Forecast Evaluation and Selection in Unstable Environments*

Stefan Richter
Heidelberg University

Ekaterina Smetanina
University of Chicago

November 3, 2020

Abstract

Out-of-sample tests are widely used for evaluating and selecting between models' forecasts in economics and finance. Although widely used, underlying these tests is often the assumption of constant relative performance between competing models which is invalid for many practical applications. We propose a new two-step methodology designed specifically for forecast evaluation and selection in a world of changing relative performance. In the first step we estimate the time-varying mean and variance of the series for forecast loss differences, and in the second step we use these estimates to compare and rank models in a changing world. We show that our tests have high power against a variety of fixed and local alternatives.

Keywords: forecasting, unstable environments, locally stationary processes
J.E.L Codes: C22, C52, C53

1 Introduction

In a field such as economics where the majority of data available is non-experimental, an important way to judge competing models is by comparing their forecasting performance. Out-of-sample forecast evaluation tests, which are broadly variations of the Diebold-Mariano test from their 1995

*Corresponding author: Ekaterina Smetanina. Email: e.smetanina@chicagobooth.edu. We would like to thank Tim Bollerslev, River Chen, Frank Diebold, Raffaella Giacomini, Bruce Hansen, Dennis Kristensen, Oliver Linton, Jason Lu, Alexei Onatski, Andrew Patton, Rasmus Søndergaard Pedersen, Anders Rahbek, Barbara Rossi, George Tauchen, Steve Thiele, Wei Biao Wu and seminar participants at various seminars this paper has been presented. The second author kindly acknowledges the support of Asness Junior Faculty Fellowship at the University of Chicago Booth School of Business. The `R code` for implementing the methodology in the paper can be found here.

seminal paper, are currently the benchmark for comparing models' forecasting performance. It is known however that in an environment where relative performance between models can change over time, these tests can generate spurious and potentially misleading results. An example of this is the well-known problem of choosing a sample splitting point. The splitting point is used in Diebold-Mariano-type tests to split the sample into the first part, data used for estimation, versus the second part, data used for evaluation. The commonly adopted approach advocates a late sample splitting point which leaves relatively little data for evaluation and consequently leads to these tests having low power.[1] Simultaneously, there is the opportunity for practitioners to search for a favorable splitting point that supports their desired outcome, see e.g. Rossi and Inoue (2012) and Hansen and Timmermann (2012). These issues become more troublesome in a world of changing relative performance. In such a setting, one model may outperform its competition for some *window* of data, but under-perform for a different window. As the splitting point controls the division of the data into estimation and evaluation, the choice of the splitting point controls implicitly the window of evaluation, and manipulating the splitting point can allow one to influence the outcome of the test.[2]

To present an example that demonstrates both the potential issue of low power and the arbitrary dependence of the test conclusion on the splitting point, consider the following real-world example. We forecast the daily variance of IBM returns spanning 2006-2017 using two models: GARCH(1,1) model with Standard normal errors and GARCH(1,1) model with Student-t errors. Variance forecasts are produced via a standard recursive scheme, 5 minute realized variance calculated from the data is used as a proxy for the "true" variance, and mean squared errors are calculated by averaging squared errors after a particular splitting point. We present the difference in the mean squared errors, ΔMSE_t, and the associated critical values[3] at 5% significance level across a range of splitting point choices. The out-of-sample data starts in December 2010, leaving at most 1500 data points for evaluation. Each slice of the graph therefore represents the result of a Diebold-Mariano test at that particular splitting point. In this example, for many splitting point choices the test is not powerful enough to distinguish between the two models. For other choices of the splitting point we obtain a rejection in one direction, and for yet other choices we obtain a rejection in the opposite direction. Hence, depending on the choice of splitting point all possible conclusions of the test are possible.

[1] See Diebold (2015) for a discussion on this issue and a more recent study by Hirano and Wright (2017) that concludes that current out-of-sample tests perform poorly due to large estimation errors.

[2] Despite these drawbacks, out-of-sample tests are still often preferred to their alternative, in-sample tests, which uses all available data for both estimation and evaluation. See Hansen (2010) for an analysis the tendency for in-sample tests to select models that overfit and underperform in forecasting.

[3] We use HAC variance estimator for the calculation of the test statistic.

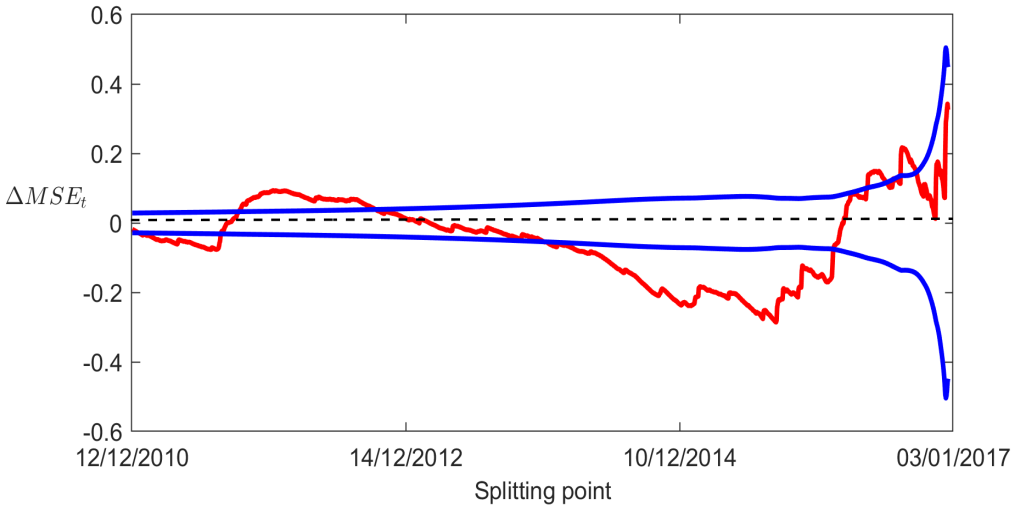

Figure 1: The figure displays the *difference* in MSE calculated for GARCH(1,1)-N and GARCH(1,1)-St-t for IBM data, 2006-2017 using recursive forecasting scheme. The MSE for each of the models is taken with respect to 5min RV calculated from the data. For a given splitting point, the red line represents the value of the Diebold-Mariano test, while the blue lines represent the associated critical values at $\alpha = 0.05$ significance level.

Our previous example demonstrates the potential perils of using tests designed for a world of constant relative performance when we are in a changing world. Giacomini and Rossi (2010) were the first to systematically study the issue of changing relative performance in the context of out-of-sample forecasting. They propose two tests in this unstable environment, the first being a "one-time reversal test" which considers a null of equal performance against a joint alternative that one of the two models was always better or that there was a one-time reversal, i.e. single permanent reversal in the relative performance of the models. They also propose a "fluctuation test" of which the null is equal relative performance between two models at all points in time. This latter test is in contrast to the first test, and more broadly the literature on structural breaks, as they instead allow for smoothly changing relative performance where models can overtake each other many times. This smooth type of instability presents its own challenges, for instance one is required to compose an estimate for the *local* relative forecasting performance between models. The above two tests can detect a wide range of instabilities in relative performance. This of course poses a natural question of what a practitioner should do once instabilities are detected. The authors state in their article that they "do not investigate this issue in depth, although possible strategies can be devised if one is willing to specify the nature of the instability". Importantly, if one detects instabilities, the notions of forecast evaluation and forecast selection do not overlap as past performance may not indicate future performance, however it is nonetheless important that practitioners have guidance

on how to evaluate and select models in such cases.

Regarding the topic of forecast selection, in the context of conditional predictive ability Giacomini and White (2006) offer a heuristic approach for forecast selection by using their model to forecast the conditional mean of forecast loss differences and comparing it with a user-specified threshold. They apply this procedure to select the best method among several parameter-reduction methods in the context of forecasting macroeconomic indicators. The authors acknowledge that their method is not formally developed since forecast selection is not a primary focus of their paper and further emphasize that "the subject of forecast selection is a significant topic that deserves extensive attention...".

In this paper, we develop new forecast evaluation and selection methodology for a world of smoothly changing relative performance. Our forecast evaluation test compares the average historical performance of models to test for *overall* equal predictive ability. This is in contrast to the fluctuation test null in Giacomini and Rossi (2010) of equal predictive ability at all points in time. Comparing with existing out-of-sample tests, our metric for overall equal predictive ability is specifically designed for a world of changing relative performance, and we incorporate the whole sample of forecast losses hence we do not anchor our test on a particular sample splitting point. For the purpose of forecast selection, we rank models based on which model we expect to outperform in the next period. We do this by constructing forecasted probabilities of how likely the forecast loss of one model will be smaller than the forecast loss of another model, which a practitioner may use to select a model for forecasting next period.

Our overall methodology is summarized by a two-step procedure. In the first step, we nonparametrically estimate the time-varying mean and variance for the series of forecast loss differences. In the second step, we utilize these estimates to compare and rank competing models using our two proposed approaches. For the first approach, we define a test statistic which aggregates the time-varying normalized losses across the entire sample. One therefore can interpret the new test as an aggregated t-test across the whole sample, which is reminiscent to the weighted least squares idea in the standard regression framework. In the second approach, we estimate the probability of how likely one model will outperform the other. In addition, we construct *forecast intervals*, which measure the confidence interval of the forecasted probability. In general, our two approaches can coincide in their conclusions, and they overlap in the special case of constant relative performance. However in some applications model A that performed on average equal to model B over the overall sample may outperform towards the end of the sample. In such a situation, our first approach will indicate historical equal predictive ability, however our second approach will propose to select model A for future forecasts. The second approach is more relevant for the purpose of ranking

models for future forecasts, but because it is only concerned with the next period performance the resulting ranking is noisier and contingent always on the end point of a particular sample. On the other hand, a practitioner may be interested in which model provided a *better explanation* to the data, from the point of view of out-of-sample performance, and here our first approach will be more appropriate to address the question. We believe both approaches are insightful for different situations and we leave it to the practitioner to select the appropriate methodology for their application.

Related to this work is topic studied more broadly by Giacomini and White (2006), who develop a conditional version of the unconditional equal predictive ability test of Diebold and Mariano (1995). Acknowledging the possible dependence of relative performance on the information set at a given point in time, Giacomini and White (2006) condition their test on a set of covariates. This therefore enables one to use their test to detect the possible variation of relative performance over time. For example, their test rejects their null when models' relative performance depends on a "state of the world" variable, even if the unconditional relative performance is equal. In this case the dependence on the state of the world variable leads to variation in relative performance over time, and we may use their test as a check for whether we are in a changing world or a constant world, where a rejection of their test is indicative of changing relative performance.

In addition there are the papers by Rossi and Inoue (2012) and Hansen and Timmermann (2012). They look to tackle the splitting point problem in a constant relative performance context by accounting for the potential for data mining of practitioners who search for favorable splitting points. They propose to explicitly mine over all splitting points for the one that is the most favorable for the *alternative* hypothesis, and they reevaluate their test statistic at this splitting point with adjusted critical values that account for the bias introduced by mining.

The rest of the paper is organized as follows. In section 2 we further discuss the world of changing relative performance and the two approaches we propose. In section 3 we present our theoretical results. Section 4 addresses the issue of bandwidth selection for our two-step nonparametric procedure. Section 5 describes the bootstrap procedure that is used to approximate the distribution of our new statistics in applications. In section 6 we investigate the size and the power of our test under a variety of alternatives as well as the performance of the sign forecasts. We present our applications in section 7 and conclude in section 8. All proofs of the theoretical results are collected in Appendix B in the Supplementary Material.

