similar to Assumptions 1-3.

Assumption 1*. $\{\boldsymbol{h}_t\}$ *and* $\{\boldsymbol{W}_t\}$ *are* ϕ*−mixing with* $\phi(t) = O\left(t^{-r/(2r-2)-\iota}\right)$, $r \geq 2$, *or* α*−mixing with* $\alpha(t) = O\left(t^{-\frac{r}{r-2}-\iota}\right)$, $r > 2$, *for some* $\iota > 0$.

Assumption 2*. $\mathbb{E}[|\boldsymbol{d}_{m,t+\tau,i}|^{r+\delta}] < \infty$ *for some* $\delta > 0$, $i = 1,\ldots,qk$, *and for all* t, *where subscript* i *indicate the* i*'th element of* $\boldsymbol{d}_{m,t+1}$.

Assumption 3*. $\boldsymbol{\Sigma}_T \equiv T^{-1}\sum_{t=1}^{T}\mathbb{E}[\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau}] + T^{-1}\sum_{j=1}^{\tau-1}\sum_{t=1+j}^{T}\Big(\mathbb{E}[\boldsymbol{d}_{m,t+\tau} \times \boldsymbol{d}'_{m,t+\tau-j}] + \mathbb{E}[\boldsymbol{d}_{m,t+\tau-j}\boldsymbol{d}'_{m,t+\tau}]\Big)$ *is uniformly positive definite.*

Along the lines of the former section, we construct a Wald statistic for multistep multivariate conditional equal predictive ability. The test statistic is given by

$$S_{m,h,\tau} = T\bar{\boldsymbol{d}}'_m\tilde{\boldsymbol{\Sigma}}_T^{-1}\bar{\boldsymbol{d}}_m, \tag{12}$$

where $\bar{\boldsymbol{d}}_m = T^{-1}\sum_{t=1}^{T}\boldsymbol{d}_{m,t+\tau}$. Analogue to Theorems 1-3 and Proposition 2, $S_{m,h,\tau}$ is asymptotically chi-squared distributed with qk degrees of freedom under the null hypothesis, has power under the alternative hypothesis in (9), and is permutation invariant. We summarize these results in Theorem 4 below.

Theorem 4 (Multistep multivariate conditional predictive ability test)**.** *Suppose*

Assumptions 1-3* hold.*

(i) *For forecast horizon $\tau > 1$, test function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$ and under H_0 in (2),*

$$S_{m,h,\tau} \xrightarrow{d} \chi^2(qk), \quad as\ T \to \infty. \tag{13}$$

(ii) *For any $c \in \mathbb{R}_+$ and under $H_{A,h}$ in (9),*

$$\mathbb{P}[S_{m,h,\tau} > c] \to 1, \quad as\ T \to \infty. \tag{14}$$

(iii) *Let $\boldsymbol{L}^*_{t+\tau}$ be an arbitrary permutation of the forecast losses, and define $\Delta\boldsymbol{L}^*_{m,t+\tau} = \boldsymbol{D}\boldsymbol{L}^*_{t+\tau}$, $\bar{\boldsymbol{d}}^*_m = T^{-1}\sum_{t=1}^T \boldsymbol{d}^*_{m,t+\tau}$ with $\boldsymbol{d}^*_{m,t+\tau} = \boldsymbol{h}_t \otimes \Delta\boldsymbol{L}^*_{m,t+\tau}$ and $\tilde{\boldsymbol{\Sigma}}^*_T$ be the associated covariance estimator defined in equation (11). Then,*

$$S^*_{m,h,\tau} \equiv T\bar{\boldsymbol{d}}^{*'}_m(\tilde{\boldsymbol{\Sigma}}^*_T)^{-1}\bar{\boldsymbol{d}}^*_m = S_{m,h,\tau} \tag{15}$$

for all T.

Consequently, a multivariate test for equal conditional multistep ahead forecasting ability can be conducted by rejecting the null hypothesis whenever $S_{m,h,\tau} > z_{1-\alpha,qk}$.

A.4. Multivariate unconditional predictive ability test

In the unconditional test with $\mathcal{G}_t = \{\emptyset, \Omega\}$ (hence, $\boldsymbol{h}_t = 1$ for all t) and $\tau \geq 1$, the sequence $\{\Delta\boldsymbol{L}_{m,t+\tau}\}$ is not 'finitely correlated'. That is, the null does no longer restrict the serial correlation to only the forecasting window, but $\{\Delta\boldsymbol{L}_{m,t+\tau}\}$ may exhibit serial correlation of any order - including infinite. Hence, we impose a modified version of Assumption 3.

Assumption 3.** $\boldsymbol{\Sigma}_T \equiv T^{-1}\sum_{t=1}^T \mathbb{E}[\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau}] + T^{-1}\sum_{j=1}^{T-1}\sum_{t=1+j}^T \Big(\mathbb{E}[\boldsymbol{d}_{m,t+\tau} \times \boldsymbol{d}'_{m,t+\tau-j}] + \mathbb{E}[\boldsymbol{d}_{m,t+\tau-j}\boldsymbol{d}'_{m,t+\tau}]\Big)$ *is uniformly positive definite.*

To accommodate this, we adopt a covariance estimator of the form

$$\breve{\boldsymbol{\Sigma}}_T = \frac{1}{T}\Big[\sum_{t=1}^{T}\Delta\boldsymbol{L}_{m,t+\tau}\Delta\boldsymbol{L}'_{m,t+\tau} + \sum_{j=1}^{b_T}\kappa(j,b_T)\sum_{t=1+j}^{T}\Big(\Delta\boldsymbol{L}_{m,t+\tau}\Delta\boldsymbol{L}'_{m,t+\tau-j}+\Delta\boldsymbol{L}_{m,t+\tau-j}\Delta\boldsymbol{L}'_{m,t+\tau}\Big)\Big], \tag{16}$$

where $\{b_T\}$ is an integer-valued truncation point sequence satisfying $b_T \to \infty$ as $T \to \infty$ and $b_T = o(T)$ (see e.g. Newey and West (1987)). Note that we require $b_T \to \infty$ for consistency in the the unconditional case as opposed to $b_T = \tau - 1$ in the conditional case described in the former section. For a review of data driven bandwidth selection methods see Clark and McCracken (2013). Along the lines of former sections, we construct the following Wald statistic which can be used in testing for multistep unconditional equal predictive ability

$$S^{\text{unc}}_{m,h,\tau} = T\overline{\Delta\boldsymbol{L}}'_m \breve{\boldsymbol{\Sigma}}_T^{-1}\overline{\Delta\boldsymbol{L}}_m, \tag{17}$$

where $\overline{\Delta\boldsymbol{L}}'_m = T^{-1}\sum_{t=1}^{T}\Delta\boldsymbol{L}_{m,t+\tau}$. This test statistic is related to the one of Mariano and Preve (2012). However, as mentioned above, our test generalizes theirs along several dimensions. In particular, we enable comparison of nested models, allow for non-stationary data, and take explicitly into account the estimation method involved in generating the forecast series for all models. Analogously to Theorem 4, $S^{\text{unc}}_{m,h,\tau}$ is asymptotically chi-squared distributed with qk degrees of freedom under the null, has power under the alternative hypothesis in (9), and is permutation invariant. Theorem 5 summarizes these results.

Theorem 5 (Unconditional predictive ability test)**.** *Suppose Assumptions 1*-2* and Assumption 3** hold.*

(i) *For forecast horizon $\tau \geq 1$, $\mathscr{G}_t = \{\emptyset, \Omega\}$, $m < \infty$ and under H_0 in (2),*

$$S^{unc}_{m,h,\tau} \xrightarrow{d} \chi^2(k), \quad \text{as } T \to \infty. \tag{18}$$

(ii) *For any $c \in \mathbb{R}_+$ and under $H_{A,h}$ in (9),*

$$\mathbb{P}[S^{unc}_{m,h,\tau} > c] \to 1, \quad \text{as } T \to \infty. \tag{19}$$

(iii) *Let $\boldsymbol{L}^*_{t+\tau}$ be an arbitrary permutation of the forecast losses, and define $\Delta\boldsymbol{L}^*_{m,t+\tau} = \boldsymbol{D}\boldsymbol{L}^*_{t+\tau}$, $\overline{\Delta\boldsymbol{L}}^*_m = T^{-1}\sum_{t=1}^{T}\Delta\boldsymbol{L}^*_{m,t+\tau}$, and $\breve{\boldsymbol{\Sigma}}^*_T$ be the associated covariance estimator via (16). Then,*

$$S^{unc*}_{m,h,\tau} \equiv T\overline{\Delta\boldsymbol{L}}^{*'}_m(\breve{\boldsymbol{\Sigma}}^*_T)^{-1}\overline{\Delta\boldsymbol{L}}^*_m = S^{unc}_{m,h,\tau} \tag{20}$$

for all T.

Consequently, a multivariate test for equal unconditional forecasting ability can be conducted by rejecting the null hypothesis whenever $S^{\mathrm{unc}}_{m,h,\tau} > z_{1-\alpha,k}$. The permutation invariance result in Theorem 5iii) is similar to Proposition 2 in Mariano and Preve (2012), but holds under the milder Assumptions 1*-2* and Assumption 3**, and hence also applies in a setting with non-stationary data, inclusion of nested models and explicit account of estimation uncertainty.

III. Finite-sample corrections

The number of elements to be estimated in the covariance matrix is $qk(qk+1)/2$. Consequently, the dimension of the covariance matrix may become large if the objective is to test equal (un)conditional predictive ability of many methods, say, in the lower two-digits, and/or if many elements are included in the test function. Estimating a high-dimensional covariance matrix using the sample covariance matrix, when the sample size is small relative to the number of elements to be estimated, may negatively affect the size and power of the proposed tests. In this section, we provide remedies that correct the original test statistic to accommodate studies, where qk is large relative to the sample size. To fix ideas, we consider the conditional case with $\tau = 1$, but results are directly generalizable to a multistep forecast horizon as well as the unconditional case.

A. A threshold Wald statistic

To improve upon the finite-sample properties of the test statistic in (5), we utilize that we can consistently estimate $\boldsymbol{\Sigma}_T$ via the thresholding approach of Bickel and Levina (2008). Essentially, the thresholding estimator shrinks small off-diagonal elements towards zero, thus reducing the impact of the noise introduced by estimating elements that are (close to) zero. In particular, let $p_{ij}(\cdot)$

be a generalized thresholding function (Rothman, Levina, and Zhu, 2009) with threshold value $\lambda_{ij} = C(\sigma_{ii}\sigma_{jj}\log(qk)/T)^{1/2}$, for some constant $C > 0$, and where $\sigma_{ij} = T^{-1}\sum_{t=1}^{T} \boldsymbol{d}_{m,t+1,i}\boldsymbol{d}_{m,t+1,j}$ for $i, j = 1, \ldots, qk$. By choosing C sufficient large one can ensure that the estimated covariance matrix will be positive definite (see e.g. Fan, Liao, and Mincheva (2013)). The threshold covariance estimator $\hat{\boldsymbol{\Sigma}}^{\text{thr}}$ is then defined by

$$\hat{\boldsymbol{\Sigma}}_{ij}^{\text{thr}} = \begin{cases} \sigma_{ii}, & \text{if } i = j, \\ p_{ij}(\sigma_{ij}), & \text{if } i \neq j. \end{cases} \tag{21}$$

The thresholding function must satisfy for all $x \in \mathbb{R}$ the following three conditions

(i) $p_{ij}(x) = 0$ for $|x| \leq \lambda_{ij}$ (thresholding),

(ii) $|p_{ij}(x)| \leq |x|$ (shrinkage), and

(iii) $|p_{ij}(x) - x| \leq \lambda_{ij}$ (limited shrinkage).

Examples of such functions are soft thresholding, $p_{ij}(x) = \text{sgn}(x)\max\{0, |x| - \lambda_{ij}\}$, hard thresholding, $p_{ij}(x) = x\mathbb{1}\{|x| \geq \lambda_{ij}\}$, (Donoho and Johnstone, 1994), the adaptive Lasso, and the smoothly clipped absolute deviation (SCAD), which is a compromise between soft and hard thresholding (Fan and Li, 2001) defined by

$$p_{ij}(x) = \begin{cases} \text{sgn}(x)\max\{0, |x| - \lambda_{ij}\}, & \text{if } |x| \leq 2\lambda_{ij}, \\ ((b-1)x - \text{sgn}(x)b\lambda_{ij})/(b-2), & \text{if } 2\lambda_{ij} < |x| \leq b\lambda_{ij}, \\ x, & \text{if } |x| > b\lambda_{ij}, \end{cases} \tag{22}$$

for some $b > 2$. See Rothman et al. (2009) for a review of the thresholding functions' finite-sample properties. The threshold value depends on the choice of C, which is to be made by the researcher. One way to do so is to follow the recommendations put forward in Rothman et al. (2009). Alternatively, the parameter can be chosen in a data-driven manner via cross-validation as in Fan et al. (2013).

Since the number of forecasting methods and the dimension of the test function are fixed, we obtain that the asymptotic properties of the test statistic with the sample covariance matrix replaced by the threshold estimator are identical to those of $S_{m,h}$ under the null and alternative hypothesis. We henceforth refer to this as the threshold Wald (TW) statistic and summarize its asymptotic properties in the following result.

Proposition 6 (Threshold Wald statistic)**.** *Suppose Assumptions 1-3 hold.*

(i) *For forecast horizon $\tau = 1$, test function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$ and under H_0 in (2),*

$$S^{(1)}_{m,h} \equiv T\bar{\boldsymbol{d}}'_m(\hat{\boldsymbol{\Sigma}}^{thr}_T)^{-1}\bar{\boldsymbol{d}}_m \xrightarrow{d} \chi^2(qk), \quad as\ T \to \infty, \tag{23}$$

where $\hat{\boldsymbol{\Sigma}}^{thr}_T$ is a threshold estimator of the type in (21).

(ii) *For any $c \in \mathbb{R}_+$ and under $H_{A,h}$ in (9),*

$$\mathbb{P}[S^{(1)}_{m,h} > c] \to 1, \quad as\ T \to \infty. \tag{24}$$

Consequently, a multivariate test for equal conditional predictive ability across many methods can be conducted by simply replacing the empirical sample covariance with the threshold estimator, and by rejecting the null hypothesis whenever $S^{(1)}_{m,h} > z_{1-\alpha,qk}$. The following result shows that permutation of the forecasting methods will not alter the test statistic nor limiting distribution asymptotically.

Corollary 7 (Asymptotic permutation invariance)**.** *Let $\bar{\boldsymbol{d}}^*_m$ be given as in Proposition 2, $\hat{\boldsymbol{\Sigma}}^{thr*}$ be the associated threshold covariance matrix estimator, and*

$$S^{(1)*}_{m,h} \equiv T\bar{\boldsymbol{d}}^{*\prime}_m(\hat{\boldsymbol{\Sigma}}^{thr*})^{-1}\bar{\boldsymbol{d}}^*_m. \tag{25}$$

Then, $S^{(1)*}_{m,h} - S^{(1)}_{m,h} \xrightarrow{\mathbb{P}} 0$, as $T \to \infty$.

We thus conclude that the TW statistic can be used in the same manner as the standard test statistic in (5), ensuring that a single test will suffice for testing multivariate equal (un)conditional predictive ability across a set of forecasting methods. However, we stress that the finite-sample appropriateness of Corollary 7 depends on the finite-sample behavior of the chosen covariance estimator. The thresholding estimator proposed above is just one of many possible choices, and that other choices might be preferable under certain structural assumptions on the covariance matrix.

B. *Power enhancement of the threshold Wald statistic*

Tests based on the (threshold) Wald statistic may suffer from low power when the number of methods and/or elements of the test function are large relative to the sample size. This is especially true under sparse alternatives, where the number of elements that violates the null hypothesis is small relative to the dimension of $\bar{\boldsymbol{d}}_m$. To alleviate this potential issue, we introduce a power enhancement component, $S^{(0)}_{m,h}$, along the lines of Fan et al. (2015). This component boosts power in specific regions of the alternative hypothesis (e.g. in sparse alternatives), where power may be low. Consequently, we construct a power-enhanced test statistic as

$$S^{(2)}_{m,h} = S^{(1)}_{m,h} + S^{(0)}_{m,h}. \tag{26}$$

We assume that the power enhancement component satisfies the following properties

Assumption 4 (Power enhancement properties)**.**

(i) $S^{(0)}_{m,h} \geq 0$ *almost surely,*

(ii) $\mathbb{P}[S^{(0)}_{m,h} = 0|H_0] \to 1$*, and*

(iii) $S^{(0)}_{m,h}$ *diverges in probability for specific regions of the alternative hypothesis.*

Assumption 4i requires non-negativity of the power enhancement component, thus insuring that power never is adversely affected by the introduction of this component. 4iii ensures that power is enhanced in certain regions of the alternative hypothesis. Assumption 4ii ensures that size is not affected (asymptotically) by inclusion of the power enhancement component. Note that $S^{(0)}_{m,h}$ is not a test statistic on its own due to Assumption 4ii, which ensures that the asymptotic distribution of $S^{(2)}_{m,h}$ under the null hypothesis is determined by that of $S^{(1)}_{m,h}$ - it simply provides additional power (with little size distortion) by adding a non-negative component to the original test statistic in specific regions of the alternative hypothesis. We set the power enhancement component to a screening statistic (see e.g. Fan et al. (2015)), which satisfies Assumption 4,

$$S^{(0)}_{m,h} = \sqrt{qk} \sum_{i=1}^{qk} \frac{\bar{\boldsymbol{d}}^2_{m,i}}{\sigma_{ii}/T} \mathbb{1}\{|\bar{\boldsymbol{d}}_{m,i}| > \sqrt{\sigma_{ii}/T}\Lambda_{qk,T}\}, \tag{27}$$

where $\bar{\boldsymbol{d}}_{m,i}$ denotes the i'th element, $i = 1,\dots,qk$, of $\bar{\boldsymbol{d}}_m$, and $\Lambda_{qk,T}$ is a threshold that plays an important role in determining the size of the screening set, which is set to $\Lambda_{qk,T} = \log\{\log(T)\}\sqrt{\log(qk)}$. Consequently, the power enhancement component strengthens the signal of $\bar{\boldsymbol{d}}_m$ by enhancing the (sufficiently) large non-zero elements. By Assumption 4ii and Proposition 6 it follows that $S^{(0)}_{m,h}$ inherits the asymptotic properties of the TW statistic in (23). The results are summarized in Proposition 8.

Proposition 8 (Power enhanced threshold Wald statistic)**.** *Suppose Assumptions 1-3 hold.*

(i) *For forecast horizon $\tau = 1$, test function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$ and under H_0 in (2),*

$$S^{(2)}_{m,h} \xrightarrow{d} \chi^2(qk), \quad \text{as } T \to \infty. \tag{28}$$

(ii) *For any $c \in \mathbb{R}_+$ and under $H_{A,h}$ in (9),*

$$\mathbb{P}[S^{(2)}_{m,h} > c] \to 1, \quad \text{as } T \to \infty. \tag{29}$$

Consequently, a multivariate test for equal conditional predictive ability with potentially improved finite-sample properties can be conducted in the usual way by rejecting the null hypothesis whenever $S^{(2)}_{m,h} > z_{1-\alpha,qk}$.

IV. Simulation study

To examine the finite sample properties of the test statistics, we perform a Monte Carlo study. The study covers both the conditional and unconditional case. We also study the impact of the finite-sample corrections put forward in Section III. In general, we document two important findings. First, the proposed test statistics have good size and power properties. Secondly, one is allowed to use a relatively large number of conditioning variables, for instance macroeconomic and financial indicators, while maintaining good finite-sample properties of the tests.

A. Simulation design

We consider the case where the data-generating process of the vector loss differential series $\{\Delta \boldsymbol{L}_{t+1}\}$ is given by

$$\Delta \boldsymbol{L}_{t+1} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{t+1}, \tag{30}$$

with $\boldsymbol{\varepsilon}_{t+1}$ being random vectors drawn from a multivariate, k-dimensional, normal distribution, $\boldsymbol{\varepsilon}_{t+1} \sim N_k(\mathbf{0}, \boldsymbol{\Gamma}_k)$. Here, $\boldsymbol{\Gamma}_k$ denotes the $k \times k$ contemporaneous covariance matrix with equi-off-diagonal entry generated from $U(0, 1/2)$ along the lines of Fan et al. (2015). Moreover, we introduce a regime-shift in the data-generating process by setting the diagonal in $\boldsymbol{\Gamma}_k$ equal to 1.25 in the first half of the sample and equal to 0.75 in the second half of the sample, thereby introducing a structural break in the loss series, which divides the data into a high-variance and low-variance regime typically observed in empirical studies (see e.g. So, Lam, and Li (1998)). Consequently, the loss differential series has on average approximately unit variances comparable to the simulation study in Mariano and Preve (2012), and we allow for contemporaneous correlation in the loss differential series. In the size and power experiments we set $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\mu} \neq \mathbf{0}$, respectively. When implementing the threshold estimator of $\boldsymbol{\Sigma}_T$ we employ the soft thresholding function with a value of $C = 2/3$ consistent with the recommendations in Fan et al. (2013) to minimize the number of non-positive definite covariance matrices. To facilitate comparison, we set the truncation lag of the HAC estimator to zero as in Giacomini and White (2006). In all experiments, we examine three sample sizes, $T = \{250, 500, 1000\}$. Given a reasonable initialization period (estimation window), the sample size $T = 250$ is, for instance, comparable to a case with a long time series of quarterly macroeconomic data, whereas $T = 500$ and $T = 1000$ are comparable to, for instance, a case with a shorter time series of monthly, weekly or daily data of stock returns. We do 10,000 Monte Carlo replications and set the nominal size to 10%.

B. Size properties

We first examine the size properties of the original test statistic and the TW statistic. For the former case, we let $k \in \{1, 2, 3, 4\}$. This is comparable to the setting considered in Mariano and Preve (2012). In the latter case, we extend the maximum number of methods to 10. For the conditional tests, we set $\boldsymbol{h}_t = (1, \Delta \boldsymbol{L}_t)'$ corresponding to the case of testing whether past predictive performance can inform about future performance. In the many-methods case, we report results using the TW statistic without inclusion of the power enhancement component.[4] This simulation design thus resembles situations often encountered in empirical exercises, and simultaneously constitutes a challenging setup by including contemporaneous correlation, regime-switching variance and many instruments. Table 1 reports the results for the uncorrected test statistic.

≪ Insert Table 1 about here ≫

We observe that the unconditional and conditional tests are generally well-sized, though the conditional test become moderately oversized, when, not surprisingly, the number of dimension (qk) increases. For large qk, the modified test statistic thus become relevant. Table 2 reports results obtained using the TW statistic for the conditional test.

≪ Insert Table 2 about here ≫

We observe that the proposed thresholding approach improves noticeable upon the size distortion that occurs when the total number of methods and dimension of the test function increase. For the conditional test statistic, empirical sizes are good for all sample sizes and for all number of methods, occasionally showing only a slight undersizing. Hence, by employing the TW statistic the test can be applied even with a rather large number of forecasting methods (and/or instruments) under examination while maintaining good size.

C. Power properties

Next, we turn to studying the power properties of the test statistic with or without the finite-sample corrections with the same range of methods, respectively. We let

$$\mu_j = \begin{cases} 0.25, & \text{if } j = 1, \\ 0, & \text{otherwise,} \end{cases} \tag{31}$$

[4] When including the power enhancement component, results in Table 3 reveal a moderate size distortion mainly for $k = 2$, which decreases in sample size and dimension (qk). Since the power-enhancement component will typically only be included in cases where k is not small, we consider this potential issue a minor concern.

which resembles a situation where the alternative hypothesis is true due to lower predictive ability of the first method relative to the remaining methods. In particular, the first method is 25% worse than the remaining methods in line with the simulation study in Mariano and Preve (2012). Table 4 reports the results for the uncorrected test statistic.

$\ll$ Insert Table 4 about here $\gg$

The unconditional test statistic has good power properties for all sample sizes and number of forecasting methods. The conditional test have good power properties for $T = 500, 1000$, and reasonable power in the low sample size case with moderate values of qk. As expected, power decreases with the number of methods in the $T = 250$ case, motivating the use of the finite-sample corrections for larger qk. Using the same structure of $\boldsymbol{\mu}$ as in (31), we report results of the TW statistic without the power enhancement component in Table 5 and including the power enhancement component in Table 6.

$\ll$ Insert Table 5 and 6 about here $\gg$

The power enhancement provides a noticeable increase in power in the conditional case, leading to good power properties of the conditional test statistic. In general, power increases in sample size and decreases in the number of methods and elements of the test function.

C.1. Different predictive ability driven by state variables

To put the multivariate conditional test statistic into an economic perspective, we consider a situation where one method is more accurate in a given state of the economy, and less accurate in another state relative to the remaining methods in the method set, but unconditionally the methods are equally accurate. Following Giacomini and White (2006), we define a state variable V_t with $\mathbb{P}[V_t = 1] = \rho$ and $\mathbb{P}[V_t = 0] = 1 - \rho$. For $T = 500$ we generate 10,000 loss difference sequences according to

$$\Delta \boldsymbol{L}_{t+1} = \boldsymbol{\mu} \frac{V_t - \rho}{\rho(1-\rho)} + \boldsymbol{\varepsilon}_{t+1}, \tag{32}$$

where the first element of $\boldsymbol{\mu}$ is set equal to r and zero otherwise, and the error terms are generated according to the procedure explained in the former section, incorporating contemporaneous correlation and regime-switching variance. It is clear that $\mathbb{E}[\Delta \boldsymbol{L}_{t+1}] = \boldsymbol{0}$ and $\mathbb{E}[\Delta L^1_{t+1}|V_t = 1] = r/\rho$ in contrast to $\mathbb{E}[\Delta L^1_{t+1}|V_t = 0] = -r/(1-\rho)$. We consider a range of $r \in [0, 0.3]$ with $\rho = 0.5$. In the conditional test, we set $\boldsymbol{h}_t = (1, V_t)'$. We report results for the TW statistic, but for ease of exposition restrict ourselves to considering the case of $k = 4$ and $k = 9$ and plot the power curves in two figures. The curves are depicted in Figure 1.

≪ Insert Figure 1 about here ≫

It is clear that the conditional test quickly achieves power to detect different performance in different states and that the unconditional test is, as expected, close to the nominal size of 10%.

V. A rule for ranking and selection of forecasting methods

Rejection of the null hypothesis suggests that one or more of the forecasting methods possess better predictive ability, however, it provides no guidance towards which method(s) that causes the rejection. The identification of these method(s) might be of practical interest. In this section, we provide an algorithm that ranks forecasting methods into sets with equal conditional predictive ability. This procedure can be utilized dynamically to select forecast methods that is expected (conditional on $\mathscr{G}_t$) to yield the lowest loss at time $\overline{T} + \tau$ and conditional combination techniques within, for instance, the best set may be of practical relevance.[5]

In formulating the algorithm, we utilise a MCS-type procedure (Hansen et al., 2011) to eliminate methods according to some elimination rule and rank forecasting methods into $\mathscr{K} \leq k+1$ sets, henceforth referenced as "method confidence sets", whose elements have equal conditional predictive ability. Let M_0 be the set of the $k+1$ forecasting methods under consideration and M^* a preliminary set of best forecasting methods (in terms of some loss function). We propose the following three-step procedure:

- **Step 0:** Set $M = M_0$. Regress $\Delta L^j_{m,t+1}$ on $\boldsymbol{h}_t$ over some rolling window for $j = 1, \ldots, k$. The conditional expectation, $\mathbb{E}[\Delta L^j_{m,t+\tau}|\mathscr{G}_T]$, is approximated by

[5] See e.g. Aiolfi and Timmermann (2006), who exploit forecasting combination within "clusters" of models to improve forecasting ability.

the predicted value from the j'th regression. Based on $\hat{\beta}^{j'}\boldsymbol{h}_{\overline{T}}$ rank all $k+1$ methods, where $\hat{\beta}^{j}$ is the vector of regression coefficients. The forecasting method with lowest predicted loss is ranked first, and similarly the method with highest fitted value is ranked at last.

- **Step 1:** Run the multivariate test for equal conditional predictive ability.
- **Step 2:** If the test is not rejected, set $M^* = M$. Otherwise, eliminate the lowest ranked forecasting method from M and iterate Steps 1-2 until the null is not rejected.

Repeating the steps once leads to a set M_1 containing the best forecasting methods statistically indistinguishable in terms of conditional predictive ability. Repeating the procedure until no additional method confidence set can be found finalizes the algorithm. Consequently, the method confidence sets are ordered from those yielding least expected loss to those yielding the highest expected loss, $M_1, \ldots, M_{\mathcal{K}}$. Effectively, this is a multivariate extension of the decision rule proposed in Giacomini and White (2006). Besides leading to the same ranking across different forecasting methods, it is also clear that (asymptotic) permutation invariance is important for conducting the algorithm, because in each iteration with rejection of the test we eliminate a method, leading to a reordering of the methods. Due to the permutation invariance, this reordering has no impact on whether we reject or not in the following iteration.