Throughout this paper, the following notation is used. Let $f : \mathbb{R}^d \to \mathbb{R}$, then $\partial_x f := \frac{\partial f}{\partial x}$ and $\partial_x^2 f := \frac{\partial^2 f}{\partial x \partial x'}$ denote the first and the second derivatives with respect to the argument x respectively. For some vector $v \in \mathbb{R}^d$ we write v' to denote the transpose of v and define $|v|_2 := (\sum_{j=1}^d v_j^2)^{1/2}$ and

$|v|_\infty := \max_{j=1,\dots,d} |v_j|$. For some matrix $A \in \mathbb{R}^{d\times d}$ we write $\text{tr}(A)$ to denote its trace and define $|A|_\infty := \max_{i,j} |A_{ij}|$. In addition, $\otimes$ denotes the Kronecker product of matrices. Finally, $\xrightarrow{p}$ denotes the convergence in probability and $\xrightarrow{d}$ denotes convergence in distribution. All convergences are considered when the sample size $T \to \infty$.

2 Forecast evaluation and selection in unstable environments

This discussion is structured in the context of point forecasts, although the methodology is more general and can be applied in the context of probability or density forecasts. Let A, B be two models and $\{y_t\}_{t=1}^T$ be the original data. Let $\widehat{y}^A_{t+k|t}$ and $\widehat{y}^B_{t+k|t}$ denote the forecasts made at time t for y_{t+k} using models A and B respectively, which reflect the models as well as the estimation procedures. We denote the difference in forecast losses at time $t+k$ by $\Delta\mathcal{L}^{AB}_{t+k} = \mathcal{L}\left(y_{t+k}, \widehat{y}^A_{t+k|t}\right) - \mathcal{L}\left(y_{t+k}, \widehat{y}^B_{t+k|t}\right)$, where $\mathcal{L}(\cdot)$ denotes the loss function chosen by the practitioner. In what follows we shall refer to $\Delta\mathcal{L}^{AB}_{t+k}$ as $\Delta\mathcal{L}_{t+k}$ for simplicity of notation. We take forecasts as given and therefore in what follows treat loss differences as primitives.

We consider a world of changing relative performance, hence we also allow the mean and variance of $\Delta\mathcal{L}_{t+k}$ to be time-varying. In particular, we define $\mu_{t+k} = \mathbb{E}\left[\Delta\mathcal{L}_{t+k}|\mathbb{X}_t\right]$, where $\mathbb{X}_t$ denotes a set of conditioning regressors. This therefore makes μ_{t+k} a conditional mean of the loss difference at time $t+k$. Since our test is a weighted average of μ_{t+k} it makes it therefore a conditional test, albeit in a different sense than in Giacomini and White (2006). Specifically, for our methodology we consider $\mathbb{X}_t$ to be lags of $\Delta\mathcal{L}_t$. In general, our theory can be extended to allow for general $\mathbb{X}_t$ so long as its functional dependence (see Wu (2005)) decays geometrically fast with further modification to the proofs. We discuss our motivation to choose lags of $\Delta\mathcal{L}_t$ as our conditioning variable in the next section when we introduce our model.

In a world of *constant relative* forecasting performance, i.e. $\mu_{t+k} = \mu$, for all $\mathbb{X}_t$ and all $t \in \{1, \cdots, T\}$, the tasks of evaluating versus selecting models overlap. Supposing that we knew μ with certainty, if $\mu = 0$ we say the models have equal predictive ability, and if $\mu < 0$ model A performs better than model B both in the past and in the future. Here, evaluating which model performed better in the past informs directly which model shall perform better in the future and the tasks of evaluation and selection are therefore the same. Furthermore, in such a world the conclusion of the standard out-of-sample tests does not depend on the evaluation window and neither on the choice of the splitting point, although with a too short evaluation window the test shall suffer from low power. Studies that assume constant relative performance include Diebold and Mariano (1995), West (1996), White (2000),McCracken (2000), Clark and McCracken (2001),

Clark and McCracken (2005), Inoue and Kilian (2005), Hansen (2005), Corradi and Swanson (2007), Hansen et al. (2011), Rossi and Inoue (2012), Hansen and Timmermann (2012) and Li and Patton (2018), among others.

In an unstable environment[4], the tasks of forecast evaluation and selection no longer overlap, and a model that performed better in the past may not perform better in the future. Simultaneously, it is often not the case that one model will dominate another model across all points in time, hence it is not immediately obvious how one should compare historical performance between models whose relative forecasting performance might have switched signs more than once over a historical sample. To address these questions we require approaches specifically designed for such unstable environments.

Giacomini and Rossi (2010) propose the fluctuation test to investigate whether models' relative forecasting performance has been stable over time by using a measure of *local* relative performance. In particular, they propose to measure local relative performance between two models by calculating the mean of loss differences over rolling windows of (fixed) size m and then testing the null hypothesis that this local measure is equal to zero at all points in time. In designing their test this way, they allow for time-variation in the mean relative performance between models. A rejection of their fluctuation test is then indicative of either changing relative performance, or non-zero constant relative performance.

In addition, Giacomini and White (2006) consider testing the following conditional moment condition: $\mathbb{E}\left[\Delta\mathcal{L}_{t+1}|\mathcal{F}_t\right] = 0$, where $\mathcal{F}_t$ is the information set available to forecaster at time t. Provided that $\{\Delta\mathcal{L}_t, \mathcal{F}_t\}$ is a martingale difference sequence, we may test the more lenient in-sample moment condition[5]: $\mathbb{H}_0 : \mathbb{E}\left[\Delta\mathcal{L}_{t+1}h_t\right] = 0$, such that $h_t \in \mathcal{F}_t$. The authors recommend to set $h_t = (1, \Delta\mathcal{L}_t)'$. With such a specification, in a regression framework this translates to the following:

$$\Delta\mathcal{L}_t = \alpha + \beta\Delta\mathcal{L}_{t-1} + \varepsilon_t, \quad \text{and} \quad \mathbb{H}_0 : \alpha = 0 \quad \cap \quad \beta = 0.$$

That is to say, their test evaluates the null of the forecast losses having zero mean and an absence of serial correlation at first lag. If we reject their null, this could be due to a dependence of the above moment condition on the first lag, or due to constant relative performance but with $\mu_t = \mu \neq 0$ for all t. Specifically, serial correlation at the first lag is indicative of changing relative performance, as it is no longer the case that $\mu_t = \mu$ for all t.

The purpose of this paper is to provide a methodology for forecast evaluation *and* forecast selection designed specifically for unstable environments. To address forecast evaluation, we compare

[4]This can occur even when the data generating process is stationary, see an example presented in Appendix A1.

[5]Meaning that rejection of the null $\mathbb{H}_0 : \mathbb{E}\left[\Delta\mathcal{L}_{t+1}h_t\right] = 0$ leads to rejection of $\mathbb{H}_0 : \mathbb{E}\left[\Delta\mathcal{L}_{t+1}|\mathcal{F}_t\right] = 0$.

past performance and test for *overall* equal predictive ability over the whole sample. Specifically our notion of overall equal predictive ability does not require one model to always outperform another. Instead we incorporate historical relative performance between models into a summary metric that looks to answer the question of which model did better overall in the given sample. Our first innovation is to use nearly the entire series of forecast losses to construct our statistic, which extends the evaluation window to nearly the whole sample. By using the near-whole sample we eliminate the choice of the splitting point. Furthermore, by using more data we dramatically improve the power of our tests. However, to do this one needs to account for the variance of the loss differences, as the losses constructed earlier in the sample are likely more noisy than those constructed at the end due to decreasing estimation error. We therefore propose to calculate the average *overall* forecasting performance as follows:

$$\sum_{t=1}^{T-k} w_{t+k}\mu_{t+k}, \tag{1}$$

where w_{t+k} are weights that we select to be inversely proportional to the standard deviation of forecast losses. With our weighting, the forecast losses at the beginning of the sample, which usually tend to come with the largest estimation error, will be naturally down weighted. Moving towards the end of the sample, losses tend to naturally receive a larger weight. In what follows, the time-varying standard deviation of loss differences is nonparametrically estimated from the data and therefore the corresponding weights are data-driven.

With regards to forecast selection in the context of unstable environments, we propose to directly forecast the probability that one model shall outperform in the next immediate period, i.e. the probability that the sign of the next period loss difference is negative. We choose to forecast the sign as opposed to the level because levels can depend on arbitrary factors such as a factor of scaling to the loss function, and furthermore it is not clear what kind of a difference in levels constitutes a significant deviation (see Giacomini and White (2006) for a simple application of their framework to level forecasting). Meanwhile, the sign of the loss difference reflects a binary comparison, and the sign for a particular comparison is stable across all symmetric loss functions. In addition to forecasting the probability of outperforming, we also construct what we call *forecast intervals* which measure the $(1-\alpha)\%$ confidence interval of the forecasted probability, where α is a chosen significance level. Specifically, we predict a model to strictly out-perform another when the forecasted probability is greater than 0.5 and the forecast interval does not contain 0.5. Importantly, we select models by forecasting truly out of the sample, which is in contrast to comparing past performance using pseudo out-of-sample methods in the existing literature.

We construct losses following the standard recursive scheme, however in contrast to the existing out-of-sample tests, we need to construct losses for nearly the entire sample and not just a short evaluation window towards the end of sample. We describe our loss construction below.

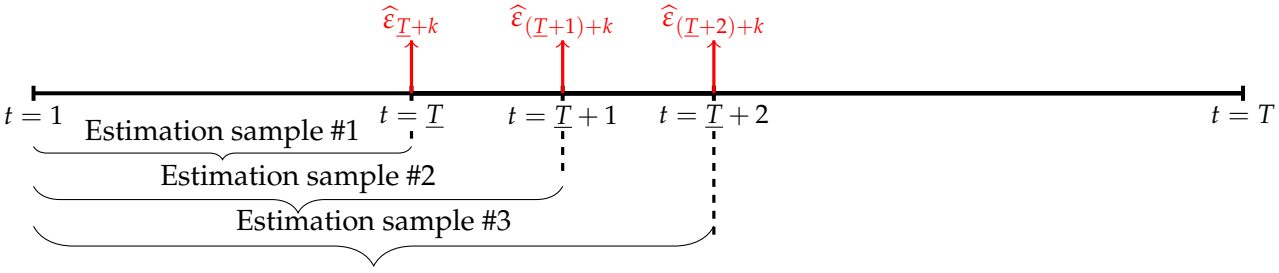

Figure 2: Construction of the time series of the forecast errors for a single model.

The pseudo-out-of-sample forecast made at time t for period $t+k$ is compared with the realized value in period $t+k$, which when differenced gives the forecast error of period t. The loss function is then applied to this error which gives the forecast loss of period $t+k$. The recursive scheme calculates the forecast loss using parameter estimates based on all data up until time t. It is recursive because with each new period the model is re-estimated to include the new data. We use all of the forecast losses except for a small sample of length $\underline{T}$ in the beginning, which is reserved for initial estimation. We recommend that practitioner always sets $\underline{T} = 100$.

Remark 1. The only recommendation for $\underline{T}$ is that it should be in the beginning of the sample. In our simulations and applications we experimented with $\underline{T} = 150$ and $\underline{T} = 200$ vs our proposed $\underline{T} = 100$ and the conclusions of our tests did not differ. Indeed, since the estimates at the beginning of the sample tend to come with a larger estimation error, they will tend to be downweighted. It is therefore intuitive that our tests are not sensitive to this choice.