VI. Forecasting conditional variance of stock returns

To illustrate the workings of the multivariate test and the ranking rule proposed in former sections, we focus our empirical investigation on forecasting (one day ahead) the daily open-to-close conditional variance of the S&P 500 Index' returns. To this end, we suppose that the efficient (log) price process is an Itô semimartingale of the form

$$p_t = p_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + J_t, \tag{33}$$

where $\{b_t\}_{t\geq 0}$ is a locally bounded and predictable drift process, $\{\sigma_t\}_{t\geq 0}$ is a cádlág process, $\{W_t\}_{t\geq 0}$ is a Brownian motion, and $\{J_t\}_{t\geq 0}$ is a jump process. The quadratic variation of this process over one day is

$$QV_t = \int_{t-1}^{t} \sigma_s^2 ds + \sum_{t<s\leq t} (\Delta p_s)^2. \tag{34}$$

Due to its latent nature, we turn to the realized variance, which is a natural estimator of quadratic variation (Andersen, Bollerslev, Diebold, and Labys, 2001), and hence a good proxy for the conditional variance (Patton, 2011). For a specific business day t, the realized variance is given by the sum of squared intraday returns, $r_{t,j} = p_{t,j} - p_{t,j-1}$,

$$RV_t = \sum_{j=1}^{n} r_{t,j}^2, \quad t = 1,\ldots,N, \tag{35}$$

where $1/n$ is the sampling frequency. As n increases, this estimator converges (in probability) to the quadratic variation of the price process. In practice, however, we only observe a noisy version of the efficient price due to the presence of market microstructure effects such as bid-ask bounce and rounding. In order to avoid problems introduced by the presence of this noise, we sample the price every 5 minutes (Hansen and Lunde, 2006; Liu, Patton, and Sheppard, 2015), thus leaving us with 78 returns for each full trading day.

A. Data

The data set consists of 5-minute observations of the liquid SPY exchange traded fund that tracks the S&P 500 Index, which is used in several other studies on variance measurement, modeling, and forecasting. We collect data for the period February 2001 to December 2013 and restrict attention to the official trading hours 9:30:00 and 16:00:00 local New York time. We remove days with shortened trading sessions. In total, we obtain data for 3,232 business days. Figure 2 depicts the evolution of daily returns and relevant realized measures used in constructing the models considered in this section, but restricts attention to the ranking sample defined below (approximately last five years of the original sample).[6] They all show the expected patterns with noticeable moves during periods of market stress in 2010 and 2011.

[6] For a complete overview of the high-frequency based measures used in this section please see Appendix C.

≪ Insert Figure 2 about here ≫

B. Set of forecasting methods and unconditional results

We consider a set of forecasting methods summarized in Table 7. The set consists of two families of models used in the variance forecasting literature, namely the (G)ARCH framework and the more recent Heterogeneous Autoregressive (HAR) framework initiated by Corsi (2009), as well as a hybrid specification in the form of the Realized GARCH (RGACRH) by Hansen et al. (2012). Details and specifications of the models and their estimation can be found in Appendix C.[7] Consequently, we examine a mixture of old and recent, simple and complex, nested and non-nested models, with or without estimation uncertainty arising from different estimation methods. Up until now, a general joint examination of the (state dependent) differences in the performance of such methods that accounts for parameter uncertainty and non-stationarity in data has not been possible without resorting to multiple testing procedures. We estimate the models using a rolling window of $m = 1000$ business days consistent with most empirical studies, hence use the first 1,022 observations for estimation, leading to 2,210 forecasts.

≪ Insert Table 7 about here ≫

We note that the distributional and permutation invariance results proposed in the former sections hold for a general loss function, however since the conditional variance is unobserved, the losses have to be calculated using a proxy. As shown in Patton (2011), this limits the set of loss functions that can be used to compare the models' forecasting ability to the class of so-called robust loss functions. One such robust loss function is the Quasi Likelihood (QLIKE) loss function, which is the one we will adopt in this paper. The QLIKE loss function is given by

$$L^i(RV_{t+1}, \widehat{RV}^i_{t+1}) = \frac{RV_{t+1}}{\widehat{RV}^i_{t+1}} - \log\left(\frac{RV_{t+1}}{\widehat{RV}^i_{t+1}}\right) - 1, \quad i = 1, \ldots, k+1, \tag{36}$$

[7] The set of models considered in the present paper has been chosen based on the fact that they require different estimation methods and represent different levels of complexity - a feature which our theory is able to accommodate. Thus, the focus has not necessarily been on finding the best possible specification of each of the given models, but rather to illustrate the flexibility of our approach. Moreover, to facilitate comparison between the (G)ARCH and HAR frameworks, daily returns are computed via open-to-close prices.

where RV_{t+1} is the realized volatility, which is our proxy for the conditional variance, and $\widehat{RV}_{t+1}$ is the forecast generated by the i'th forecasting method. We make this choice of loss function as opposed to the Squared Prediction Error, which is also contained in the class of robust loss functions of Patton (2011), since the former leads to more power when comparing losses across different regimes, which arguably is relevant in our data set. We apply an "insanity filter" like e.g. Bollerslev, Patton, and Quaedvlieg (2016) and Patton and Sheppard (2015) and replace negative forecasts with the forecasts generated by the Random Walk, which only happens six times in the entire sample, hence playing no role on results besides enabling evaluation of the QLIKE loss function.

The rightmost column in Table 7 reports the average QLIKE loss for each forecasting method over the ranking sample. On average, the HAR forecasting methods appear to perform better than the traditional (G)ARCH specifications, confirming the findings in e.g. Andersen et al. (2003).[8] The inclusion of a realized measure of volatility as in the realized GARCH model of Hansen et al. (2012) appears to improve substantially on the performance of the GARCH framework, making it comparable to the best HAR type models. Due to the large differences in forecasting performance it is not surprising that an unconditional test on the entire set leads to a strong rejection of the null hypothesis of equal predictive ability. This also applies if we take out the AR(1) and ARCH(1) that perform particularly bad on this sample, indicating the relevance of capturing the long-memory feature of the variance process. Furthermore, it appears that the models proposed by Patton and Sheppard (2015) perform almost equally well, which is supported by no rejection (p-value of 0.3893) of an unconditional test of equal predictive ability among these four forecasting methods. Likewise, an unconditional test of equal predictive ability between the RGARCH and HARQ do not reject with a p-value of 0.9880. In the following we investigate what drives some of these (in)differences in performances among the forecasting methods by means of the ranking rule in Section V and multivariate conditional tests of predictive ability.

[8] Andersen et al. (2004) and Sizova (2011) provide theoretical justifications for this result by showing that model misspecification and estimation errors of the realized measure used as proxy for conditional variance may cause model-based forecasts (such as the ones from G(ARCH) specifications) to be inferior relative to reduced-form forecasts (such as the ones from HAR specifications).

C. Conditional results

To investigate the conditional predictive ability of the individual forecasting methods we perform Step 0 of the ranking rule on a rolling basis with 1,000 observations and use a significance level of 10% in the implementation of Step 1. This leaves a total of 1,210 days for examination below, henceforth referenced as the “ranking sample”.

We introduce two classes of state variables. First, we construct a bivariate state variable indicating whether daily returns were negative at time t, i.e. $\mathscr{R}_t = \mathbb{1}\{r_t < 0\}$ for $t = 1,\dots,N$, where r_t denotes daily returns computed via open-close prices to avoid any overnight and weekend effects. We will refer to this as the “leverage state variable” for obvious reasons. Secondly, we construct a set of state variables indicating whether a negative jump, no jump or positive jump occurred at day t. Let $\hat{J}_t$ denote the jump test statistic in Barndorff-Nielsen and Shephard (2006)[9], which enables testing for jumps in intraday returns at time t and obeys $\hat{J}_t \xrightarrow{d} N(0,1)$ under the null hypothesis of no jumps. We refer to the paper for additional details. Then, we define

$$\mathscr{J}_t^{(1)} = 1_{\{SJ_t>0\}}\mathbb{1}\{\hat{J}_t > z_{1-\alpha}\}, \tag{37}$$

$$\mathscr{J}_t^{(2)} = 1_{\{SJ_t<0\}}\mathbb{1}\{\hat{J}_t > z_{1-\alpha}\}, \tag{38}$$

where $SJ_t = RS^+ - RS^-$ is the signed jump variation, measuring the variation in intraday returns attributable to jumps of either positive or negative sign and $RS^+ = \sum_{t,j}^n r_{t,j}^2 \mathbb{1}\{r_{t,j} > 0\}$ and $RS^- = \sum_{t,j}^n r_{t,j}^2 \mathbb{1}\{r_{t,j} < 0\}$ are the positive and negative realized semi-variances (Barndorff-Nielsen, Kinnebrouk, and Shephard, 2010). The $(1-\alpha)$ quantile of the standard normal distribution is denoted by $z_{1-\alpha}$. We will refer to these variables as the “jump state variables” and use a significance level of 1% for determination of the presence of a jump as in Barndorff-Nielsen and Shephard (2006). By construction, the jump state variables are equal to zero if there is no jump at day t. If there is one or more jumps during day t, then $\mathscr{J}_t^{(1)}$ ($\mathscr{J}_t^{(2)}$) will equal unity if the positive (negative) jumps contribute the most to the

[9] The test statistic is given by $\hat{J}_t = \sqrt{n}\frac{RV_t - BPV_t}{\sqrt{(\pi^2/4+\pi-5)TQ_t}}$, where $BPV_t = \frac{\pi}{2}\sum_{j=2}^n |r_{t,j}||r_{t,j-1}|$ and $TQ_t = n\left(\frac{\Gamma(1/2)}{2^{2/3}\Gamma(7/6)}\right)^3 \sum_{j=3}^n |r_{t,j}|^{4/3}|r_{t,j-1}|^{4/3}|r_{t,j-2}|^{4/3}$.

daily price movements. The leverage state variable is just moderately correlated with the jump state variables (-23% with $\mathscr{J}_t^{(1)}$ and 19% with $\mathscr{J}_t^{(2)}$), suggesting they represent distinct market states.

We examine three cases separately. First, we set $\boldsymbol{h}_t = (1, \mathscr{R}_t)'$ to investigate in isolation the impact of a negative return on the previous day. Secondly, we set $\boldsymbol{h}_t = (1, \mathscr{J}_t)'$, with $\mathscr{J}_t = (\mathscr{J}_t^{(1)}, \mathscr{J}_t^{(2)})$ to isolate the impact of a jump on the previous day and, finally, we set $\boldsymbol{h}_t = (1, \mathscr{R}_t, \mathscr{J}_t)'$ to examine the impact of jumps and the leverage effect in conjunction. Figure 3 depicts M_1 (the best method confidence set) over calendar time decomposed into leverage states and non-leverage states.[10]

≪ Insert Figure 3 about here ≫

A few things stand out. Firstly, there is a remarkable persistence in which models are included in the set M_1 during both states. Secondly, the figure confirms that HAR specifications outperform the traditional (G)ARCH specifications both in normal and leverage states. Even though the GJR specification is build to capture leverage effects, it is not included in the best set during leverage states. Instead, the first three leverage models of Patton and Sheppard (2015) are preferred until early 2012, at which point the HARQ and RGARCH models take over during the leverage state. Interestingly, RGARCH is included only in the no-leverage state during the time until the beginning of 2012. By the third quarter of 2012 the HARQ specification appears to take over the role as the most commonly included model in M_1 during no-leverage states.

The vertical lines in the figure mark two events in the U.S. stock market that appear to have a large influence on the forecasting performance of the models under consideration. On May 6, 2010, the Flash Crash occurred with the S&P 500 Index collapsing and rebounding rapidly resulting in turmoil in the following months. January 3, 2012, was the first trading day of 2012 and marked the beginning of a period of lower volatility. The periods following these events are thus characterized by very different volatility regimes as documented in Figure 2

[10] We use here and for the remainder of the empirical section (unless otherwise stated) the power-enhanced TW statistic with soft thresholding and $C = 2/3$. We have also experimented with different orderings of the forecasting methods, but results are unaltered, which is in line with the (asymptotic) permutation invariance property.

above. Interestingly, during normal states, HAR and occasionally HAR-RS-II are included in the best set up until the Flash Crash, but drop out after this event. After the markets have calmed again, the HARQ, HAR-RS-I, HAR-RS-II, and HAR-SJ-I specifications appear to make a comeback for the first few months of 2012. Such structural breaks in (relative) conditional predictive ability of the forecast methods highlight the importance of having a test that is valid even if the data is non-stationary.[11] These findings are robust to different estimation windows, m, and inclusion of a HAR model estimated using a short window of 250 days, suggesting that the identified structural breaks are not attributable to rigidity in parameters of the HAR model, but rather a regime shift in volatility.

Interestingly, we observe that the RGARCH model mainly is preferred in the normal states, whereas the leverage models HAR-RS-I, HAR-RS-II, HAR-SJ-I as well as the HARQ model for the second half of the sample are preferred in the leverage states. This indicates that the superior performance on average in Table 7 of the RGARCH and HARQ models originates from distinctively different states.

Lastly, despite the fact that the four models of Patton and Sheppard (2015) perform equally well on average, our analysis reveals that the (relative) performance of these models differ in important ways. In particular, inclusion of the HAR-SJ-II specification mainly occurs during normal states, whereas the remaining three models primarily are included during the leverage state. Interestingly, this suggests that average unconditional superiority (relative to simpler models) of HAR-SJ-II in Table 7 is driven by performance in normal states, whereas the gains in the remaining three models of Patton and Sheppard (2015) are derived from leverage states. In Table 8, we report a summary of M_1 conditional of the relevant states and whether M_1 for each time period contains only a single forecasting method.

≪ Insert Table 8 about here ≫

The table confirms the overall ranking of the forecasting methods from Table 7 in the first column, and confirms that the forecasting gain of HARQ, HAR-RS-

[11] The presence of the regime shifts is robust to a different choice of rolling window used in Step 0 equal to 500 observations, indicating that it is not caused by observations from the 2008 market turmoil dropping out of the rolling window.

I, HAR-RS-II, and HAR-SJ-I relative to the simpler methods is derived from leverage states, whereas the HAR-SJ-II and RGARCH perform particularly well during normal states. In fact, whenever M_1 is a singleton (29% of the sample), it is most often either HARQ or RGARCH.

Figure 4 depicts a corresponding plot when we condition only on the jump variables, and Table 9 provides a summary of the resulting M_1.