After the time series of forecast losses is constructed for each model, we may now compute the loss differences for a pair of models, A and B:

$$\Delta\mathcal{L}_{t+k} = \mathcal{L}\left(y_{t+k}, \widehat{y}^{A}_{t+k|t}\right) - \mathcal{L}\left(y_{t+k}, \widehat{y}^{B}_{t+k|t}\right),$$

where $\mathcal{L}(\cdot)$ is the chosen (by the researcher) loss function. For example, for the conventional squared error loss function it simply becomes

$$\Delta\mathcal{L}_{t+k} = \left(\widehat{\varepsilon}^{A}_{t+k}\right)^2 - \left(\widehat{\varepsilon}^{B}_{t+k}\right)^2, \quad \text{where} \quad \widehat{\varepsilon}^{i}_{t+k} = y_{t+k} - \widehat{y}_{t+k|t}, \quad i = A, B.$$

In addition, due to our use of a recursive estimation scheme, our methodology is not applica-

ble to nested models. This happens because the estimation error of parameters is asymptotically vanishing, and therefore whenever two models are nested their forecasts become asymptotically equivalent. This leads to the asymptotic variance-covariance matrix of the vector of their respective losses to become degenerate and lose full rank.

3 Theoretical results

We now present the key theoretical results for this paper. The code required for the implementation is available here. Having constructed the time series for $\Delta\mathcal{L}_t$, we now work with this time series and not the original data. For ease of notation we shall say that $\Delta\mathcal{L}_t$ ranges from $t = 1, 2, \cdots, T$, although the length of this series T is different to the length of the original data y_t. The new T is equal to the original $T - \underline{T} - k + 1$.

We propose to model $\Delta\mathcal{L}_t$ as a locally stationary AR(d) process with white noise errors which allows its mean and variance to change smoothly over time. The rationale behind such a modelling framework is as follows. It is a well-established fact that estimation error causes $\mathcal{L}_t$ to exhibit serial correlation, see e.g. Bollerslev et al. (2016). One way to account for serial correlation in $\Delta\mathcal{L}_t$ is to allow it to depend on its own lags, and an AR(d) structure on $\Delta\mathcal{L}_t$ is a simple yet tractable model that achieves that goal. In particular, it is well-known that any correlated stationary processes can be approximated by white-noise AR or ARMA models, see e.g. Pourahmadi (1992) and Brockwell and Davis (2006). Since locally stationary processes behave approximately stationary over short periods of time (i.e. locally in time) the above argument straightforwardly extends to locally stationary AR processes with white noise errors. In this paper we develop the theory for general time-varying AR(d) processes and briefly discuss how the lag selection of d can be performed in such a framework. That said, we believe that the simple AR(1) model is often sufficient for practitioners to capture features of the data while remaining practically simple to implement.

As in the literature on locally stationary processes, we make $\Delta\mathcal{L}_t$ depend on the rescaled time points t/T rather than real time t, forming therefore a triangular array, $\{\Delta\mathcal{L}_{t,T} : t = 1, \cdots, T\}$. This rescaling is necessary to justify the properties of the resulting estimation procedures with so-called infill asymptotics. The model on $\Delta\mathcal{L}_{t,T}$ then reads:

$$\Delta\mathcal{L}_{t,T} = \rho^0_{t,T} + \sum_{j=1}^{d} \rho^j_{t,T} \Delta\mathcal{L}_{t-j,T} + \xi_{t,T}, \qquad t = 1, \cdots, T, \tag{2}$$

where ξ_t has zero mean and is independent of $\Delta\mathcal{L}_{s,T}$ for $s \leq t - 1$. For mathematical reasons, we assume that the recursion (2) is also valid for the unobserved time points $t < 0$ and $t > T$. We

further take the time-varying functions $\rho_{t,T}^j$, $j = 0, \cdots, d$ to be deterministic functions of time. We use the following rescaling method. Let $\rho_j(\cdot)$ be a function on $[0,1]$ and let for $t = 1, \cdots, T$

$$\rho_{t,T}^j = \rho_j\left(t/T\right), \quad j = 0, 1, \cdots, d.$$

The notation $\rho_{t,T}^j$ indicates that $\rho_{t,T}^j$ depends on the sample size T and the domain of $\rho^j(\cdot)$ becomes more dense in t/T as $T \to \infty$. In other words, the time-varying coefficient functions $\rho^j(\cdot)$ do not depend on the real time t but rather on the rescaled time points t/T. For simplicity, we impose a deterministic structure on $\rho_{t,T}$ and $\sigma_{t,T}$ since we are primarily interested in capturing time variation in the mean and variance of $\Delta\mathcal{L}_t$. In addition, since the coefficients of the autoregression in (2) are time-varying, $\Delta\mathcal{L}_{t,T}$ is no longer stationary, but locally stationary in the sense of Dahlhaus et al. (2019). In addition, we assume that the error process $\{\xi_{t,T}, t = 1, \cdots, T\}$ has the following structure:

$$\xi_{t,T} = \sigma(t/T)\varepsilon_t, \tag{3}$$

where ε_t is an i.i.d. process having the property that ε_t is independent of $\Delta\mathcal{L}_{s,T}$ for $s \leq t-1$.

Remark 2. Note that our model is for forecast loss differences as a whole, which allows for serial correlation in forecast errors due to, for example, estimation error. The error term ε_t corresponds to the innovation to the relative forecasting performance at time t, which we model as white noise based on our rationale from before. Nonetheless, the theory developed in this paper can be extended to non-i.i.d. error ε_t. In fact, it is not hard to see in the proofs that we can allow for errors of the form:

$$\varepsilon_t = \mathcal{G}(\eta_t, \eta_{t-1}, ...),$$

where $\mathcal{G}(\cdot)$ is a well-defined function, η_t is i.i.d. and ε_t has mean 0 and is an uncorrelated sequence. Furthermore, one has to impose a condition that both the β-mixing coefficients (Doukhan et al. (1995)) and the functional dependence measure (Wu (2005)) decay geometrically fast. In principle, it is also possible to extend the theory to polynomially decaying dependence coefficients but then some additional proofs are required.

For the ease of exposition we compactly write the model in (2) as follows:

$$\Delta\mathcal{L}_{t,T} = \mathbb{X}_{t,T}'\rho\left(t/T\right) + \sigma\left(t/T\right)\varepsilon_t, \tag{4}$$

where $\rho(t/T) = (\rho_0(t/T), \rho_1(t/T), \cdots, \rho_d(t/T))'$ and $\mathbb{X}_{t,T} = (1, \Delta\mathcal{L}_{t-1,T}, \Delta\mathcal{L}_{t-2,T}, \cdots, \Delta\mathcal{L}_{t-d,T})'$. Finally we assume that $\rho(u) = \rho(0)$ and $\sigma(u) = \sigma(0)$ for $u \leq 0$, while $\rho(u) = \rho(1)$ and $\sigma(u) = \sigma(1)$ for $u \geq 1$. In what follows, we estimate the time-varying coefficient function $\rho(t/T)$ and time-

varying volatility function $\sigma(t/T)$ by nonparametric kernel techniques. In particular, using the notation $K_h(\cdot) = K(\cdot/h)/h$ for the kernel function, we estimate model (4) in the following way:

Step 1: First estimate the mean function via the local linear nonparametric estimator. In particular, define the following locally weighted least-squares objective:

$$\widehat{\theta}(u) = (\hat{\rho}(u)', \widehat{\partial_u \rho}(u)')' = \underset{\theta \in \mathbb{R}^{2(d+1)}}{\arg\min} \sum_{t=1}^{T} K_{h_1}(t/T - u) \left(\Delta \mathcal{L}_{t,T} - (\mathbb{Z}_t^u)'\theta\right)^2, \tag{5}$$

where $\mathbb{Z}_t^u = \mathbb{Z}_{t,T}^u = (\mathbb{X}_{t,T}, \mathbb{X}_{t,T}(t/T - u))'$. The explicit solution to (5) is given by

$$\widehat{\theta}(u) = \hat{\Sigma}_T(u)^{-1} \cdot \frac{1}{T} \sum_{t=1}^{T} K_{h_1}(\frac{t}{T} - u) \cdot \mathbb{Z}_t^u \cdot \Delta \mathcal{L}_t, \qquad \hat{\Sigma}_T(u) := \frac{1}{T} \sum_{t=1}^{T} K_{h_1}(\frac{t}{T} - u) \cdot \mathbb{Z}_t^u (\mathbb{Z}_t^u)', \tag{6}$$

Step 2: Define estimated residuals $\widehat{\xi}_t(u) = \Delta \mathcal{L}_{t,T} - \mathbb{X}_{t,T}' \widehat{\rho}(u)$ which serves as an estimator of ξ_t as long as $\frac{t}{T} \approx u$. Estimate the conditional variance and its derivative $\varsigma(u) = (\sigma^2(u), \partial_u(\sigma^2(u)))'$ by running a local linear nonparametric regression of $\widehat{\xi}_{t,T}(u)$ on the rescaled time $\frac{t}{T}$. Define $\mathbb{F}_t^u := (1, \frac{t}{T} - u)'$, and

$$\hat{\varsigma}(u) := (\widehat{\sigma^2}(u), \widehat{\partial_u \sigma^2}(u))' = \underset{\varsigma \in (0,\infty) \times \mathbb{R}}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} K_{h_2}(t/T - u) \cdot \{\hat{\xi}_t(u)^2 - (\mathbb{F}_t^u)'\varsigma\}^2. \tag{7}$$

The explicit solution of (7) is given by

$$\hat{\varsigma}(u) = \hat{\kappa}_T(u)^{-1} \hat{z}_T(u), \tag{8}$$

where

$$\hat{\kappa}_T(u) := \frac{1}{T} \sum_{t=1}^{T} K_{h_2}(\frac{t}{T} - u) \cdot \mathbb{F}_t^u (\mathbb{F}_t^u)', \qquad \hat{z}_T(u) := \frac{1}{T} \sum_{t=1}^{T} K_{h_2}(\frac{t}{T} - u) \cdot \mathbb{F}_t^u \hat{\xi}_t(u)^2.$$

Remark 3. In Lemma 10 in Supplementary Material it is shown that $\widehat{\sigma^2}(u) \xrightarrow{p} \sigma^2(u)$, which shows that for Th_2 large enough and h_2 small enough, $\widehat{\sigma^2}(u)$ is positive. However, for finite samples it may occur in rare cases that $\widehat{\sigma^2}(u) < 0$. This issue can be dealt with as follows: fix some small $\sigma_0 > 0$. If $\widehat{\sigma^2}(u) < \sigma_0^2$, then replace $\widehat{\sigma^2}(u)$ by σ_0^2. In practice, one may simply choose σ_0^2 as 10^{-3}.

The corresponding estimator of the standard error can then be constructed by talking the square root of $\widehat{\sigma_2}(u)$. In other words, we take $\widehat{\sigma}(u) = \{\widehat{\sigma^2}(u)\}^{1/2}$. Furthermore, given a band-

width h, one may further choose the number of lags d based on the Akaike information criterion as follows: let $\widehat{\sigma^2}(u)_d$ denote the estimator based on d lags, then choose

$$\hat{d} = \underset{d \in \mathbb{N}}{\arg\min}\, \mathrm{AIC}(d), \qquad \mathrm{AIC}(d) := \sum_{t=1}^{T} \{ \log(\widehat{\sigma^2}(\frac{t}{T})_d) + 1 \} + 2(d+1).$$

Next, to guarantee formal existence of the theoretical model (2), we impose the following standard assumptions for time-varying autoregressive processes.

Assumption 1. *(i) Let $\sigma_{min} > 0$. The functions $\rho = (\rho_0, ..., \rho_d) : [0,1] \to \mathbb{R}^{d+1}$ and $\sigma : [0,1] \to [\sigma_{min}, \infty)$ are three times continuously differentiable.*

(ii) The characteristic polynomial fulfills $\vartheta(u,z) = 1 - \rho_1(u)z - \cdots - \rho_d(u)z^d \neq 0$ for all $u \in [0,1]$ and all $0 < |z| \leq 1 + \delta$ for some $\delta > 0$.

Assumption 1 guarantees the existence of a moving average representation $\Delta\mathcal{L}_{t,T} = \sum_{j=0}^{\infty} a_{t,T}(j)\varepsilon_{t-j}$ of the recursively defined process (2), c.f. Dahlhaus and Polonik (2009) (Proposition 2.4 therein), where $a_{t,T}(j)$ are real numbers satisfying $|a_{t,T}(j)| \leq C\rho^j$ with some $C > 0$, $\rho \in (0,1)$. Furthermore, they allow one to introduce a family of stationary processes $\Delta\mathcal{L}_t(u)$ which approximate $\Delta\mathcal{L}_{t,T}$ in a suitable way for $\frac{t}{T} \approx u$ and arise in the limit distributions of our theorems. For each $u \in (0,1)$, the process $\Delta\mathcal{L}_t(u)$ follows the recursion:

$$\Delta\mathcal{L}_t(u) = \mathbb{X}_t(u)'\rho(u) + \sigma(u)\varepsilon_t, \qquad t \in \mathbb{Z}, \tag{9}$$

where $\mathbb{X}_t(u) := (1, \Delta\mathcal{L}_{t-1}(u), ..., \Delta\mathcal{L}_{t-d}(u))'$. The process $\Delta\mathcal{L}_t(u)$ is not observed in practice and is a theoretical construct which is needed to provide the bias and variance terms in our asymptotic results. Lastly, we ask the kernel to fulfill the following smoothness assumptions.

Assumption 2. *The kernel $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is nonnegative and Lipschitz continuous, i.e. there exists some $L_K > 0$ such that for all $v_1, v_2 \in \mathbb{R}$: $|K(v_1) - K(v_2)| \leq L_K|v_1 - v_2|$, with compact support $\subset [-1,1]$. Furthermore, K is symmetric and fulfills $\int K(z)dz = 1$. We set for $j \in \{0,1,2\}$, $\lambda_j = \int z^j K(z)dz$ and $\nu_j = \int z^j K^2(z)dz$.*

Conditions on the bandwidths are stated in the main theorems. In general, we ask h_1 and h_2 to be of the same order, i.e. the ratio h_1/h_2 is bounded away from 0 and ∞ for $T \to \infty$ and fulfill $h_i \to 0$, $Th_i \to \infty$ for $i = 1,2$. Since we are using a local linear estimation approach for $\rho(\cdot)$ in (5), we are able to obtain asymptotic results not only for $\hat{\rho}$, but also for its derivative $\hat{\partial_u \rho}$. To state this

in a concise way, let us define

$$H = \begin{bmatrix} I_{d+1} & 0 \\ 0 & h_1 I_{d+1} \end{bmatrix},$$

where I_{d+1} is the identity matrix of dimension $(d+1) \times (d+1)$. The following theorem is a special case of Theorem 7 which is proven in the Appendix B in the Supplementary Material.

Theorem 1. *Let Assumptions 1, 2 hold. Fix $u \in (0,1)$. Assume that $\mathbb{E}|\varepsilon_0|^q < \infty$ for some $q > 2$ and $Th_1^5 = O(1)$, $Th_1 \to \infty$, $h_1 \to 0$. Then*

$$\sqrt{Th_1}\Big(H(\widehat{\theta}(u) - \theta(u)) - \frac{h_1^2}{2}\begin{pmatrix} \lambda_2 \partial_u^2 \rho(u) \\ 0 \end{pmatrix}\Big) \xrightarrow{d} N\Big(0,\ \sigma^2(u)\begin{pmatrix} \nu_0 & 0 \\ 0 & \lambda_2^{-2}\nu_2 \end{pmatrix} \otimes \Omega(u)^{-1}\Big),$$

where $\Omega(u) := \mathbb{E}[\mathbb{X}_0(u)\mathbb{X}_0(u)']$ and $\mathbb{X}_t(u)$ is from (9).

Given that our test statistics, which we describe in the next section, aggregates $\widehat{\mu}_t(u)$, i.e. $\hat{\rho}(u)$, $\widehat{\sigma^2}(u)$ over $u \in [0,1]$, it becomes necessary to establish the uniform convergence of $\hat{\rho}(u)$ and $\widehat{\sigma^2}(u)$ over the whole support of u rather than just establishing pointwise consistency. To proceed we have to impose the following additional Assumption.

Assumption 3. *Suppose that for some $q > 2$, $\mathbb{E}|\varepsilon_0|^{2q} < \infty$. Furthermore assume that h_1 and h_2 fulfill*

$$\frac{\log(T)^3}{T^{1-\frac{2}{q}}h_1} = o(1), \qquad \frac{\log(T)}{T^{1/2}h_1} = o(1) \tag{10}$$

and

$$h_1/h_2 \to c \in (0,\infty). \tag{11}$$

Remark 4. While the first condition (10) is needed to provide the correct rate for the uniform convergence in Theorem 2, the second condition is needed later in the proof of Theorem 3 to guarantee the neglegibility of several arising terms, especially r^2_{T,h_1} defined in (12). Condition (11) means that h_1, h_2 have the same order of magnitude (although they can be different). By (11), the conditions stated in (10) for h_1 automatically also hold for h_2. Condition (10) formally imposes lower bound on the rate with which h_1 tends to 0. If $h_1 = cT^{-\alpha}$, then (10) asks for $\alpha < \min\{1/2, 1-2/q\}$. As long as $q < 4$, the second condition in (10) is therefore negligible. Our statements also hold for the MSE-optimal bandwidth $h_1 = c \cdot T^{-1/5}$ if $q > 5/2$. In what follows to

formulate the convergence rates of the estimators we use the following notation:

$$r_{T,h} = \sqrt{\frac{\log(T)}{Th}}. \tag{12}$$

The next theorem states the uniform convergence rates for $\widehat{\theta}(u)$ and $\widehat{\sigma^2}(u)$.

Theorem 2. *Let Assumptions 1, 2, 3 hold. Then*

$$\sup_{u\in[h_1,1-h_1]} \left| H(\hat{\theta}(u) - \theta(u)) - h_1^2 \begin{pmatrix} \lambda_2 \partial_u^2 \rho(u) \\ 0 \end{pmatrix} \right|_\infty = o_p(h_1^2) + O_p(r_{T,h_1}).$$

and

$$\sup_{u\in[h_2,1-h_2]} \left| \widehat{\sigma^2}(u) - \sigma^2(u) - h_2^2 \lambda_2 \left\{ \partial_u \rho(u)' \Omega(u) \partial_u \rho(u) + \frac{1}{2} \partial_u^2 (\sigma^2(u)) \right\} \right| = o_p(h_1^2) + O_p(r_{T,h_1}).$$

Remark 5. In fact, the uniform statements of Theorem 2 are proven for the whole interval, that is, for $u \in [0,1]$. However, the bias terms on the boundary become more complicated and therefore we only state the result for $u \in [h_1, 1-h_1]$ or $u \in [h_2, 1-h_2]$, respectively and the rest of the results can be found in the Appendix B in the Supplementary Material.

3.1 Test Statistics

Having obtained the estimates of μ_t and σ_t we can now form an estimate of our new metric, introduced in (1). In line with the discussion in section 2, we choose the weights w_t to be inversely proportional to the standard error of the estimate of μ_t. We might want to make it slightly more general by allowing some extra (given) weighting ϕ_t such that $w_t \propto \phi_t/\sigma_t$. An example of ϕ_t is $\phi_t = \mathbb{1}(t \in \mathcal{I})$, where $\mathcal{I}$ is a period of interest, e.g. recession times. With this in mind, define s to be the following quantity:

$$s := \int_0^1 \phi(u) \frac{\mathbb{E}\mathbb{X}_0(u)'\rho(u)}{\sigma(u)} du,$$

which represents the deterministic version of the metric introduced in (1). Then the null of equal predictive ability EPA_w and the associated fixed alternative hypotheses can be stated formally as follows:

$$\mathbb{H}_0 : \quad s = 0 \quad \text{vs} \quad \mathbb{H}_1 : \quad s \neq 0, \tag{13}$$

The notation EPA_w explicitly acknowledges that these are rather a class of null hypothesises, depending on the chosen weighting scheme. We then form the test statistic corresponding to the

above null by replacing s with a good estimator. First define the local t-statistic, which we denote by $\widehat{\tau}_t$:

$$\widehat{\tau}_t = \frac{\widehat{\mu}_t}{\widehat{se}_t} = \frac{\mathbb{X}'_{t,T}\widehat{\rho}(t/T)}{\widehat{\sigma}(t/T)}. \tag{14}$$

We then propose the corresponding test statistic $\mathcal{S}_T$, which we take to be an estimator of s, given by:

$$\mathcal{S}_T = \frac{1}{T}\sum_{t=1}^{T}\phi(t/T)\widehat{\tau}_t = \frac{1}{T}\sum_{t=1}^{T}\phi(t/T)\frac{\mathbb{X}'_{t,T}\widehat{\rho}(t/T)}{\widehat{\sigma}(t/T)}. \tag{15}$$

Importantly, in what follows in Theorem 3 below we show that $\mathcal{S}_T$ is indeed a good estimator of s. Note that the null and associated hypotheses can only be formulated in terms of s which should be thought of as a theoretical quantity of interest corresponding to the metric in (1). The null can not be formulated in terms of $\mathcal{S}_T$ directly (with true $\rho(\cdot)$ and $\sigma(\cdot)$ functions) since this will result in random null and alternative hypotheses due to the fact that $\mathbb{X}_{t,T}$ itself is random.

Theorem 3. *Let Assumptions 1, 2 and 3 hold with $q > \frac{5}{2}$. Suppose that $Th_1^5 = O(1)$. Then it holds that*

$$\sqrt{T}\Big((\mathcal{S}_T - s) - \mathbb{B}(h_1, h_2)\Big) \xrightarrow{d} N(0, \mathbb{V}),$$

where the bias term $\mathbb{B}(h_1, h_2)$ is given by

$$\begin{aligned}\mathbb{B}(h_1,h_2) \;:=\; & -\frac{1}{2}\lambda_2 h_2^2 \cdot \int_0^1 \frac{\phi(u)}{2\sigma^3(u)}\mathbb{E}\mathbb{X}_0(u)'\rho(u)\partial_u^2(\sigma^2(u))du \\ & + \frac{h_1^2}{2}\lambda_2\int_0^1 \frac{\phi(u)}{\sigma(u)}\mathbb{E}\mathbb{X}_0(u)'\partial_u^2\rho(u)du \\ & - \lambda_2 h_2^2\int_0^1 \frac{\phi(u)}{2\sigma^3(u)}\mathbb{E}\mathbb{X}_0(u)'\rho(u)\cdot\partial_u\rho(u)'\Omega(u)\partial_u\rho(u)du,\end{aligned} \tag{16}$$

and $\mathbb{V}$ is defined in (125) in Appendix B.

Remark 6. The bias term $\mathbb{B}(h_1, h_2)$ is of the order h_1^2, h_2^2. If $h_1 = T^{-\alpha}$, then Assumption 3 and condition $Th_1 = O(1)$ are simultaneously fulfilled for $\min\{\frac{1}{2}, 1 - 2/q\} > \alpha \geq 1/5$. Specifically, the MSE-optimal choice $h_1 = h_2 = cT^{-1/5}$ for estimating $\rho(\cdot), \sigma^2(\cdot)$ with $\hat{\rho}$, $\widehat{\sigma^2}$ is always covered.