≪ Insert Figure 4 about here ≫

≪ Insert Table 9 about here ≫

Interestingly, and in contrast to the clear differences across states in the leverage case, it appears that the same models to a large degree are chosen almost independently of the jump states and sign. That is, jumps appear to have little effect in general on the relative forecasting ability among the forecasting methods. They do, however, play a noticeable role in the 2010-2011 period, which has been characterized by a large degree of market turmoil and, according to Figure 2, several large jumps. From the Flash Crash in May 2010 and until the end of 2011, the HAR-RS-I and HAR-RS-II are generally excluded from the best set in jump states. Instead, this period is dominated by specifications explicitly accounting for jumps, i.e. the HAR-J, HAR-SJ-I and HAR-SJ-II, as well as the RGARCH model. Furthermore, as it was the case in the leverage scenario considered previously, the baseline HAR model is only included during the initial part of the sample. Despite the fact that the RGARCH model is included in M_1 74% of the time, it is never chosen when M_1 is a singleton. Instead, we find that in the case where M_1 is a singleton (10% of the sample), it consists of either HARQ or HAR-SJ-II. The fact that the HARQ model is the most likely one to be chosen in this case is interesting because the model not directly accounts for the jumps, although as argued in Bollerslev et al. (2016) the jumps are indirectly accounted for through the inclusion of realized quarticity.

Table 10 reports results from the joint case with the test function containing both the leverage and jump state variables.

≪ Insert Table 10 about here ≫

The table generally confirms the findings from the separate cases above, however, we derive an additional conclusion. In the negative jump days the Random Walk is included in the best set around 38-39% of the times independently of being in a leverage or no leverage state. That is, forecasting the conditional variance following days with negative jumps and beating a Random Walk in predictive ability appears to be particularly challenging. Furthermore, the RGARCH model remains the most commonly included model in M_1, and it is picked in 74.3% of the cases where M_1 is a singleton.

D. Dynamic forecast combination

It stands out from the former section that the forecasting methods' predictive ability is time-varying in two ways. First, two structural breaks appear to occur during the sample period. Secondly, the forecasting methods' predictive ability relative to each other depend on the state of the market as characterized by jump and/or leverage states. Based on these state-dependencies, about one-fifth of the days we were able to identify a single superior model, mainly chosen among the RGARCH of Hansen et al. (2012), HARQ of Bollerslev et al. (2016), and HAR-SJ-II of Patton and Sheppard (2015). For the remaining days, the best two or more forecasting methods provide statistically indistinguishable predictive ability, comprising a best method confidence set, M_1, at each day (whose composition varies over time). This suggests a potentially beneficial conditional, dynamic forecast combination procedure for each day as the following:

- If $M_1 = \{i\}$ (singleton), select the i'th forecasting method,
- otherwise, perform forecast combination within M_1.

This section thus evaluates the performance of such conditional forecast combination procedure, which exploits predictability of forecast losses identified by the test statistic developed in this paper. A related approach is put forward in Aiolfi and Timmermann (2006), who conduct forecast combination within 'clusters' of forecasting models with most predictable forecast errors based on lagged forecast errors. Among a very large set of models supposed to forecast quarterly macroeconomic data, they find gains relative to choosing the previous best forecasting model at each time point. Recently, Wang, Ma, Wei, and Wu (2016) forecast realized variance of the S&P 500 Index' returns using a set of HAR specifications comparable to the set examined in this paper. In an attempt to exploit (unobserved) time-varying predictive ability of the models, they implement a Dynamic Model Averaging combination. Despite its higher degree of sophistication, it leads to limited gains relative to individual forecasts generated by the models in Patton and Sheppard (2015).

To capture potential persistence in the forecast losses, we include lagged loss differentials as a state variable in the test function, hence set $\boldsymbol{h}_t = (1, \mathscr{R}_t, \mathscr{J}_t, \Delta \boldsymbol{L}_t)'$. For each time period, we compute combination weights using a window of m_w past days via the following expression

$$\hat{w} = K^{-1}\iota_K' + \frac{1}{1+g}\left(\hat{w}_{\mathrm{LS}} - K^{-1}\iota_K'\right), \tag{39}$$

where K denotes the number of elements in M_1 and ι_K is a K-dimensional vector of ones. The combination weights in (39) are the conventional g-prior shrinkage weights (see e.g. Zellner (1986)). Here, $\hat{w}_{\mathrm{LS}}$ denotes the estimated time-varying least squares (LS) weights (Bates and Granger, 1969; Granger and Ramanathan, 1984; Diebold and Pauly, 1987) using the restrictions that the weights are non-negative (to ensure non-negative forecasts) and contain no intercept in the regression specification.[12] Effectively, g controls the shrinkage towards equal weights away from the LS estimator and, thus, controls impact of estimation error. This weight estimation is motivated by the simulation study in Elliott and Timmermann (2005), which reveals that a rolling window least squares estimator may be preferable when combination weights are subject to a structural break, whereas a simple equal-weighted average may be preferable in cases with frequent regime shifts. Both instances may be present qua the findings in the former section, motivating a weighting scheme that enables the presence of both.

A direct competitor for the conditional forecast combination (FC) procedure is a naive forecast combination, which utilizes information in all forecasting methods at each time point by combining within the entire M_0. By pre-selection at each

[12] Imposing the additional restrictions that the weights are (weakly) less than unity and sum to one corresponds to the mean square optimal weights in Bates and Granger (1969). However, they may lead to inferior results according to e.g. Granger and Ramanathan (1984); Holmen (1987), hence we proceed without this restriction.

time point a relevant set of forecasting methods (the best method confidence set, M_1), the conditional forecast combination trades off information from methods for less estimation error in combination weights, potentially leading to superior performance. We examine both cases in the following.

We consider $g = 0.33, 1, 3$ corresponding to a case with approximately 75%, 50%, 25% weight put on the LS weights, respectively. We choose a medium length of the rolling window equal to two years, $m_w = 375$, to ensure a fair basis of comparison between the conditional and naive methods, though conclusions are qualitatively unaltered if a shorter window of 250 days (one year) or a longer window of 750 days (three years) is used. To make the relative gains stand out more clearly, we normalize the QLIKE loss measures of the relevant forecast combination procedures by the QLIKE loss of each individual model such that a number below unity indicates superiority of the forecast combination procedure relative to the individual models. Table 11 reports the results.

≪ Insert Table 11 about here ≫

The conditional forecast combination procedure systematically improves upon the individual forecasting methods' performances for all values of g. Specifically, it provides a gain relative to the HAR model of approximately 17-18%, about 10-13% relative to the leverage models of Patton and Sheppard (2015) and about 4-5 % to the RGARCH and HARQ of Hansen et al. (2012) and Bollerslev et al. (2016), respectively. We consider this finding an interesting and promising result for the ranking algorithm proposed above, considering that i) the RGARCH and HARQ models are chosen almost one-fourth of the times whenever M_1 is a singleton, and ii) the forecast errors of the methods in M_1 whenever it is not a singleton arises from similar models leading to highly correlated forecast errors and, hence, a limit on the gain of forecast combination. Despite this, an unconditional test of equal predictive ability between the HARQ specification and the conditional forecast combination procedure (for each value of g) rejects on conventional levels with p-values of 0.0060, 0.0029, and 0.0295 for $g = 0.33, 1, 3$, respectively, documenting a statistically significant gain in terms of predictive ability of the proposed procedure. Relative to the naive combination, the benefit of narrowing down M_0 to a relevant set of forecasting methods at each time point is clear. Estimation error seems to dominate the performance of the naive combination strategy, leading to inferior results in general relative to most HAR models and a significant 18% relative to the conditional forecast combination procedure proposed here.

VII. Conclusion

Our new statistical tests for equal conditional predictive ability among a set of two or more forecasting methods may be seen as a multivariate generalization of the Giacomini-White tests (Giacomini and White, 2006), and in a special case provide an extension of the multivariate Diebold-Mariano test statistic in Mariano and Preve (2012) for equal unconditional predictive ability. They apply in a setting that allows for a mixture of nested and non-nested models as well as non-stationarity in data, and explicitly accounts for estimation uncertainty in parameters used to make predictions. All our tests hold for a general loss function, have chi-squared limiting distributions, and are generally invariant to any reordering of the forecasting methods, thus facilitating easy implementation.

Simulations suggest that our tests have good finite-sample size and power. To potentially improve upon statistical properties of the test statistics in the case with many methods and/or instruments, we introduce two finite-sample adjustments. First, we developed test statistics employing a threshold estimator of the covariance matrix, and secondly, we introduced a power enhancement component along the lines of Fan et al. (2015). The simulation study confirms that the finite-sample corrections succeed in improving both size and power.

A new Model Confidence Set (Hansen et al., 2011) inspired rule allows for ranking the forecasting methods into sets containing forecasting methods of indistinguishable conditional predictive ability. In an empirical application to forecasting the conditional variance of the S&P 500 Index' returns, we provide evidence of what drives (in)differences in forecasting performance between a diversified set of forecasting methods including (G)ARCH, Realized GARCH, and HAR specifications. The results show, among other things, that exploiting the ranking rule in a novel conditional forecast selection procedure leads to significant gains in predictive ability relative to individual forecasting methods and competing forecast combination methods. Finally, our empirical work shows that there is room for further improvement in the forecasting of return variance following days with negative jumps.

References

AIOLFI, M. AND A. TIMMERMANN (2006): "Persistence in forecasting performance and conditional combination strategies," *Journal of Econometrics*, 135, 31–53.

ANDERSEN, T. G., T. BOLLERSLEV, AND F. X. DIEBOLD (2007): "Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility," *Review of Economics and Statistics*, 89, 701–720.

ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND P. LABYS (2001): "The distribution of realized exchange rate volatility," *Journal of the American Statistical Association*, 96, 42–55.

——— (2003): "Modeling and forecasting realized volatility," *Econometrica*, 71, 579–625.

ANDERSEN, T. G., T. BOLLERSLEV, AND N. MEDDAHI (2004): "Analytical evaluation of volatility forecasts," *International Economic Review*, 45, 1079–1110.

ANDREWS, D. W. K. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59, 817–858.

BARNDORFF-NIELSEN, O. E., S. KINNEBROUK, AND N. SHEPHARD (2010): "Measuring downside risk: realised semivariance," *In Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle (Edited by T. Bollerslev, J. Russel and M. Watson)*, Oxford University Press, 117–136.

BARNDORFF-NIELSEN, O. E. AND N. SHEPHARD (2006): "Econometrics of testing for jumps in financial economics using bipower variation," *Journal of Financial Econometrics*, 4, 1–30.

BATES, J. M. AND C. W. J. GRANGER (1969): "The combination of forecasts," *Operational Research Quarterly*, 20, 451–468.

BICKEL, P. J. AND E. LEVINA (2008): "Covariance regularization by thresholding," *Annals of Statistics*, 36, 2577–2604.

BOLLERSLEV, T. (1986): "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

BOLLERSLEV, T., A. J. PATTON, AND R. QUAEDVLIEG (2016): “Exploiting the errors: A simple approach for improved volatility forecasting,” *Journal of Econometrics*, 192, 1–18.

CLARK, T. E. AND M. MCCRACKEN (2001): “Tests of equal forecast accuracy and encompassing for nested models,” *Journal of Econometrics*, 105, 85–110.

——— (2012): “Reality checks and comparisons of nested predictive models,” *Journal of Business and Economic Statistics*, 30, 53–66.

——— (2013): “Advances in forecast evaluation,” *In Handbook of Economic Forecasting Volume 2, Elsevier B.V.*, 1107–1201.

CORSI, F. (2009): “A simple approximate long-memory model of realised volatility,” *Journal of Financial Econometrics*, 7, 174–196.

DIEBOLD, F. X. AND R. S. MARIANO (1995): “Comparing Predictive Accuracy,” *Journal of Business and Economic Statistics*, 13, 253–263.

DIEBOLD, F. X. AND P. PAULY (1987): “Structural change and the combination of forecasts,” *Journal of Forecasting*, 6, 21–40.

DONOHO, D. L. AND I. M. JOHNSTONE (1994): “Ideal spatial adaption by wavelet shrinkage,” *Biometrika*, 81, 425–455.

ELLIOTT, G. AND A. TIMMERMANN (2005): “Optimal forecast combination under regime switching,” *International Economic Review*, 46, 1081–1102.

ENGLE, R. F. (1982): “Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation,” *Econometrica*, 50, 987–1008.

FAMA, E. AND K. R. FRENCH (1997): “Industry cost of equity,” *Journal of Financial Economics*, 43, 153–193.

FAN, J. AND R. LI (2001): “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.

FAN, J., Y. LIAO, AND M. MINCHEVA (2013): "Large covariance estimation by thresholding principal orthogonal complements," *Journal of Royal Statistical Society*, 75, 603–680.

FAN, J., Y. LIAO, AND J. YAO (2015): "Power enhancement in high-dimensional cross-sectional tests," *Econometrica*, 83, 1497–1541.

GIACOMINI, R. AND B. ROSSI (2010): "Forecast comparisons in unstable environments," *Journal of Applied Econometrics*, 25, 595–620.

GIACOMINI, R. AND H. WHITE (2006): "Tests of conditional predictive ability," *Econometrica*, 74, 1545–1578.

GLOSTEN, L. R., R. JAGANNATHAN, AND D. E. RUNKLE (1993): "On the relation between the expected value and the volatility of the nominal excess return on stocks," *The Journal of Finance*, 48, 1779–1801.

GONÇALVES, S., M. MCCRACKEN, AND B. PERRON (2017): "Tests of equal accuracy for nested models with estimated factors," *Journal of Econometrics*, 198, 231–252.

GOYAL, A. AND I. WELCH (2003): "Predicting the equity premium with dividend ratios," *Management Science*, 49, 639–654.

GRANGER, C. W. J. AND M. J. MACHINA (2006): "Forecasting and decision theory," *In Handbook of Economic Forecasting, Elsevier B.V.*, 1, 82–98.

GRANGER, C. W. J. AND R. RAMANATHAN (1984): "Improved method of combining forecasts," *Journal of Forecasting*, 3, 197–204.

GRANZIERA, E., K. HUBRICH, AND H. R. MOON (2014): "A predictability test for a small number of nested models," *Journal of Econometrics*, 182, 174–185.

HANSEN, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

HANSEN, P. R., Z. HUANG, AND H. H. SHEK (2012): "Realixed GARCH: a joint model for rreturn and realized mesures of volatility," *Journal of Applied Econometrics*, 27, 877–906.

HANSEN, P. R. AND A. LUNDE (2006): "Realized variance and market microstructure noise," *Journal of Business and Economic Statistics*, 24, 127–161.

HANSEN, P. R., A. LUNDE, AND J. M. NASON (2011): "The model confidence set," *Econometrica*, 79, 453–497.