3.2 Behavior under local alternatives

In addition, to get an idea of the power of the test, we further examine a series of local alternatives, i.e. alternatives that converge to $\mathbb{H}_0$ as the sample size T grows. In particular, we define the

sequence of functions $\tau_t(t/T)$ given by:

$$\tau_t(t/T) = \tau_t + c_T\Delta(t/T),$$

where $c_T \to 0$ as $T \to \infty$, the function Δ is continuous and the quantity $T^{-1}\sum_{t=1}^{T}\tau_t$ satisfies the null hypothesis $\mathbb{H}_0$. Under these local alternatives the process $\Delta\mathcal{L}_{t,T}$ is given by

$$\Delta\mathcal{L}_{t,T} = \mathbb{X}'_{t,T}\rho\left(t/T\right) + c_T\Delta\left(t/T\right)\sigma\left(t/T\right) + \xi_{t,T}, \qquad t = 1, ..., T. \tag{17}$$

We now show that under (17), we move along the following sequence of local alternatives:

$$\mathbb{H}_{1,T}: \qquad s = c_T \cdot D, \tag{18}$$

where

$$D := \int_0^1 \phi(u)\mathbb{E}\mathbb{X}_0(u)'\mathbb{E}\mathbb{X}_0(u)\Delta(u)du + \int_0^1 \phi(u)\Delta(u)(1, E(u)')\rho(u)du, \tag{19}$$

is a bias term due to the changed mean in the local alternative, where

$$\Gamma(u) := \begin{pmatrix} \rho_1(u) & \rho_2(u) & \dots & \rho_d(u) \\ 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad E(u) = \Big(I_{d\times d} - \Gamma(u)\Big)^{-1}(1, 0, ..., 0)'.$$

The statistic $\mathcal{S}_T$ under $\mathbb{H}_{1,T}$ gets smaller as the sample size increases and therefore the alternatives $\mathbb{H}_{1,T}$ gets closer and closer to $\mathbb{H}_0$ as $T \to \infty$. We next examine the behaviour of $\mathcal{S}_T$ under local alternatives. Theorem 4 below states that the asymptotic power of the test against local alternatives given in (18) with $c_T = 1/\sqrt{T}$ is constant for all functions Δ.

Theorem 4. *Let the model (17) hold with $c_T = 1/\sqrt{T}$. Under the conditions of Theorem 3, we have*

$$\sqrt{T}\Big((\mathcal{S}_T - s - c_T \cdot D) - \mathbb{B}(h_1, h_2)\Big) \xrightarrow{d} N(0, \mathbb{V}),$$

where $\mathbb{B}(h_1, h_2)$ is defined in (16), D is defined in (19) and $\mathbb{V}$ is defined in (125) in Appendix B.

Theorem 4 shows that in model (17), $\mathcal{S}_T$ estimates s but it converges to a normal distribution with mean D.

3.3 Sign Forecasting

We now present the theory for sign forecasting. Given that our model (2) for $\Delta\mathcal{L}_t$ has an autoregressive structure, we may project which model is likely to forecast better in the next period in the following way. Let $\mathcal{A}_T := \sigma(\Delta\mathcal{L}_{t,T}|t \leq T)$ be the sigma-algebra generated by the history of the process $\Delta\mathcal{L}_{t,T}$. At the final point in the sample T we would like to predict the sign of $\Delta\mathcal{L}_{T+1}$, i.e. we want to know

$$\mathbb{P}\left(\Delta\mathcal{L}_{T+1} \leq 0|\mathcal{A}_T\right) = \mathbb{P}\left(\varepsilon_{T+1} \leq \frac{-\mathbb{X}'_{T+1}\rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\Big|\mathcal{A}_T\right) = F_\varepsilon\Big(\frac{-\mathbb{X}'_{T+1}\rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\Big) =: \mathcal{P}_T(\mathbb{X}_{T+1}), \tag{20}$$

where $F_\varepsilon(x) := \mathbb{P}(\varepsilon_1 \leq x)$ denotes the distribution of the i.i.d. errors ε_i, $i \in \mathbb{Z}$. Furthermore, let f_ε denote the density of ε_i with respect to the Lebesgue measure. To estimate (20), we use the approximations $\rho((T+1)/T) \approx \rho(1) \approx \widehat{\rho}(1)$ and $\sigma((T+1)/T) \approx \sigma(1) \approx \widehat{\sigma}(1)$, where we formally prove such an approximation is valid as part of Theorem 5. We then estimate F_ε by the corresponding empirical distribution function $\hat{F}_\varepsilon$ of the estimated residuals $\hat{\varepsilon}_t$, that is,

$$\hat{F}_\varepsilon(y) := \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{\hat{\varepsilon}_t \leq y\}}, \qquad \hat{\varepsilon}_t := \frac{\Delta\mathcal{L}_t - \mathbb{X}'_{t,T}\hat{\rho}(\frac{t}{T})}{\widehat{\sigma}(\frac{t}{T})}. \tag{21}$$

Our final estimator of $\mathcal{P}_T(\mathbb{X}_{T+1})$ is given by $\hat{\mathcal{P}}_T(\mathbb{X}_{T+1})$, where

$$\hat{\mathcal{P}}_T(x) := \hat{F}_\varepsilon\Big(\frac{-x'\hat{\rho}(1)}{\widehat{\sigma}(1)}\Big).$$

The following definition is only to state the next theoretical result but is not needed for the corresponding application (see Remark 7 below). For $x = (x_j)_{j\in\{1,\dots,d+1\}} \in \mathbb{R}^{d+1}$, $m \in \mathbb{R}_{>0}$, define the truncation of x by

$$x^{\wedge m} := (\max\{-m, \min\{x_j, m\}\})_{j\in\{1,\dots,d+1\}},$$

that is, all entries of $x^{\wedge m}$ are at most m in absolute value. Therefore for a given sample $\{\Delta\mathcal{L}_t\}_{t=1}^T$ the practitioner can calculate the probability of $\Delta\mathcal{L}_{T+1}$ of being negative. We state the theoretical result in Theorem 5 below, which allows the practitioner to calculate the probability as well as the confidence intervals for this probability, which we call **forecast intervals**.

Theorem 5. *Suppose that Assumptions 1, 2, 3 hold for some $q > \frac{5}{2}$. Suppose that f_ε is twice continuously differentiable with $\sup_{y\in\mathbb{R}}(f_\varepsilon(y)y^2) < \infty, \sup_{y\in\mathbb{R}}|\partial_y f_\varepsilon(y)y^2| < \infty$. Let $\delta \in (0,1)$ be arbitrarily small,*

and assume that

$$\frac{\log(T)}{Th^{3+2\delta}} = o(1), \qquad Th_1^5 = o(1). \tag{22}$$

Then there exists $\gamma > 0$ small enough such that

$$\left[\sqrt{Th_1}(\hat{\mathcal{P}}_T(x^{\wedge T^\gamma}) - \mathcal{P}_T(x^{\wedge T^\gamma}))\right]_{x\in\mathbb{R}^{d+1}} \xrightarrow{d} \left[G(x)\right]_{x\in\mathbb{R}^{d+1}}, \tag{23}$$

where $G(x)$ is a centered Gaussian process with covariance function

$$Cov(G(x_1), G(x_2)) = f_\varepsilon\left(x_1'\frac{\rho(1)}{\sigma(1)}\right)\cdot f_\varepsilon\left(x_2'\frac{\rho(1)}{\sigma(1)}\right)\cdot x_1' V_z x_2,$$

where V_z is defined in the proof in (137) in Appendix B.

Remark 7. Theorem 5 is applied to $\hat{\mathcal{P}}_T(\mathbb{X}_{T+1})$ by inserting $x = \mathbb{X}_{T+1}$ into (23) which is possible due to $\mathbb{P}(|\mathbb{X}_{T+1}|_\infty > T^\gamma) \leq \mathbb{E}|\mathbb{X}_{T+1}|_\infty / T^\gamma = O(T^{-\gamma}) = o(1)$ as $T \to \infty$ and gives the following approximation in distribution

$$\sqrt{Th_1}(\hat{\mathcal{P}}_T(\mathbb{X}_{T+1}) - \mathcal{P}_T(\mathbb{X}_{T+1})) \approx G(\mathbb{X}_{T+1}).$$

Note that in general, $G(\mathbb{X}_{T+1})$ is not Gaussian distributed. In (23), there is a bias term of order $\sqrt{Th_1^5}$ present. The exact derivation of this term is tedious and we omit it due to its limited use in practice. Provided our conditions on the bandwidths, this bias term vanishes asymptotically. The condition (22) narrows the set of possible bandwidths provided by Assumption 3. Allowable bandwidths are $h_1 = cT^{-\alpha}$ with $\min\{1/3, 1-2/q\} > \alpha > 1/5$.

In the simulations, see Figure 7, we see that the sign forecasts perform quite well, forecasting near the true probability. In particular, the sign forecasts improve as we go later in the sample. Because the bandwidth for the first estimation step for ρ is quite small in this particular example, this improvement is not due to estimating ρ more precisely; instead it is due to approximating the c.d.f. of ε_t better as we go later in the sample, using more data. In general, it looks as if the difficulty of approximating the c.d.f. of ε_t is greater than the issues surrounding estimating ρ imperfectly. Also, because we are not interested in forecasting the level of the forecast loss difference next period, but rather its sign, our results are somewhat less sensitive to the imprecision caused by using a two-sided kernel. In general, if the p.d.f. at the particular $\widehat{\varepsilon}_t$ threshold is small, the probability will not respond much to inaccuracies in ρ and σ. One way to improve forecasts even further would be for instance to use the derivative of $\widehat{\rho}(1)$.

In practice, we are only concerned about making predictions at the last point in time T, however

we can nonetheless produce pseudo out-of-sample sign forecasting to assess the quality of our procedure. In our simulation, we derive the true probabilities explicitly and compare it with our forecasted probabilities. In addition, in our simulations we use the following criterion to assess the quality of our forecasts:

$$\widehat{C} := \frac{1}{T-\underline{T}} \sum_{t=\underline{T}}^{T} \left[\mathbb{1}\left(\Delta\mathcal{L}_{t+1} \leq 0\right) - \widehat{\mathbb{P}}^{bc}\left(\Delta\mathcal{L}_{t+1} \leq 0 | \mathcal{A}_T\right) \right], \tag{24}$$

where $\underline{T} = 100$ is the first splitting point where we begin our evaluation, and $\widehat{\mathbb{P}}^{bc}(\cdot)$ denotes the bias-corrected probability. If the forecasted probabilities were correct, then the criterion above should on average equal to zero. The bias as well as the forecast intervals can be obtained via bootstrap which we discuss in detail in section 5.

4 Bandwidth selection

In this section we briefly describe how we choose bandwidths h_1 and h_2. We start with the selection of the first stage estimation bandwidth h_1. The conventional way to choose the optimal bandwidth is to construct the asymptotic mean squared error given by:

$$\text{AMSE}(h_1) = \frac{h_1^4}{4}\lambda_2^2 \int_0^1 |\partial_u^2 \rho(u)|_2^2 du + \frac{\nu_0}{T h_1} \int_0^1 \sigma^2(u) \text{tr}(\Omega(u)^{-1}) du,$$

(cf. Theorem 1). Then minimizing $\text{AMSE}(h_1)$ with respect to h_1 provides the optimal bandwidth h_1^{opt} given by:

$$h_1^{opt} = \lambda_2^{-2} T^{-1/5} \int_0^1 \left\{ \text{tr}(\Omega(u)^{-1}) |\partial_u^2 \rho(u)|_2^{-2} \right\}^{-1/5} du \tag{25}$$

However, note that (25) involves the unknown quantity $\partial_u^2 \rho(u)$ that therefore has to be estimated first before the optimal bandwidth can be computed. Several other methods has been proposed in the literature, one of which is multi-fold cross-validation see e.g. Cai et al. (2000b), Cai et al. (2000a) which takes into account the time-series structure of the data. More precisely, we first partition the data into Q groups (usually $Q = 20$), with the jth group consisting of the data points with indices:

$$d_j = \{Qk + j, \;\; k = 1, 2, 3, \cdots\}, \;\; j = 0, 1, 2, \cdots, Q-1.$$

We then fit the model and obtain the estimate of $\widehat{\theta}^{-j}(u)$ given in (6) by using the remaining data after deleting the jth group. Now denote by $\widehat{\Delta\mathcal{L}}_{-d_j,T}$ the fitted values of $\Delta\mathcal{L}_{t,T}$ using the data with

the jth group deleted. Then the cross-validation criterion has the following form:

$$\mathrm{CV}(h_1) = \sum_{j=0}^{Q-1} \sum_{i \in d_j} \left[\Delta \mathcal{L}_{i,T} - \widehat{\Delta \mathcal{L}}_{-d_j,T} \right]^2.$$

Minimizing the $\mathrm{CV}(h_1)$ with respect to h_1 then yields the estimated bandwidth $h_1^\star$. In practice, and in general, as established by Cai et al. (2000a) the cross-validation is not particularly sensitive to the way the data is partitioned. The second-stage estimation procedure of estimating the conditional variance $\sigma^2(u)$ via the local linear estimator is standard, and the optimal bandwidth $h_2^\star$ is estimated via conventional least-squares cross-validation, see e.g. Li and Racine (2007) for details.

5 Bootstrapping $\mathcal{S}_T$

Theorem 3 allows one to conduct inference for $\mathcal{S}_T$ as the distribution of the test statistics is a simple normal distribution. Note also, that the test statistic $\mathcal{S}_T$ is a nonparametric statistic, however through aggregation it converges to the limit distribution with the standard parametric $\sqrt{T}$ rate.

The bias term $\mathbb{B}(h_1, h_2)$ and the variance term $\mathbb{V}$ in Theorem 3 however contain unknown quantities, such as $\partial_u^2 \rho(u)$. Although it is possible to estimate these unknown quantities, replacing them with consistent estimates will result in additional approximation errors. We therefore propose a bootstrap approach, which is capable of estimating the unknown variance $\mathbb{V}$. In what follows, we discuss the bootstrap procedure in the context of Theorem 3, however the same methodology will be applied to obtain the bias and the forecast intervals in Theorem 5.

The asymptotics of $\mathcal{S}_T$ is determined by the asymptotic distributions of the three random terms $\mathbb{X}_{t,T}$, $\hat{\rho}(t/T)$ and $\hat{\sigma}(t/T)$. A fixed regressor bootstrap is only capable of estimating the contributions of the last two terms $\hat{\rho}(t/T)$ and $\hat{\sigma}(t/T)$. We therefore propose a time series bootstrap which we set up in the following way. Create the bootstrap sample $\{\Delta \mathcal{L}_{t,T}^*\}_{t=1}^T$ as follows:

$$\Delta \mathcal{L}_{t,T}^* = (\mathbb{X}_{t,T}^*)' \widehat{\rho}_g(t/T) + \widehat{\sigma}_g(t/T) \cdot \varepsilon_{t,T}^*, \tag{26}$$

where $\widehat{\rho}_g(\cdot)$, $\widehat{\sigma}_g(\cdot)$ are the estimators of $\rho(\cdot), \sigma(\cdot)$, respectively from the original sample with a bandwidth $g \gg h_1$ and where the bootstrap residuals $\varepsilon_{t,T}^*$ are constructed as follows: Define the standardized residuals

$$\tilde{\varepsilon}_{t,T} := \frac{\bar{\varepsilon}_{t,T}}{\left(\frac{1}{T}\sum_{j=1}^T \bar{\varepsilon}_{j,T}^2\right)^{1/2}}, \qquad \bar{\varepsilon}_{t,T} = \hat{\varepsilon}_{t,T} - \frac{1}{T}\sum_{j=1}^T \hat{\varepsilon}_{j,T}, \qquad \hat{\varepsilon}_{t,T} := \frac{\Delta \mathcal{L}_t - \mathbb{X}_{t,T}' \hat{\rho}_g(t/T)}{\widehat{\sigma}_g(t/T)}. \tag{27}$$

Next construct $\varepsilon^*_{t,T}$ by randomly drawing with replacement from $\{\tilde{\varepsilon}_{t,T}, t = 1, ..., T\}$. The estimators $\hat{\rho}^*(u), \widehat{\sigma}^*(u)$ are then constructed as in (6) and (8) but using $\Delta\mathcal{L}^*_{t,T}$ and $\mathbb{Z}^{u*}_t = \mathbb{Z}^{u*}_{t,T} = (\mathbb{X}^*_{t,T}, \mathbb{X}^*_{t,T}(t/T - u))'$ with $\mathbb{X}^*_{t,T} = (1, \Delta\mathcal{L}^*_{t-1,T}, ..., \Delta\mathcal{L}^*_{t-d,T})'$ instead of the corresponding terms with $\Delta\mathcal{L}_{t,T}$ and using the original optimal bandwidths h_1 and h_2. As bootstrap statistics of $\mathcal{S}_T$, we consider

$$\mathcal{S}^*_T := \frac{1}{T}\sum_{t=1}^{T} \phi(t/T) \frac{(\mathbb{X}^*_{t,T})'\hat{\rho}^*_{h_1}(t/T)}{\widehat{\sigma}^*_{h_2}(t/T)}.$$

Let d_2 denote Mallow's distance, i.e. for random variables X, Y with distribution functions F_X, F_Y,

$$d_2(X, Y) = \Big(\int_0^1 (F_X^{-1}(x) - F_Y^{-1}(x))^2 \Big)^{1/2},$$

see Bickel and Freedman (1981). We are now able to state an asymptotic result for $\mathcal{S}^*_T$. To discuss the bias of the bootstrap version $\mathcal{S}^*_T$, we have to pose the following assumption.

Assumption 4. *The kernel K is three times continuously differentiable.*

We next state our next result stating the validity of the bootstrap set-up described above.

Theorem 6. *Let the conditions of Theorem 3 hold, and additionally suppose that Assumption 4 holds. Suppose that* $Th_1^5 = o(1)$ *and that the pre-estimation bandwidth g satisfies Assumption 3 and*

$$\frac{\log(T)}{Tg^5} = O(1), \qquad \frac{\log(T)}{Tg^3h_1^2} = O(1). \tag{28}$$

Then it holds that

$$d_2(\sqrt{T}(\mathcal{S}^*_T - \tilde{\mathcal{S}}_T), \sqrt{T}(\mathcal{S}_T - \mathbb{E}\mathcal{S}_T)) \xrightarrow{p} 0,$$

where $\tilde{\mathcal{S}}_T = \frac{1}{T}\sum_{t=1}^{T} \phi(\frac{t}{T}) \cdot \frac{\mathbb{E}[\mathbb{X}^*_{t,T}|\mathcal{A}_T]'\hat{\rho}(t/T)}{\widehat{\sigma}(t/T)}$ *and* $\mathcal{A}_T := (\Delta\mathcal{L}_{1,T}, ..., \Delta\mathcal{L}_{T,T})$.

We next comment on conditions on the bandwidths as well as give guidance on how Theorem 6 can be utilized in practice.

Remark 8. If $h_1 = cT^{-\alpha}$, $g = c_gT^{-\beta}$, then it has to hold by (28) that

$$0 < \beta < \min\{\frac{1}{5}, \frac{1-2\alpha}{3}\} \leq \frac{1}{5} < \alpha < \min\{\frac{1}{2}, 1 - \frac{2}{q}\},$$

This implies that the pre-estimation bandwidth $g \gg h_1$ has to be chosen larger than h_1, h_2. In practice, one may simply choose $g = 2h_1$. The quantity $\tilde{\mathcal{S}}_T$ does not correspond to $\mathcal{S}_T$ since $\mathbb{E}[\mathbb{X}^*_{t,T}|\mathcal{A}_T]$

is a smoothed version of $\mathbb{X}_{t,T}$. In this sense, using $\tilde{\mathcal{S}}_T \approx \mathcal{S}_T$ may asymptotically not lead to correct bootstrap quantiles. Instead, we propose to use $\tilde{\mathcal{S}}_T \approx \tilde{\mathcal{S}}_T^{\circ}$, where

$$\tilde{\mathcal{S}}_T^{\circ} = \frac{1}{T}\sum_{t=1}^{T} \phi(\frac{t}{T}) \cdot \frac{(1, \hat{E}_t')'\hat{\rho}(t/T)}{\widehat{\sigma}(t/T)},$$

where

$$\hat{E}_t := \left(I_{d\times d} - \widehat{\Gamma}(t/T)\right)^{-1} (\widehat{\rho}_0(t/T), 0, \cdots, 0)',$$

where $\Gamma(\cdot)$ matrix is defined in Theorem 4. Note that $\tilde{\mathcal{S}}_T^{\circ}$ yields a valid approximation of $\tilde{\mathcal{S}}_T$ since $\mathbb{E}[\mathbb{X}_{t,T}^{*}|\mathcal{A}_T] - (1, \hat{E}_t')' = O_p(T^{-1})$.

6 Simulations

In this section we provide the simulation results for the size and the power of our test statistic $\mathcal{S}_T$ as well as demonstrating the sign forecasting methodology. Throughout this section we shall be simulating $\Delta\mathcal{L}_t$, $t = 1, \cdots, T$ directly as the main purpose of this section is to investigate the properties of our methodology under the known behavior for $\Delta\mathcal{L}_t$. We shall however produce $\Delta\mathcal{L}_t$ using various forecasting methods in our application in section 7. Given that Giacomini and White (2006) also consider the question of forecast selection, throughout this section we replicate several of their simulation exercises to make for easy comparison when possible.

Once the bootstrap is set up, the size and the power of the test is then calculated as follows. We denote by $\mathcal{S}_{T,n}$ the value of the test statistic $\mathcal{S}_T$ in the n-th simulation, and let $\mathcal{S}_{T,n,b}^{\star}$ be the value of the bootstrap statistics $\mathcal{S}_T^{\star}$ in the b-th bootstrap sample generated in the n-th simulation. We denote by $G_n^{\star}$ the empirical distribution function calculated from the sample of the bootstrap values in n-th simulation, i.e. of $\{\mathcal{S}_{T,n,b}^{\star} - \tilde{\mathcal{S}}_T^{\circ}\}_{b=1}^{B}$. Then the actual size of the test statistics can be calculated as follows. Given a fixed nominal size α, for each simulated sample $n \in \{1, \cdots, N\}$, calculate the $(1-\alpha)$-quantile of $G_n^{\star}$, denoted by $q_{\alpha,n}^{\star}$. Finally we compute the actual size and power corresponding to the nominal level α as

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(\mathcal{S}_{T,n} > q_{\alpha,n}^{\star}\right).$$

We start by investigating the size of our test statistic $\mathcal{S}_T$.