HANSEN, P. R. AND A. TIMMERMANN (2015): "Equivalence between out-of-sample forecast comparisons and Wald statistics," *Econometrica*, 83, 2485–2505.

HOLMEN, J. S. (1987): "A note on the value of combining short-term earnings forecasts: a test of Granger and Ramanathan," *International Journal of Forecasting*, 3, 239–243.

HUBRICH, K. AND K. D. WEST (2010): "Forecast evaluation of small nested model sets," *Journal of Applied Econometrics*, 25, 574–594.

LIU, L. Y., A. J. PATTON, AND K. SHEPPARD (2015): "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes," *Journal of Econometrics*, 187, 293–311.

MARIANO, R. S. AND D. PREVE (2012): "Statistical tests for multiple forecast comparison," *Journal of Econometrics*, 169, 123–130.

MCCRACKEN, M. (2007): "Asymptotics for out of sample tests of Granger causality," *Journal of Econometrics*, 140, 719–752.

NEWEY, W. K. AND K. D. WEST (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance marix," *Econometrica*, 55, 703–708.

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

PATTON, A. J. AND K. SHEPPARD (2015): "Good volatility, bad volatility: Signed jumps and the persistence of volatility," *Review of Economics and Statistics*, 97, 683–697.

PAYE, B. S. AND A. TIMMERMANN (2006): "Instability of return prediction models," *Journal of Empirical Finance*, 13, 274–315.

PHILLIPS, P. C. B. AND S. JIN (2014): "Testing the Martingale Hypothesis," *Journal of Business and Economic Statistics*, 32, 537–554.

ROMANO, J. P., A. M. SHAIKH, AND M. WOLF (2010): "Hypothesis testing in Econometrics," *Annual Review of Economics*, 2, 75–104.

ROTHMAN, A. J., E. LEVINA, AND J. ZHU (2009): "Generalized thresholding of large covariance matrices," *Journal of American Statistical Association*, 104, 177–186.

SCHRIMPF, A. AND Q. WANG (2010): "A reappraisal of the leading indicator properties of the yield curve under structural instability," *International Journal of Forecasting*, 26, 836–857.

SIZOVA, N. (2011): "Integrated variance forecasting: Model based vs. reduced form," *Journal of Econometrics*, 162, 294–311.

SO, M. E. P., K. LAM, AND W. K. LI (1998): "A stochastic volatility model with Markov switching," *Journal of Business and Economic Statistics*, 16, 244–253.

STOCK, J. H. AND M. W. WATSON (1999): "Forecasting inflation," *Journal of Monetary Economics*, 44, 293–335.

——— (2003): "Forecasting output and inflation: the role of asset prices," *Journal of Economic Literature*, 41, 788–829.

WANG, Y., F. MA, Y. WEI, AND C. WU (2016): "Forecasting realized volatility in a changing world: A dynamic model averaging approach," *Journal of Banking and Finance*, 64, 136–149.

WELCH, I. AND A. GOYAL (2008): "A comprehensive look at the empirical performance of equity premium prediction," *Review of Financial Studies*, 21, 1455–1508.

WEST, K. D. (1996): "Asymptotic inference about predictive ability," *Econometrica*, 64, 1067–1084.

——— (2006): "Forecast evaluation," *In Handbook of Economic Forecasting, Elsevier B.V.*, 1, 100–134.

——— (2008): “Heteroskedasticity and autocorrelation corrections,” *In: Durlauf, S. N., Blume, L. E. (Eds.), Macroecometrics and Time Series Analysis*, 135–144.

WHITE, H. (1994): *Estimation, inference and specification analysis*, New York: Cambridge University Press.

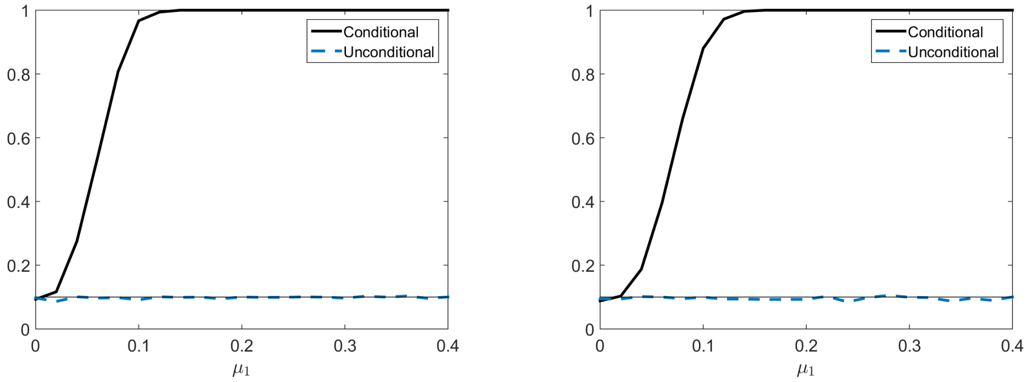
——— (2001): *Asymptotic theory for econometricians*, San Diego: Academic Press.

WOOLDRIDGE, J. M. AND H. WHITE (1988): “Some invaraince principles and central limit theorems for dependent heterogeneous processes,” *Econometric Theory*, 4, 210–230.

ZELLNER, A. (1986): “On assessing prior distributions and Bayesian regression analysis with g -prior distributions,” *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6, 233–43.

A. Figures

Figure 1: Power curves for the conditional and unconditional test

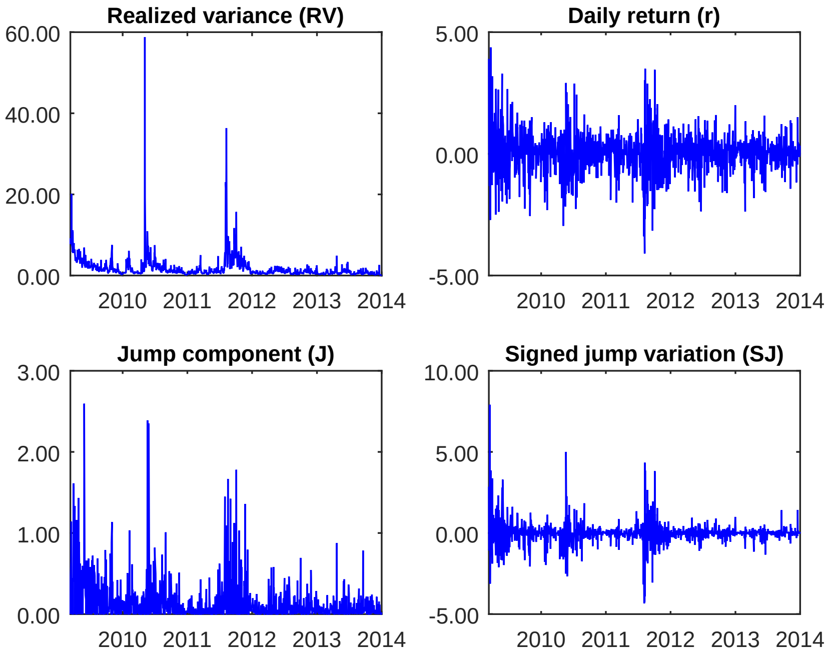


(a) Five methods ($k = 4$)

(b) Ten methods ($k = 9$)

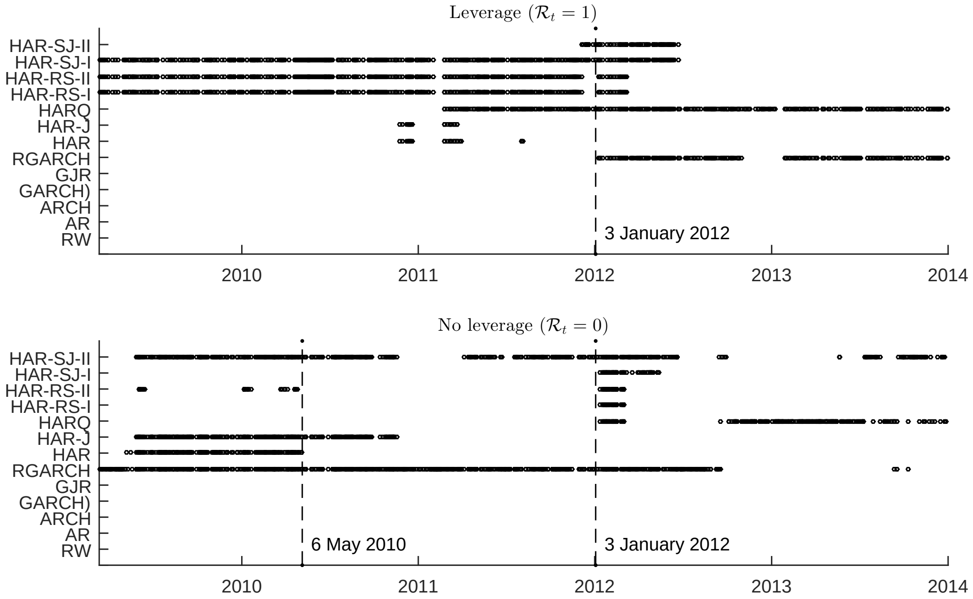
Figure a) and b) depict the power curves for the unconditional and conditional test statistic with $T = 500$ observations and five or ten forecasting methods, respectively, and data generating process according to (32). The solid horizontal line marks the significance level of 10%.

Figure 2: Evolution of realized measures



This figure depicts the evolution of relevant realized measures and daily returns of the S&P 500 Index over the ranking sample period (approximately 2009-2014). See Appendix C for a definition of the variables. Realized measures are annualized and shown in percentages, whereas daily returns are shown in percentages.

Figure 3: Diagnostics of M_1 when $\boldsymbol{h}_t = (1, \mathscr{R}_t)'$

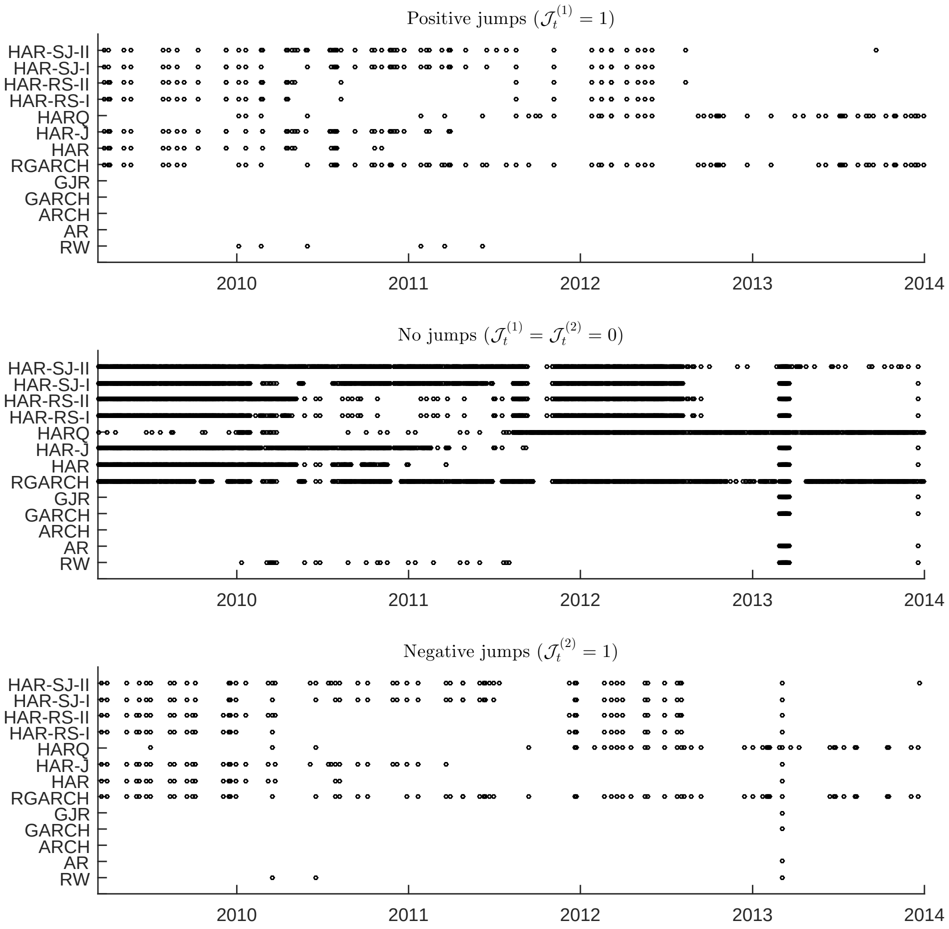


Leverage ($\mathcal{R}_t = 1$)

No leverage ($\mathcal{R}_t = 0$)

This figure depicts the inclusions in M_1 (best method confidence set) of each forecasting method from Table 7 conditional on being in a leverage (upper figure) or non-leverage state (lower figure). A circle represents inclusion in M_1. The vertical dashed lines mark two important events for the U.S. stock market. The figure is generated by means of the ranking rule in Section V with rolling estimation windows of 1,000 days for generating the forecasts and ranking the forecasting methods. We implement Step 1 using the power-enhanced TW statistic with soft thresholding and $C = 2/3$ and a significance level of 10%.

Figure 4: Diagnostics of M_1 when $\boldsymbol{h}_t = (1, \mathscr{J}_t)'$



This figure depicts the inclusions in M_1 (best method confidence set) of each forecasting method from Table 7 conditional on being in a positive jump (upper figure), no jump (middle figure) or negative jump state (lower figure). A circle represents inclusion in M_1. The figure is generated by means of the ranking rule in Section V with rolling estimation windows of 1,000 days for generating the forecasts and ranking the forecasting methods. We implement Step 1 using the power-enhanced TW statistic with soft thresholding and $C = 2/3$ and a significance level of 10%.

B. Tables

Table 1: Empirical size with no finite-sample corrections

This table reports the rejection frequencies (empirical sizes) of the multivariate test for equal predictive density with a nominal size of 10%, data generating process given by (30) with $\boldsymbol{\mu} = \mathbf{0}$ and 10,000 Monte Carlo replications. Panel A reports results for the unconditional case, whereas Panel B reports results for the conditional case with $\boldsymbol{h}_t = (1, \Delta \boldsymbol{L}_t)'$.