6.1 Test Statistics: Size

For all simulations we set the number of simulations $N = 1000$ and we vary the number of bootstrap replications, B, between $B = 500$, $B = 750$, and $B = 1000$. We start with replicating two alternatives from Giacomini and White (2006) that constitute our null hypothesis. In particular we simulate the loss difference $\Delta\mathcal{L}_t$ as the following AR(1) process:

$$\mathbb{H}_0^{(1)}: \qquad \Delta\mathcal{L}_t = \mu(1-\rho) + \rho\Delta\mathcal{L}_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d.\mathcal{N}(0,1) \tag{29}$$

For each $n \in \{1, \cdots, N\}$ simulations we generate a sequence of loss differences $\Delta\mathcal{L}_t$ of length $T = 150$ according to (29), starting from the initial value of $\Delta\mathcal{L}_t$ that equals the difference of squared errors for forecasts of the second log difference of the monthly U.S. consumer price index (CPI), $\text{CPI}_{2016:12}$ implied by two models: i) a white noise; and ii) an AR(1) model for CPI estimated over a window of size $m = 150$ using the data up to 2016:11. Moreover, we consider the scenario with zero unconditional mean and $\rho = (0, 0.05, \cdots, 0.9)$.[6]

For the second null hypothesis, also considered in Giacomini and White (2006), for $T = 150$ we generate the sequence of loss differences as follows:

$$\mathbb{H}_0^{(2)}: \qquad \Delta\mathcal{L}_t = \frac{\mu}{p(1-p)}(S_t - p) + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.\mathcal{N}(0,1), \tag{30}$$

where $S_t = 1$ with probability p and $S_t = 0$ with probability $1-p$, with $p = 0.5$. We thus have that the unconditional mean $\mathbb{E}[\Delta\mathcal{L}_t] = 0$ irrespective of the value of μ, however

$$\mathbb{E}[\Delta\mathcal{L}_t | S_t] = \begin{cases} \mu/p & \text{if} \quad S_t = 1 \\ -\mu/(1-p) & \text{if} \quad S_t = 0. \end{cases}$$

We set $\mu = 1$ in this example. Finally, we simulate the data for $\Delta\mathcal{L}_{t,T}$ for the sample of length $T = 1000$ under $\mathbb{H}_0^{(3)}$ such that mean is time-varying:

$$\mathbb{H}_0^{(3)}: \qquad \Delta\mathcal{L}_{t,T} = \rho_0(t/T)\left(1 - \rho_1(t/T)\right) + \rho_1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

where $\sigma(t/T) = 1$ for all $t = 1, \cdots, T$ and

$$\rho_0(t/T) = \sin(8\pi t/T), \qquad \rho_1(t/T) = \frac{1}{4}\big(\sin(4\pi t/T) + 1\big).$$

[6] We present the results for $\rho = 0.3$ only as varying ρ virtually leaves the results unchanged.

Under $\mathbb{H}_0^{(3)}$ the mean of $\Delta\mathcal{L}_t$ is time-varying, however the overall mean of $\Delta\mathcal{L}_{t,T}$ is still zero. For simplicity we set the variance to be constant throughout. For each of the aforementioned nulls Table 1 shows the simulated actual size for different levels of the nominal size $\alpha = 0.01,\ 0.05,\ 0.10,\ 0.15$. From Table 1 we can see that the actual size is very close to the nominal size for all levels and for all nulls under consideration. The results are also stable regardless of the number of bootstrap replications B.

Table 1: Actual size versus nominal size of $\mathcal{S}_T$ for various nulls.

Actual size versus nominal size of $\mathcal{S}_T$ for null $\mathbb{H}_0^{(1)}$

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.011	0.056	0.092	0.144
$B = 750$	0.010	0.054	0.096	0.152
$B = 1000$	0.009	0.051	0.104	0.151

Actual size versus nominal size of $\mathcal{S}_T$ for null $\mathbb{H}_0^{(2)}$

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.020	0.062	0.108	0.168
$B = 750$	0.018	0.057	0.107	0.160
$B = 1000$	0.013	0.050	0.103	0.153

Actual size versus nominal size of $\mathcal{S}_T$ for null $\mathbb{H}_0^{(3)}$

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.013	0.055	0.107	0.141
$B = 750$	0.012	0.051	0.105	0.145
$B = 1000$	0.012	0.050	0.102	0.148

Note: In this table B denotes the number of bootstrap replications. Number of simulations N in all cases is always fixed to be $N = 1000$.

6.2 Test Statistics: Power

We start by replicating two alternatives from Giacomini and White (2006) that also constitute alternatives for our test. The first alternative, which we denote by $\mathbb{H}_{1,1}^{GW}$, simulates the loss differences $\Delta\mathcal{L}_t$ according to (29) such that $\rho = 0$ and $\mu = (0, 0.05, \cdots, 1)$. We fix the nominal size of the test to be 5%. In Figure 3 below we show the power curves when applying our test as well as Giacomini and White (2006) test.

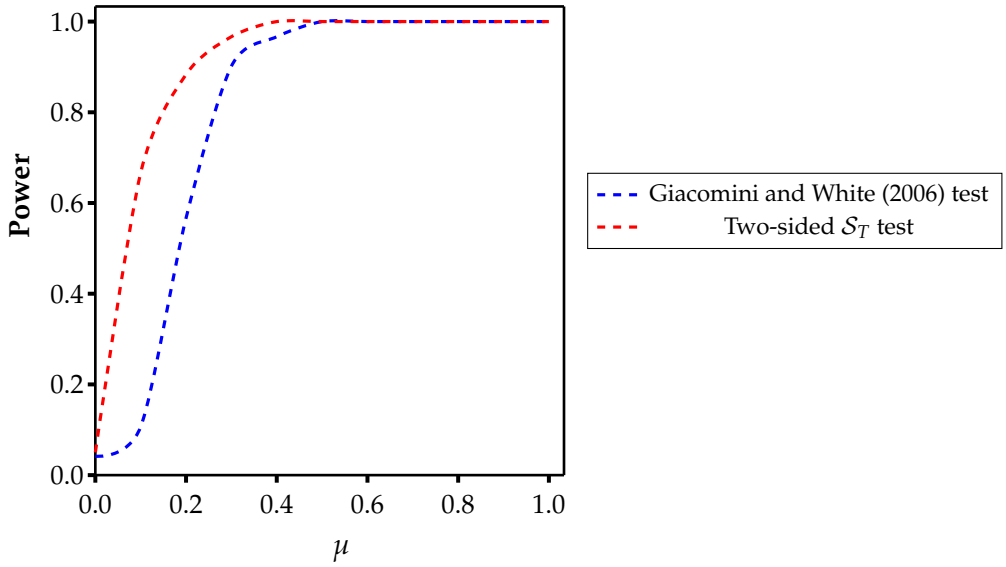

Figure 3: Power curves under alternative $\mathbb{H}_{1,1}^{GW}$.

We next consider another alternative from Giacomini and White (2006) paper, which we further denote by $\mathbb{H}_{1,2}^{GW}$. In particular, we again generate the loss differences $\Delta\mathcal{L}_t$ according to (30), where we vary $d = \frac{\mu}{p(1-p)} = (0, 0.1, \cdots, 1)$. Note that d represents the difference in expected loss between two states. We apply our general test $\mathcal{S}_T$ by setting the chosen weighting function to be the state of the world, i.e. we set $\phi_t = S_t$. In this case (30) constitutes an alternative for our null as well. We plot the power curves in Figure 4 below.

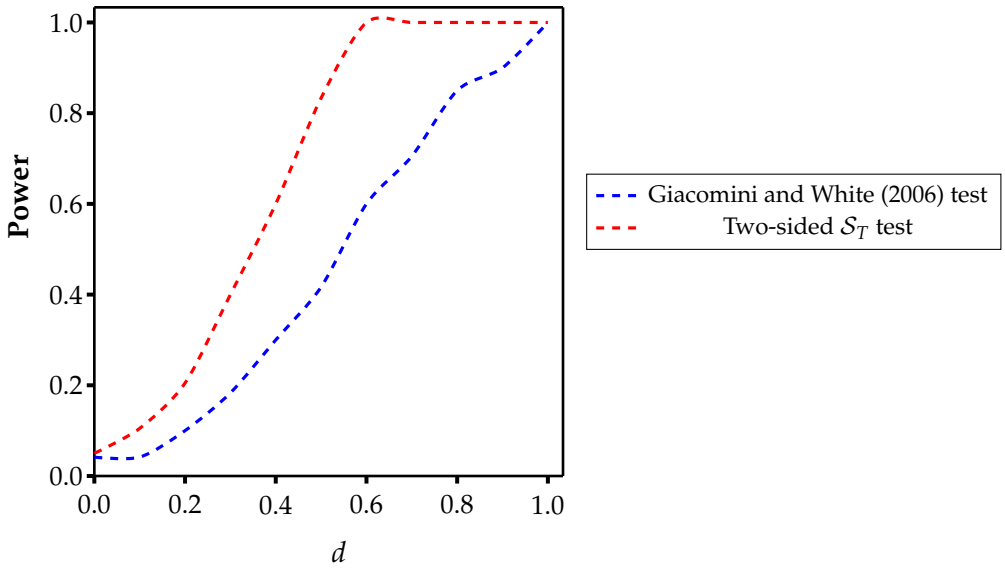

Figure 4: Power curves under alternative $\mathbb{H}_{1,2}^{GW}$.

We now investigate the power of the test under several fixed alternatives that exhibit time variation of the mean/variance process. We deliberately design the set of these alternatives to be similar to our earlier time-varying null $\mathbb{H}_0^{(3)}$, however we add one additional feature that makes for a deviation from the null. Under the first alternative $\mathbb{H}_1^{(1)}$ we simulate the data as follows:

$$\Delta\mathcal{L}_{t,T} = \rho_0(t/T)(1-\rho_1(t/T)) + \rho_1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

where $\sigma(t/T) = 1$ for all $t = 1, \cdots, T$ and

$$\rho_0(t/T) = \sin(8\pi t/T) + 0.1, \qquad \rho_1(t/T) = \frac{1}{4}\big(\sin(4\pi t/T) + 1\big).$$