	Panel A: Unconditional test		
No. of methods	T = 250	T = 500	T = 1000
2	0.102	0.103	0.099
3	0.112	0.103	0.096
4	0.116	0.093	0.093
5	0.121	0.092	0.112
	Panel B: Conditional test		
No. of methods	T = 250	T = 500	T = 1000
2	0.102	0.099	0.102
3	0.107	0.102	0.107
4	0.132	0.122	0.111
5	0.173	0.116	0.113

Table 2: Empirical size with threshold Wald statistic

This table reports the rejection frequencies (empirical sizes) of the multivariate test for equal conditional predictive ability using the TW statistic in (23) with a nominal size of 10%, data generating process given by (30) with $\boldsymbol{\mu} = \mathbf{0}$ and 10,000 Monte Carlo replications. The table reports results for the conditional case with $\boldsymbol{h}_t = (1, \Delta\boldsymbol{L}_t)'$.

		Conditional test	
No. of methods	T = 250	T = 500	T = 1000
2	0.103	0.102	0.101
3	0.096	0.103	0.103
4	0.093	0.089	0.099
5	0.093	0.088	0.086
6	0.088	0.087	0.083
7	0.082	0.083	0.085
8	0.086	0.080	0.088
9	0.088	0.088	0.083
10	0.122	0.094	0.088

Table 3: Empirical size with TW statistic and power enhancement

This table reports the rejection frequencies (empirical sizes) of the multivariate test for equal conditional predictive ability using the TW statistic including the power enhancement component ($S^{(2)}_{m,h}$ in (26)) with a nominal size of 10%, data generating process given by (30) with $\boldsymbol{\mu} = \mathbf{0}$ and 10,000 Monte Carlo replications. The table reports results for the conditional case with $\boldsymbol{h}_t = (1, \Delta \boldsymbol{L}_t)'$.

	Panel B: Conditional test		
No. of methods	T = 250	T = 500	T = 1000
2	0.294	0.240	0.204
3	0.159	0.127	0.109
4	0.138	0.108	0.096
5	0.121	0.105	0.094
6	0.114	0.096	0.084
7	0.102	0.095	0.089
8	0.108	0.086	0.082
9	0.117	0.090	0.090
10	0.130	0.099	0.097

Table 4: Empirical power with no finite-sample corrections

This table reports the rejection frequencies (empirical powers) of the multivariate test for equal predictive ability with a nominal size of 10%, data generating process given by (30) with $\boldsymbol{\mu}$ determined via (31) and 10,000 Monte Carlo replications. Panel A reports results for the unconditional case, whereas Panel B reports results for the conditional case with $\boldsymbol{h}_t = (1, \Delta \boldsymbol{L}_t)'$.

	Panel A: Unconditional test		
No. of methods	T = 250	T = 500	T = 1000
2	0.986	1.000	1.000
3	0.984	1.000	1.000
4	0.978	1.000	1.000
5	0.975	1.000	1.000
	Panel B: Conditional test		
No. of methods	T = 250	T = 500	T = 1000
2	0.974	0.999	1.000
3	0.938	0.998	1.000
4	0.886	0.997	1.000
5	0.880	0.996	1.000

Table 5: Empirical power with threshold Wald statistic

This table reports the rejection frequencies (empirical powers) of the multivariate test for equal conditional predictive ability using the TW statistic in (23) with a nominal size of 10%, data generating process given by (30) with $\boldsymbol{\mu}$ determined via (31) and 10,000 Monte Carlo replications. The table reports results for the conditional case with $\boldsymbol{h}_t = (1, \Delta\boldsymbol{L}_t)'$.

		Conditional test	
No. of methods	T = 250	T = 500	T = 1000
2	0.972	1.000	1.000
3	0.930	1.000	1.000
4	0.870	0.996	1.000
5	0.784	0.982	1.000
6	0.713	0.970	1.000
7	0.624	0.948	1.000
8	0.576	0.918	1.000
9	0.527	0.887	1.000
10	0.525	0.847	1.000

Table 6: Empirical power with TW statistic and power enhancement

This table reports the rejection frequencies (empirical powers) of the multivariate test for equal conditional predictive ability using the TW statistic including the power enhancement component ($S^{(2)}_{m,h}$ in (26)) with a nominal size of 10%, data generating process given by (30) with $\boldsymbol{\mu}$ determined via (31) and 10,000 Monte Carlo replications. The table reports results for the conditional case with $\boldsymbol{h}_t = (1, \Delta \boldsymbol{L}_t)'$.

	Panel B: Conditional test		
No. of methods	T = 250	T = 500	T = 1000
2	0.995	1.000	1.000
3	0.967	1.000	1.000
4	0.937	0.999	1.000
5	0.901	0.996	1.000
6	0.855	0.994	1.000
7	0.821	0.988	1.000
8	0.797	0.986	1.000
9	0.768	0.983	1.000
10	0.752	0.978	1.000

Table 7: Set of forecasting models

This table summarizes the set of forecasting methods considered with details given in Appendix C. The third column shows the estimation procedure associated with each forecasting method, where "OLS" refers to Ordinary Least Squares and "ML" to Maximum Likelihood. The rightmost column reports the average QLIKE loss measure over the ranking sample.

Model name	Reference	Estim. method	QLIKE
RW	N/A	N/A	0.1925
AR(1)	N/A	OLS	0.2775
ARCH(1)	Engle (1982)	ML	0.4739
GARCH(1,1)	Bollerslev (1986)	ML	0.1998
GJR(1,1,1)	Glosten, Jagannathan, and Runkle (1993)	ML	0.1969
RGARCH(1,1)	Hansen et al. (2012)	ML	0.1522
HAR	Corsi (2009)	OLS	0.1755
HAR-J	Andersen, Bollerslev, and Diebold (2007)	OLS	0.1756
HARQ	Bollerslev et al. (2016)	OLS	0.1518
HAR-RS-I	Patton and Sheppard (2015)	OLS	0.1665
HAR-RS-II	Patton and Sheppard (2015)	OLS	0.1652
HAR-SJ-I	Patton and Sheppard (2015)	OLS	0.1684
HAR-SJ-II	Patton and Sheppard (2015)	OLS	0.1621

Table 8: Inclusion frequencies of M_1 with $\boldsymbol{h}_t = (1, \mathscr{R}_t)'$

This table reports the inclusion frequencies (numbers are in percentages) of the forecasting methods under consideration in the ranking sample (last 1,210 observations in the sample). The second column provides frequencies over the entire ranking sample, whereas the third and fourth columns conditions on being in a leverage and non-leverage state, respectively. The rightmost column provides the inclusion frequency given the best method confidence set, M_1, is a singleton, i.e. contains a single element. We use the power-enhanced TW statistic from Section III with soft thresholding and $C = 2/3$.

Method	F	$F\|\mathscr{R} = 1$	$F\|\mathscr{R} = 0$	$F\|M_1 = \{i\}$
RW	0.00	0.00	0.00	0.00
AR(1)	0.00	0.00	0.00	0.00
ARCH(1)	0.00	0.00	0.00	0.00
GARCH(1,1)	0.00	0.00	0.00	0.00
GJR(1,1,1)	0.00	0.00	0.00	0.00
RGARCH(1,1)	58.86	36.14	72.63	40.63
HAR	13.55	4.40	20.52	0.57
HAR-J	18.51	3.44	29.99	0.00
HARQ	39.92	61.38	23.58	44.03
HAR-RS-I	28.10	60.61	3.35	0.00
HAR-RS-II	30.00	60.61	6.70	0.00
HAR-SJ-I	33.47	69.79	5.82	0.00
HAR-SJ-II	38.60	12.24	58.66	14.77
Pct of sample	100.00	45.84	54.16	29.09

Table 9: Inclusion frequencies of M_1 with $\boldsymbol{h}_t = (1, \mathscr{J}_t)'$

This table reports the inclusion frequencies (numbers are in percentages) of the forecasting methods under consideration in the ranking sample (last 1,210 observations in the sample). The second column provides frequencies over the entire ranking sample, whereas the third, fourth and fifth columns conditions on being in a positive, no and negative jump state, respectively. The rightmost column provides the inclusion frequency given the best method confidence set, M_1, is a singleton, i.e. contains a single element. We use the power-enhanced TW statistic from Section III with soft thresholding and $C = 2/3$.

Method	F	$F\|\mathscr{J}^{(1)} = 1$	$F\|\mathscr{J} = 0$	$F\|\mathscr{J}^{(2)} = 1$	$F\|M_1 = \{i\}$
RW	3.88	5.94	3.71	3.49	0.00
AR(1)	1.24	0.00	1.37	1.16	0.00
ARCH(1)	0.00	0.00	0.00	0.00	0.00
GARCH(1,1)	1.24	0.00	1.37	1.16	0.00
GJR(1,1,1)	1.24	0.00	1.37	1.16	0.00
RGARCH(1,1)	74.38	74.26	74.19	76.74	0.00
HAR	28.93	23.76	29.72	25.58	0.00
HAR-J	40.50	38.61	40.96	37.21	0.00
HARQ	50.17	50.50	50.44	46.51	69.67
HAR-RS-I	39.34	28.71	40.76	34.88	0.00
HAR-RS-II	43.55	32.67	44.97	39.53	0.00
HAR-SJ-I	51.82	44.55	52.69	50.00	0.00
HAR-SJ-II	67.19	57.43	68.52	62.79	30.33
Pct of sample	100.00	7.96	85.07	6.97	10.08

Table 10: Inclusion frequencies of M_1 with $\boldsymbol{h}_t = (1, \mathscr{R}_t, \mathscr{J}_t)'$

This table reports the inclusion frequencies (numbers are in percentages) of the forecasting methods under consideration in the ranking sample (last 1,210 observations in the sample). The notation is similar to Tables 8-9. We use the power-enhanced TW statistic from Section III with soft thresholding and $C = 2/3$.

		Panel A: $\mathscr{R} = 0$		
Method	F	$F \mid \mathscr{J}^{(1)} = 1$	$F \mid \mathscr{J} = 0$	$F \mid \mathscr{J}^{(2)} = 1$
RW	9.50	5.26	7.73	39.13
AR(1)	5.62	4.21	4.92	17.39
ARCH(1)	0.00	0.00	0.00	0.00
GARCH(1,1)	5.62	4.21	4.92	17.39
GJR(1,1,1)	5.79	4.21	4.92	17.39
RGARCH(1,1)	70.99	20.00	98.24	100.00
HAR	29.67	24.21	39.19	17.39
HAR-J	29.34	43.16	39.19	17.39
HARQ	47.85	4.21	47.80	65.22
HAR-RS-I	40.91	65.26	18.45	17.39
HAR-RS-II	42.73	72.63	19.68	17.39
HAR-SJ-I	39.26	4.21	20.39	17.39
HAR-SJ-II	34.46	51.58	44.99	17.39
Pct of sample	100.00	7.85	47.02	1.90

		Panel B: $\mathscr{R} = 1$		
Method	$F \mid M_1 = \{i\}$	$F \mid \mathscr{J}^{(1)} = 1$	$F \mid \mathscr{J} = 0$	$F \mid \mathscr{J}^{(2)} = 1$
RW	0.00	33.33	6.83	38.10
AR(1)	0.00	33.33	5.07	11.11
ARCH(1)	0.00	0.00	0.00	0.00
GARCH(1,1)	0.00	33.33	5.07	11.11
GJR(1,1,1)	0.00	33.33	5.51	11.11
RGARCH(1,1)	74.32	33.33	47.14	66.67
HAR	0.00	66.67	19.60	25.40
HAR-J	0.00	83.33	16.52	11.11
HARQ	2.70	33.33	49.56	96.83
HAR-RS-I	0.00	100.00	65.20	34.92
HAR-RS-II	0.00	100.00	67.84	28.57
HAR-SJ-I	6.76	33.33	71.81	36.51
HAR-SJ-II	16.22	33.33	21.59	12.70
Pct of sample	18.35	0.50	37.52	5.21

Table 11: Forecast selection/combination evaluation

This table reports the QLIKE loss measures of the conditional forecast combination or naive forecast combination procedures divided by the QLIKE loss measures of each individual forecasting method under consideration. The last row provides the QLIKE loss measures of the proposed procedures over the raking sample. We consider three values of the shrinkage parameter/g-prior of Zellner (1986) equal to $g = 0.33, 1, 3$ and a rolling window of 375 observations for estimation of the combination weights.

Method	Cond. FC ($g = 0.33$)	Naive FC ($g = 0.33$)	Cond. FC ($g = 1$)	Naive FC ($g = 1$)	Cond. FC ($g = 3$)	Naive FC ($g = 3$)
RW	0.7536	0.8993	0.7544	0.9002	0.7637	0.9091
AR(1)	0.5226	0.6237	0.5232	0.6242	0.5296	0.6304
ARCH(1)	0.3061	0.3653	0.3064	0.3656	0.3102	0.3693
GARCH(1,1)	0.7260	0.8664	0.7267	0.8672	0.7357	0.8758
GJR(1,1,1)	0.7367	0.8791	0.7374	0.8799	0.7465	0.8886
RGARCH(1,1)	0.9529	1.1371	0.9538	1.1381	0.9655	1.1494
HAR	0.8264	0.9862	0.8273	0.9871	0.8374	0.9969
HAR-J	0.8259	0.9856	0.8267	0.9865	0.8369	0.9962
HARQ	0.9558	1.1405	0.9567	1.1416	0.9685	1.1529
HAR-RS-I	0.8713	1.0397	0.8721	1.0406	0.8828	1.0510
HAR-RS-II	0.8779	1.0476	0.8788	1.0486	0.8896	1.0590
HAR-SJ-I	0.8614	1.0279	0.8622	1.0288	0.8728	1.0390
HAR-SJ-II	0.8951	1.0682	0.8960	1.0691	0.9070	1.0797
QLIKE	0.1451	0.1731	0.1452	0.1733	0.1470	0.1750

C. Set of forecasting models

This section very briefly introduces each element of the set of forecasting methods under consideration in Section VI. First, the forecast of the Random Walk (RW) is trivially the previous period's realization of the realized variance, whereas the autoregressive model of order one, AR(1), is defined by

$$RV_{t+1} = \beta_0 + \beta_d RV_t + \varepsilon_{t+1}, \tag{C.1}$$

which can be consistently estimated by Ordinary Least Squares (OLS).

The (G)ARCH framework of Engle (1982) and Bollerslev (1986) directly models the daily conditional variance, σ_t^2. In this paper, we consider a first-order autoregressive daily return specification (mean equation) as suggested in e.g. Andersen et al. (2003), $r_{t+1} = \varphi_0 + \varphi_1 r_t + \eta_{t+1}$, where $\eta_{t+1} = \sigma_{t+1} u_{t+1}$ and $u_{t+1} \sim N(0,1)$ i.i.d., leading to the ARCH(1) specification

$$\sigma_{t+1}^2 = \omega + \alpha \eta_t^2, \tag{C.2}$$

and the GARCH(1,1) specification

$$\sigma_{t+1}^2 = \omega + \alpha \eta_t^2 + \beta \sigma_t^2. \tag{C.3}$$

To capture the well-known property of leverage effect in financial markets, Glosten et al. (1993) considers the following extension of the GARCH(1,1) model, which is known as a GJR(1,1,1) model:

$$\sigma_{t+1}^2 = \omega + \alpha \eta_t^2 + \beta \sigma_t^2 + \gamma \eta_t^2 \mathbb{1}\{\eta_t < 0\}, \tag{C.4}$$

where the third argument in GJR(1,1,1) refers to the number of interaction components. Under the distributional assumption on u_{t+1}, the ARCH(1), GARCH(1,1) and GJR(1,1,1) models can be estimated consistently by Maximum Likelihood.

To relate realized measures such as the realized variance to the conditional variance while maintaining the features of the GARCH specifications, Hansen et al. (2012) propose the Realized GARCH(1,1) (RGARCH). The model is given by

$$r_{t+1} = \varphi_0 + \sigma_{t+1}u_{t+1} \tag{C.5}$$

$$\log(\sigma^2_{t+1}) = \omega + \beta\log(\sigma^2_t) + \gamma\log(RV_t) \tag{C.6}$$

$$\log(RV_{t+1}) = \xi + \phi\log(\sigma^2_{t+1}) + \delta(u_{t+1}) + z_t, \tag{C.7}$$

where $z_t \sim N(0,\sigma^2_z)$, u_t and z_t independent, and the asymmetric leverage function is given by $\delta(u_{t+1}) = \delta_1 u_{t+1} + \delta_2(u^2_{t+1} - 1)$. Note that we use the specification where the (vector) of realized measures is constituted only by the realized variance, but acknowledge that the RGARCH framework can be readily extended to include information from additional realized measures. The model is estimated by Maximum Likelihood and one-step forecasts of the conditional variance are obtained directly from the GARCH equation in (C.6).

Recently, the Heterogeneous Autoregressive realized variance model (HAR) by Corsi (2009) has emerged as a popular specification due to its simplicity and ability to mimic the long-memory property of conditional variance. It is given by

$$RV_{t+1} = \beta_0 + \beta_d RV_t + \beta_w RV_{w,t} + \beta_m RV_{m,t} + \varepsilon_{t+1}, \tag{C.8}$$

where $RV_{w,t}$ and $RV_{m,t}$ are the average realized variance from day $t-4$ to t and from day $t-21$ to t, respectively. The model can be consistently estimated by OLS. This is also true for the remaining models considered in this section. In the HAR-J modification, Andersen et al. (2007) add a jump component such that

$$RV_{t+1} = \beta_0 + \beta_d RV_t + \beta_j J_t + \beta_w RV_{w,t} + \beta_m RV_{m,t} + \varepsilon_{t+1}, \tag{C.9}$$

where the jump component is defined by $J_t = \max(RV_t - BPV_t, 0)$ with $BPV_t = u^{-2}\sum_{j=2}^{n}|r_{t,j-1}||r_{t,j}|$ and $u = \sqrt{(2/\pi)}$.

Exploiting the asymptotic theory for realized variance estimation, Bollerslev et al. (2016) introduce the Q-family of HAR models by introducing time-varying parameters that vary with the degree of estimation error in the realized variance measure. Specifically, we include the standard HARQ specification given by

$$RV_{t+1} = \beta_0 + (\beta_d + \beta_Q RQ_t)RV_t + \beta_w RV_{w,t} + \beta_m RV_{m,t} + \varepsilon_{t+1}, \tag{C.10}$$

where $RQ_t = n/3\sum_{j=1}^{n} r_{t,j}^4$ is the realized quarticity at time t.[13]

To capture the leverage effect, Patton and Sheppard (2015) introduce signed realized measures in the HAR specification. The first model, HAR-RS-I, decomposes daily realized variance into two semi-variances, leading to the specification

$$RV_{t+1} = \beta_0 + \beta_d^+ RS_t^+ + \beta_d^- RS_t^- + \beta_w RV_{w,t} + \beta_m RV_{m,t} + \varepsilon_{t+1}, \tag{C.11}$$

where $RS_t^+ = \sum_{j=1}^{n} r_{t,j}^2 \mathbb{1}\{r_{t,j} > 0\}$ and, correspondingly, $RS_t^- = \sum_{j=1}^{n} r_{t,j}^2 \mathbb{1}\{r_{t,j} < 0\}$. The second model, HAR-RS-II, adds an interaction term supposed to capture the leverage effect arising from the previous day's return via

$$RV_{t+1} = \beta_0 + \beta_d^+ RS_t^+ + \beta_d^- RS_t^- + \beta_l RV_t \mathbb{1}\{r_t < 0\} + \beta_w RV_{w,t} + \beta_m RV_{m,t} + \varepsilon_{t+1}. \tag{C.12}$$

The third model, HAR-SJ-I, instead introduces signed jump variation along with bi-power variation. The model is given by

$$RV_{t+1} = \beta_0 + \beta_j SJ_t + \beta_b BPV_t + \beta_w RV_{w,t} + \beta_m RV_{m,t} + \varepsilon_{t+1}, \tag{C.13}$$

where $SJ_t = RS_t^+ - RS_t^-$. The last model, HAR-SJ-II, decomposes signed jump variation into positive and negative jumps via

$$RV_{t+1} = \beta_0 + \beta_j^+ SJ_t^+ + \beta_j^- SJ_t^- + \beta_b BPV_t + \beta_w RV_{w,t} + \beta_m RV_{m,t} + \varepsilon_{t+1}, \tag{C.14}$$

where $SJ_t^+ = SJ_t \mathbb{1}\{SJ_t > 0\}$ and, correspondingly, $SJ_t^- = SJ_t \mathbb{1}\{SJ_t < 0\}$.

[13] Given the assumption on the efficient (log) price in (33), it may be more appropriate to use the (jump robust) tri-power quarticity of Barndorff-Nielsen and Shephard (2006) given by $TQ_t = n\left(\frac{\Gamma(1/2)}{2^{2/3} * \Gamma(7/6)}\right)^3 \sum_{j=3}^{n} |r_{t,j}|^{4/3} |r_{t,j-1}|^{4/3} |r_{t,j-2}|^{4/3}$. However, it leads to similar results for the HARQ specification and to remain close to the primary specification in Bollerslev et al. (2016), we proceed with the use of RQ_t.

D. Proofs

Proof of Theorem 1. Let $\boldsymbol{d}_{m,t+1} = \boldsymbol{h}_t \otimes \Delta \boldsymbol{L}_{m,t+1}$ and write

$$\boldsymbol{d}_{m,t+1}\boldsymbol{d}'_{m,t+1} = g(\boldsymbol{h}_t, W_{t+1}, \ldots, W_{t-m}) \tag{D.1}$$

for some measurable function g. Since $m < \infty$, and $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{W}_t\}$ are mixing of the same size according to Assumption 1, it follows from Theorem 3.49 in White (2001) that $\{\boldsymbol{d}_{m,t+1}\boldsymbol{d}'_{m,t+1}\}$ is mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{W}_t\}$.

By Assumption 2 there exists $\bar{C} \in \mathbb{R}_+$ and $\delta > 0$ such that $\mathbb{E}[|\boldsymbol{d}_{m,t+1,i}|^{2(r+\delta)}] < \bar{C} < \infty$ for $i = 1, \ldots, qk$ and for all t, where subscript i indicates the i'th element in $\boldsymbol{d}_{m,t+1}$. Hence, by the Cauchy-Schwartz inequality, we obtain

$$\mathbb{E}[|\boldsymbol{d}_{m,t+1,i}\boldsymbol{d}_{m,t+1,j}|^{r+\delta}] \leq \mathbb{E}[|\boldsymbol{d}^2_{m,t+1,i}|^{r+\delta}]^{1/2}\mathbb{E}[|\boldsymbol{d}^2_{m,t+1,j}|^{r+\delta}]^{1/2} < \bar{C} \tag{D.2}$$

for $i, j = 1, \ldots, qk$ and for all t. By Corollary 3.48 in White (2001), it then follows that $\hat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. Furthermore, by Assumption 2 it follows that $\boldsymbol{\Sigma}_T$ is finite and by Assumption 3 it is uniformly positive definite.

Next, let $\boldsymbol{\lambda} \in \mathbb{R}^{qk}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and consider

$$\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_{m,t+1} = T^{-1/2}\sum_{t=1}^{T-1}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+1}. \tag{D.3}$$

Let $\tilde{\boldsymbol{\lambda}}_i$ denote the i'th element of the product $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}$, such that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+1} = \sum_{i=1}^{qk}\tilde{\boldsymbol{\lambda}}_i\boldsymbol{d}_{m,t+1,i}$. Hence, under the null hypothesis

$$\mathbb{E}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+1}|\mathscr{G}_t] = \mathbb{E}\left[\sum_{i=1}^{qk}\tilde{\boldsymbol{\lambda}}_i\boldsymbol{d}_{m,t+1,i}|\mathscr{G}_t\right] = \sum_{i=1}^{qk}\tilde{\boldsymbol{\lambda}}_i\mathbb{E}[\boldsymbol{d}_{m,t+1,i}|\mathscr{G}_t] = 0, \tag{D.4}$$

by measurability of $\tilde{\boldsymbol{\lambda}}_i$, such that the sequence $\{\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+1}, \mathscr{G}_t\}$ is a MDS. The asymptotic variance is

$$\begin{aligned}\sigma_d^2 &= \mathrm{Var}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_m] \\ &= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\mathrm{Var}[\sqrt{T}\bar{\boldsymbol{d}}_m]\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} \\ &= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} \\ &= 1 \end{aligned} \tag{D.5}$$

for sufficiently large T. Furthermore, since $\hat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$ it follows by the Continuous Mapping Theorem that

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}'_{m,t+1}\boldsymbol{d}_{m,t+1}\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} - \sigma_d^2 \\ &\quad = \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\hat{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} - \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} \xrightarrow{\mathbb{P}} 0.\end{aligned} \tag{D.6}$$

Lastly, we need to check that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+1}$ has absolute $2+\delta$ moment. By Minkowski's inequality and Assumption 2 we obtain

$$\begin{aligned}\mathbb{E}[|\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+1}|^{2+\delta}] &= \mathbb{E}\left[\left|\sum_{i=1}^{qk}\tilde{\boldsymbol{\lambda}}_i\boldsymbol{d}_{m,t+1,i}\right|^{2+\delta}\right] \\ &\leq \left(\sum_{i=1}^{qk}\tilde{\boldsymbol{\lambda}}_i\mathbb{E}\left[|\boldsymbol{d}_{m,t+1,i}|^{2+\delta}\right]^{1/(2+\delta)}\right)^{2+\delta} < \infty.\end{aligned} \tag{D.7}$$

Consequently, we can apply the CLT for MDS and deduce that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_m \xrightarrow{d} N(0,1)$. By the Cramér-Wold device it then follows that

$$\boldsymbol{\Sigma}^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_m \xrightarrow{d} N(0, \boldsymbol{I}_{qk}). \tag{D.8}$$

Since $\hat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$, we deduce that

$$\sqrt{T}(\hat{\boldsymbol{\Sigma}}_T^{-1/2}\bar{\boldsymbol{d}}_m)'\sqrt{T}\boldsymbol{\Sigma}_T^{-1/2}\bar{\boldsymbol{d}}_m = T\bar{\boldsymbol{d}}'_m\hat{\boldsymbol{\Sigma}}_T^{-1}\bar{\boldsymbol{d}}_m \xrightarrow{d} \chi^2_{qk}, \tag{D.9}$$

as $T \to \infty$. □

Proof of Proposition 2. Let $\boldsymbol{L}^*_{m,t+1}$ be an arbitrary permutation of the forecast losses, i.e. $\boldsymbol{L}^*_{m,t+1} = \boldsymbol{P}\boldsymbol{L}_{m,t+1}$, where $\boldsymbol{P}$ is a $(k+1)\times(k+1)$ permutation matrix and $\boldsymbol{L}_{m,t+1} = (L^1_{m,t+1},\dots,L^{k+1}_{m,t+1})'$. Define the $k\times(k+1)$ matrix $\boldsymbol{D}$ by

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix}$$

such that $\Delta\boldsymbol{L}^*_{m,t+1} = \boldsymbol{D}\boldsymbol{L}^*_{m,t+1} = \boldsymbol{D}\boldsymbol{P}\boldsymbol{L}_{m,t+1}$. In total, the number of permutations of the forecast losses at each point of time t is $(k+1)!$. Mariano and Preve (2012) show that there always exists a nonsingular matrix $\boldsymbol{B}$ of dimension $k\times k$ such that $\boldsymbol{B}\Delta\boldsymbol{L}_{m,t+1} = \Delta\boldsymbol{L}^*_{m,t+1}$. Consequently, define the $qk\times qk$ matrix $\boldsymbol{A} = (\boldsymbol{I}_q\otimes\boldsymbol{B})$, where $\boldsymbol{I}_q$ is the $q\times q$ identity matrix. By standard properties of the Kronecker product $\boldsymbol{A}$ is nonsingular, and we have that

$$\boldsymbol{d}^*_{m,t+1} = \boldsymbol{h}_t\otimes\Delta\boldsymbol{L}^*_{m,t+1} = (\boldsymbol{I}_q\boldsymbol{h}_t)\otimes(\boldsymbol{B}\Delta\boldsymbol{L}_{m,t+1}) = (\boldsymbol{I}_q\otimes\boldsymbol{B})(\boldsymbol{h}_t\otimes\Delta\boldsymbol{L}_{m,t+1}) = \boldsymbol{A}\boldsymbol{d}_{m,t+1}. \tag{D.10}$$

Since the null hypothesis implies that the asymptotic variance can be estimated consistently by the sample variance, it follows that

$$\hat{\boldsymbol{\Sigma}}^*_T \equiv \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{d}^*_{m,t+1}\boldsymbol{d}^{*'}_{m,t+1} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}\boldsymbol{d}_{m,t+1}\boldsymbol{d}'_{m,t+1}\boldsymbol{A}' = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_T\boldsymbol{A}'.$$

Due to the nonsingularity of $\boldsymbol{A}$ and $\hat{\boldsymbol{\Sigma}}_T$, it follows that

$$\begin{aligned}\bar{\boldsymbol{d}}^{*'}_{m,t+1}(\hat{\boldsymbol{\Sigma}}^*_T)^{-1}\bar{\boldsymbol{d}}^*_{m,t+1} &= \boldsymbol{d}'_{m,t+1}\boldsymbol{A}'(\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_T\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{d}_{m,t+1} \\ &= \boldsymbol{d}'_{m,t+1}\hat{\boldsymbol{\Sigma}}_T^{-1}\boldsymbol{d}_{m,t+1},\end{aligned}$$

which shows that the test is invariant to a permutation of the ordering of the forecast losses. □

Proof of Theorem 3. By the same arguments as in the proof for Theorem 1, it follows that the sequence $\{\boldsymbol{d}_{m,t+1}\}$ is mixing of the same size as $\{\boldsymbol{W}_t\}$ and $\{\boldsymbol{h}_t\}$.

Furthermore, Assumption 2 ensures that each element of $\boldsymbol{d}_{m,t+1}$ is bounded uniformly in t, such that

$$\bar{\boldsymbol{d}}_m - \mathbb{E}[\bar{\boldsymbol{d}}_m] \xrightarrow{\mathbb{P}} 0 \tag{D.11}$$

by Corollary 3.48 in White (2001). Under the alternative hypothesis there exists $\eta > 0$ such that $\mathbb{E}[\bar{\boldsymbol{d}}'_m]\mathbb{E}[\bar{\boldsymbol{d}}_m] > 2\eta$ for T sufficiently large. It follows that

$$\begin{aligned}\mathbb{P}[\bar{\boldsymbol{d}}'_m\bar{\boldsymbol{d}}_m > \eta] &\geq \mathbb{P}[\bar{\boldsymbol{d}}'_m\bar{\boldsymbol{d}}_m - \mathbb{E}[\bar{\boldsymbol{d}}'_m]\mathbb{E}[\bar{\boldsymbol{d}}_m] > -\eta] \\ &\geq \mathbb{P}[|\bar{\boldsymbol{d}}'_m\bar{\boldsymbol{d}}_m - \mathbb{E}[\bar{\boldsymbol{d}}'_m]\mathbb{E}[\bar{\boldsymbol{d}}_m]| < \eta] \to 1,\end{aligned} \tag{D.12}$$

where the convergence to unity is due to (D.11). By identical arguments as the proof of Theorem 1, $\boldsymbol{d}'_{m,t+1}\boldsymbol{d}_{m,t+1}$ is mixing with the same size as $\{\boldsymbol{W}_t\}$ and each element is uniformly bounded in t. Corollary 3.48 in White (2001) can then be applied, and it follows that $\hat{\boldsymbol{\Sigma}}_T$ is a consistent estimator of $\boldsymbol{\Sigma}_T$. By Assumption 3, $\boldsymbol{\Sigma}_T$ is uniformly positive definite. Let $c \in \mathbb{R}_+$. It then follows from Theorem 8.13 in White (1994) that

$$\mathbb{P}[S_{m,h} > c] \to 1, \quad \text{as } T \to \infty. \tag{D.13}$$

□

Proof of Theorem 4. **i)** We proceed by a similar procedure as in the proof of Theorem 1, however with modifications due to the dependency in $\boldsymbol{d}_{m,t+\tau}$ under the null hypothesis. First, by Assumptions 2* and 3*, $\boldsymbol{\Sigma}_T$ is finite and uniformly positive definite. Let $\boldsymbol{\lambda} \in \mathbb{R}^{qk}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and consider $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_m = T^{-1/2}\sum_{t=1}^{T}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+\tau}$. Since the null hypothesis imposes $\mathbb{E}[\boldsymbol{d}_{m,t+\tau}|\mathcal{G}_t] = \boldsymbol{0}$, identical arguments as in Theorem 1 imply that $\{\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+\tau}\}$ being mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{W}_t\}$. Moreover, the asymptotic variance satisfies $\sigma_d^2 = \text{Var}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_m] = \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} = 1$ for all T sufficiently large. Via Minkowski's inequality and computations as in (D.7), $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{m,t+\tau}$ has absolute $2+\delta$ moment for some $\delta > 0$. Then, by Corollary 3.1 in Wooldridge and White (1988) we deduce that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_m \xrightarrow{d} N(0,1)$. Hence, by the Cramér-Wold device it follows that $\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_m \xrightarrow{d} N(0,\boldsymbol{I}_{qk})$.

It remains to be shown that $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. Consider

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T = &\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau} - \mathbb{E}[\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau}]\right) \\ &+ \frac{1}{T}\sum_{j=1}^{\tau-1}\kappa(j,\tau)\sum_{t=1+j}^{T}\left(\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-j} - \mathbb{E}[\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-j}]\right. \\ &\left.+ \boldsymbol{d}_{m,t+\tau-j}\boldsymbol{d}'_{m,t+\tau} - \mathbb{E}[\boldsymbol{d}_{m,t+\tau-j}\boldsymbol{d}'_{m,t+\tau}]\right). \qquad \text{(D.14)}\end{aligned}$$

By Theorem 3.49 in White (2001), $\{\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-j}\}$ is mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{W}_t\}$ for each $j = 0,\dots,\tau-1$. Moreover, each of its elements are bounded uniformly in t by Assumption 2*. Hence, since $\kappa(j,\tau) \to 1$ as $T \to \infty$ and $\kappa(0,\tau) = 1$ it follows via Corollary 3.48 in White (2001) that

$$\frac{1}{T}\kappa(j,\tau)\sum_{t=1+j}^{T}\left(\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-j} - \mathbb{E}[\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-j}]\right) \xrightarrow{\mathbb{P}} 0,$$

for each $j = 0,\dots,\tau-1$. Combined with equation D.14, this implies that $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. Hence, we can deduce via similar steps as in (D.9) that $S_{m,h,\tau} \xrightarrow{d} \chi^2(qk)$ as $T \to \infty$.

ii) The result follows by arguments similar to those in the proof of Theorem 3. Hence, $\{\boldsymbol{d}_{m,t+\tau}\}$ is mixing with the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{W}_t\}$ and each element in $\boldsymbol{d}_{m,t+\tau}$ is bounded uniformly in t by Assumption 2*. Then it follows by Corollary 3.48 in White (2001) that $\bar{\boldsymbol{d}}_m - \mathbb{E}[\bar{\boldsymbol{d}}_m] \xrightarrow{\mathbb{P}} 0$, and consequently similar computations as in (D.12) applies. By arguments identical to those in the proof of Theorem 4i, $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$, where $\boldsymbol{\Sigma}_T$ is positive definite by Assumption 3*. Theorem 8.13 in White (1994) then implies that under $H_{A,h}$ in (9) and for any constant $c \in \mathbb{R}_+$, $\mathbb{P}[S_{m,h,\tau} > c] \to 1$ as $T \to \infty$.

iii) Due the arguments in the proof of Proposition 2 it suffices to show that $\tilde{\boldsymbol{\Sigma}}_T* = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{A}'$, where $\boldsymbol{A} = \boldsymbol{I}_q \otimes \boldsymbol{B}$. Thus, let

$$\tilde{\boldsymbol{\Sigma}}_T(p) \equiv \frac{1}{T}\sum_{t=1+p}^{T}\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-p},$$

for $p = 0,1,2\ldots$. It then follows that

$$\tilde{\boldsymbol{\Sigma}}_T(p)^* \equiv \frac{1}{T}\sum_{t=1+p}^{T} \boldsymbol{d}^*_{m,t+\tau}\boldsymbol{d}^{*\prime}_{m,t+\tau-p} = \frac{1}{T}\sum_{t=1+p}^{T} \boldsymbol{A}\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{m,t+\tau-p}\boldsymbol{A}' = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T(p)\boldsymbol{A}'.$$

Consequently, it follows that $\tilde{\boldsymbol{\Sigma}}^*_T = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{A}'$, which completes the proof. □

Proof of Theorem 5. **i)** Due to arguments similar to those in the proof of Theorem 4i, it suffices to show consistency of the variance estimator, $\breve{\boldsymbol{\Sigma}}_T$ in (16). Since $b_T \to \infty$ as $T \to \infty$ and $b_T = o(T)$, it follows by similar arguments as in the proof of Theorem 4 that for each $j = 0,\ldots,b_T$

$$\frac{1}{T}\kappa(j,b_T)\sum_{t=1+j}^{T}\left(\Delta\boldsymbol{L}_{m,t+\tau}\Delta\boldsymbol{L}'_{m,t+\tau-j} - \mathbb{E}[\Delta\boldsymbol{L}_{m,t+\tau}\Delta\boldsymbol{L}'_{m,t+\tau-j}]\right) \xrightarrow{\mathbb{P}} 0,$$

such that $\breve{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. See also Andrews (1991). Hence, by arguments identical to those in the proof of Theorem 4i, it follows that $S^{\text{und}}_{m,h,\tau} \xrightarrow{d} \chi^2(k)$ as $T \to \infty$.

ii) Let $\boldsymbol{h}_t = 1$ such that $\boldsymbol{d}_{m,t+\tau} = \Delta\boldsymbol{L}_{m,t+\tau}$. Under Assumption 3**, it then follows from the proof of Theorem 4ii that Theorem 8.13 in White (1994) applies, proving that under $H_{A,h}$ in (9) and for any constant $c \in \mathbb{R}_+$, it holds that $\mathbb{P}[S^{\text{und}}_{m,h,\tau} > c] \to 1$ as $T \to \infty$.

iii) The result follows from identical arguments as those in the proof of Theorem 4iii using $\boldsymbol{B}$ instead of $\boldsymbol{A}$. By Proposition 2, permutation invariance of the test statistic, $S^{\text{und}}_{m,h,\tau}$ then follows. □

Proof of Proposition 6. Due to the proof of Theorem 1, it suffices to show that $\boldsymbol{\Sigma}_T - \hat{\boldsymbol{\Sigma}}^{\text{thr}}_T \xrightarrow{\mathbb{P}} 0$ or equivalently that $\hat{\boldsymbol{\Sigma}}_T - \hat{\boldsymbol{\Sigma}}^{\text{thr}}_T \xrightarrow{\mathbb{P}} 0$. Since $\lambda_{ij} = C(\sigma_{ii}\sigma_{ij}\log(qk)/T)$ for some $C > 0$ it follows that $\lambda_{ij} \to 0$ as $T \to \infty$. Consequently, by the properties of the thresholding function $p_{ij}(\cdot)$ it follows directly that $\hat{\boldsymbol{\Sigma}}_T - \hat{\boldsymbol{\Sigma}}^{\text{thr}}_T \xrightarrow{\mathbb{P}} 0$.

ii) By identical arguments as in the former proof of Proposition 6i and Theorem 3, it follows as a direct consequence of the fact that $\hat{\boldsymbol{\Sigma}}_T - \hat{\boldsymbol{\Sigma}}_T^{\text{thr}} \xrightarrow{\mathbb{P}} 0$ that, under $H_{A,h}$ in (9), $\mathbb{P}[S_{m,h}^{(1)} > c] \to 1$ as $T \to \infty$ for any constant $c \in \mathbb{R}_+$. □

Proof of Corollary 7. It suffices to show that $\hat{\boldsymbol{\Sigma}}_T - \hat{\boldsymbol{\Sigma}}_T^{\text{thr}} \xrightarrow{\mathbb{P}} 0$, which follows directly from the proof of Proposition 6i. Hence, by Proposition 2 we deduce that $S_{m,h}^{(1)*} - S_{m,h}^{(1)} \xrightarrow{\mathbb{P}} 0$. □

Proof of Proposition 8. **i**) By Proposition 6 we have that $S_{m,h}^{(1)} \xrightarrow{d} \chi^2(qk)$, hence it suffices to show that the power enhancement component $S_{m,h}^{(0)}$ satisfies $S_{m,h}^{(0)} \xrightarrow{\mathbb{P}} 0$ as $T \to \infty$ under $H_{0,h}$. This follows if Assumption 4ii can be verified for the proposed power enhancement component in (27), which is shown in Theorem 3.1 in Fan et al. (2015).

ii) By Proposition 6ii, we have under $H_{A,h}$ in (9) that $\mathbb{P}[S_{m,h}^{(1)} > c] \to 1]$ as $T \to \infty$ for any $c \in \mathbb{R}_+$, hence by non-negativity of the power-enhancement component in (27), the result follows. □

Research Papers 2016

CREATES
Center for Research in Econometric Analysis of Time Series

2017-02: Giuseppe Cavaliere, Morten Ørregaard Nielsen and Robert Taylor: Quasi-Maximum Likelihood Estimation and Bootstrap Inference in Fractional Time Series Models with Heteroskedasticity of Unknown Form

2017-03: Peter Exterkate and Oskar Knapik: A regime-switching stochastic volatility model for forecasting electricity prices

2017-04: Timo Teräsvirta: Sir Clive Granger's contributions to nonlinear time series and econometrics

2017-05: Matthew T. Holt and Timo Teräsvirta: Global Hemispheric Temperatures and Co–Shifting: A Vector Shifting–Mean Autoregressive Analysis

2017-06: Tobias Basse, Robinson Kruse and Christoph Wegener: The Walking Debt Crisis

2017-07: Oskar Knapik: Modeling and forecasting electricity price jumps in the Nord Pool power market

2017-08: Malene Kallestrup-Lamb and Carsten P.T. Rosenskjold: Insight into the Female Longevity Puzzle: Using Register Data to Analyse Mortality and Cause of Death Behaviour Across Socio-economic Groups

2017-09: Thomas Quistgaard Pedersen and Erik Christian Montes Schütte: Testing for Explosive Bubbles in the Presence of Autocorrelated Innovations

2017-10: Jeroen V.K. Rombouts, Lars Stentoft and Francesco Violante: Dynamics of Variance Risk Premia, Investors' Sentiment and Return Predictability

2017-11: Søren Johansen and Morten Nyboe Tabor: Cointegration between trends and their estimators in state space models and CVAR models

2017-12: Lukasz Gatarek and Søren Johansen: The role of cointegration for optimal hedging with heteroscedastic error term

2017-13: Niels S. Grønborg, Asger Lunde, Allan Timmermann and Russ Wermers: Picking Funds with Confidence

2017-14: Martin M. Andreasen and Anders Kronborg: The Extended Perturbation Method: New Insights on the New Keynesian Model

2017-15: Andrea Barletta, Paolo Santucci de Magistris and Francesco Violante: A Non-Structural Investigation of VIX Risk Neutral Density

2017-16: Davide Delle Monache, Stefano Grassi and Paolo Santucci de Magistris: Does the ARFIMA really shift?

2017-17: Massimo Franchi and Søren Johansen: Improved inference on cointegrating vectors in the presence of a near unit root using adjusted quantiles

2017-18: Matias D. Cattaneo, Michael Jansson and Kenichi Nagasawa: Bootstrap-Based Inference for Cube Root Consistent Estimators

2017-19: Daniel Borup and Martin Thyrsgaard: Statistical tests for equal predictive ability across multiple forecasting methods