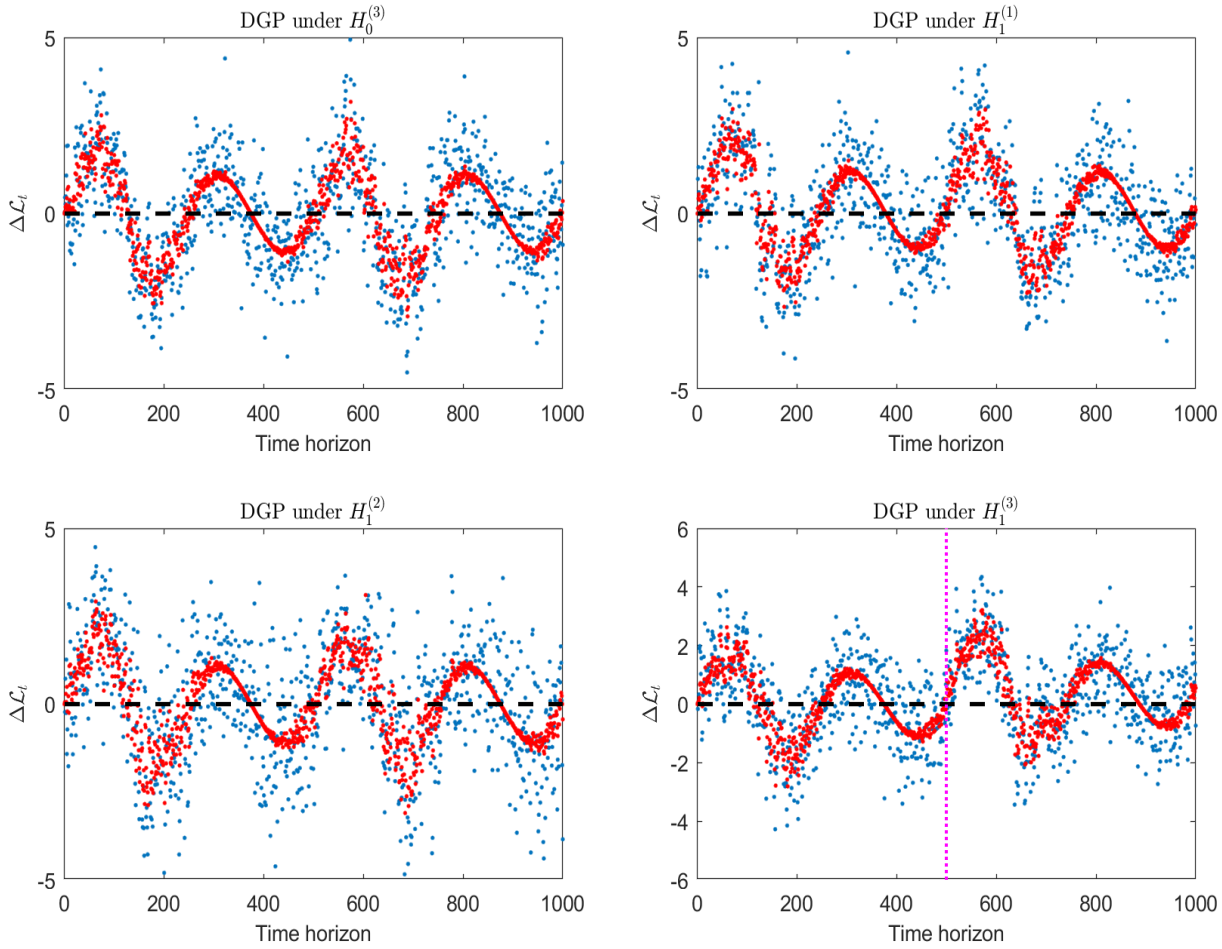

Figure 5: Data generating processes (DGP) under the null $\mathbb{H}_0^{(3)}$ and the corresponding alternatives $\mathbb{H}_1^{(1)}$, $\mathbb{H}_1^{(2)}$ and $\mathbb{H}_1^{(3)}$. Blue dots represent the data and red dots represents the true mean function μ_t.

Under $\mathbb{H}_1^{(1)}$, we shift upwards the intercept such that the overall mean of $\Delta\mathcal{L}_{t,T}$ is no longer zero. The deviation is hard to differentiate visually due to the variance around the mean, and the mean still goes above and below zero, with relative performance overtaking back and forth.

Under $\mathbb{H}_1^{(2)}$ we leave the mean the same as under the null and change the variance in a way that all upswings of the sine function are less volatile and downswings are more volatile, more precisely:

$$\Delta\mathcal{L}_{t,T} = \rho_0(t/T)\left(1-\rho_1(t/T)\right) + \rho_1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

and where

$$\rho_0(t/T) = \sin\left(8\pi t/T\right), \qquad \rho_1(u) = \frac{1}{4}\big(\sin\left(4\pi t/T\right)+1\big),$$

and setting $w = T/8$, the local variance is given by

$$\sigma(t/T) = \begin{cases} 1 \quad \forall \ t \in [1+kw, (k+1)w] \quad \text{for} \quad k = 0, 2, 4, 6. \\ 1.5 \quad \forall \ t \in [1+kw, (k+1)w] \quad \text{for} \quad k = 1, 3, 5, 7. \end{cases}$$

Note that although the mean function under $\mathbb{H}_1^{(2)}$ is the same as under $\mathbb{H}_0$, due to the changes in the variance, the upper swings shall receive more weight as they are less volatile, while the opposite shall hold for the downswings. As the result, we expect the overall statistic to be positive, pointing towards the preference of model B versus model A.

Finally, we consider the alternative $\mathbb{H}_1^{(3)}$ that allows for a break in the intercept. In particular, under $\mathbb{H}_1^{(3)}$ we simulate the data as follows:

$$\Delta\mathcal{L}_{t,T} = \rho_0(t/T)\left(1 - \rho_1(t/T)\right) + \rho_1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \ \mathcal{N}(0,1),$$

where $\rho_1(t/T) = \frac{1}{4}\big(\sin(8\pi t/T) + 1\big)$ and $\sigma(t/T) = 1$ for all $t = 1, \cdots, T$, and

$$\rho_0(t/T) = \begin{cases} \sin(8\pi t/T) & \text{for } t \in [1, T/2], \\ \sin(8\pi t/T) + 0.1 & \text{for } t \in [T/2+1, T]. \end{cases}$$

This alternative highlights the ability for our statistic from deal with breaks. Here the deviation to the null is smaller than the first alternative where the intercept added is throughout the whole sample.

Table 2: Mean of $\mathcal{S}_T$.

Alternative	$\mathbb{E}\left(\mathcal{S}_T\right)$
$\mathbb{H}_1^{(1)}$	0.13
$\mathbb{H}_1^{(2)}$	0.21
$\mathbb{H}_1^{(3)}$	0.06

Table 3: Power for fixed alternative hypotheses.

Nominal size	$\mathbb{H}_1^{(1)}$	$\mathbb{H}_1^{(2)}$	$\mathbb{H}_1^{(3)}$
$\alpha = 0.01$	0.83	0.75	0.44
$\alpha = 0.05$	0.96	0.86	0.52
$\alpha = 0.10$	0.97	0.95	0.66
$\alpha = 0.15$	0.97	0.98	0.68

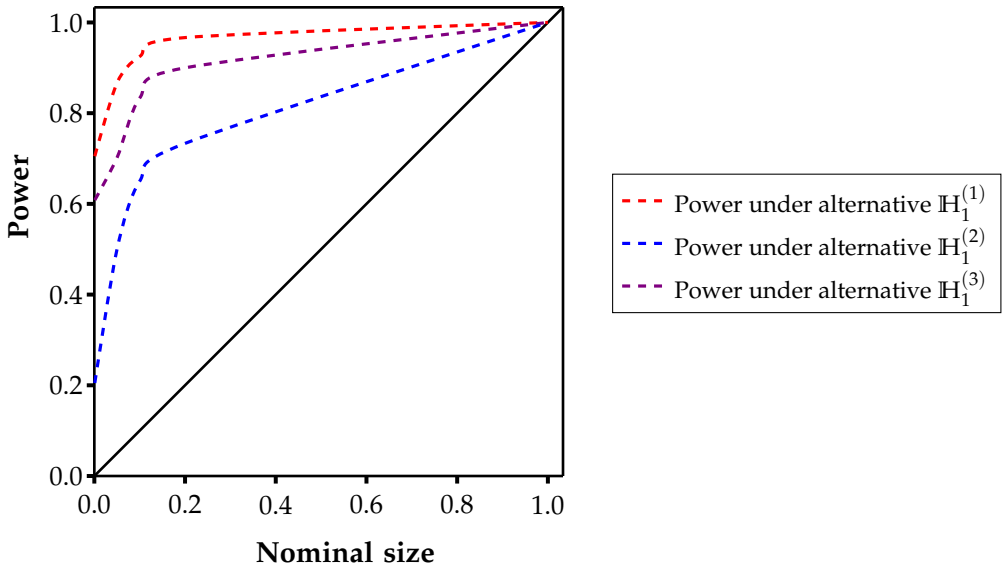

Figure 6: The figure displays the power curves for different alternatives. The dashed blue line depicts the power curve under $\mathbb{H}_1^{(3)}$, the dashed violet line depicts the power curve under $\mathbb{H}_1^{(2)}$, and the dashed red line depicts the power curve under $\mathbb{H}_1^{(1)}$.

Figure 6 shows that our test has very good power at any nominal level and is capable of detecting relatively small deviations from the null. We finish this section with the following thought experiment. Assume that the true data generating process for $\Delta\mathcal{L}_{t,T}$ is indeed as under one of the considered alternatives $\mathbb{H}_1^{(1)}$, $\mathbb{H}_1^{(2)}$ or $\mathbb{H}_1^{(3)}$. Assume that researcher applies any currently available test, e.g. Diebold and Mariano (1995) test or Giacomini and White (2006) test to decide whether competing models have equal forecasting performance. As with any existing out-of-sample test the researcher would have to choose the splitting point. Table 4 below displays the results of applying these tests as function of the cutoff point.

Table 4: $p-$values of applying existing tests under our alternatives.

Results when $\Delta\mathcal{L}_t$ is simulated according to $\mathbb{H}_1^{(1)}$.

Cutoff κ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DM (1995)	**0.005**	0.203	0.808	**0.021**	**0.016**	0.582	**0.031**	0.215	0.631
GW (2006)	**0.011**	**0.010**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Results when $\Delta\mathcal{L}_t$ is simulated according to $\mathbb{H}_2^{(1)}$.

Cutoff κ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DM (1995)	**0.001**	**0.036**	0.187	0.067	**0.012**	0.656	**0.048**	0.492	0.609
GW (2006)	0.086	0.06	0.100	**0.010**	**0.004**	**0.005**	**0.010**	**0**	**0**

Results when $\Delta\mathcal{L}_t$ is simulated according to $\mathbb{H}_3^{(1)}$.

Cutoff κ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DM (1995)	**0**	0.464	0.185	**0.018**	**0.016**	0.877	**0.039**	0.263	0.367
GW (2006)	**0**	0.102	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Note: The cutoff point κ is defined as the fraction of the estimation subsample to the length of the full sample. The values in the table present the p-values from the corresponding tests at the nominal level of $\alpha = 5\%$. The p-values in bold indicate rejection at the nominal size α. DM abbreviates Diebold and Mariano (1995) test of equal predictive ability and GW abbreviates Giacomini and White (2006) test of equal conditional predictive ability with $h_t = [1, \Delta\mathcal{L}_{t-1}]'$.

Table 4 shows that the conclusion of the tests, especially Diebold and Mariano (1995) test, can change depending on the splitting point when applied to our alternatives. Giacomini and White (2006) test suffers less from the splitting point problem and with a reasonable estimation sample delivers consistent results. Interestingly, for many splitting points Diebold and Mariano (1995) test does not reject the null of equal predictive ability, while Giacomini and White (2006) test does reject the same null. This is indicative of changing relative performance as we knew ex-ante, hence this is an example where the existing methodology based on constant relative performance is inappropriate. We stress that the presented thought experiment is not a reflection on the tests as they were not designed to deal with the world of changing relative performance, but rather to highlight the dangers that the researcher runs into when applying existing tests that rely on an assumption that does not hold.

6.3 Sign Forecasting

In this section we assess how our methodology for sign forecasting, described in section 3.3, performs with a known data-generating process. In this case the true probability $\mathcal{P}_T(\mathbb{X}_{T+1}) := \mathbb{P}\left(\Delta\mathcal{L}_{T+1} \leq 0 | \mathcal{A}_T\right)$ is known, where recall $\mathcal{A}_T := \sigma(\Delta\mathcal{L}_{t,T} | t \leq T)$ and we choose $\mathbb{H}_0^{(3)}$ as our true data generating process for $\Delta\mathcal{L}_{t,T}$. For simplicity, we also set $\rho_1(t/T) = 0$ for all $t = 1, \cdots, T$. We then forecast the probability $\mathbb{P}\left(\Delta\mathcal{L}_{t+1} \leq 0 | \mathcal{A}_t\right)$, starting from $t = \underline{T} = 100$.

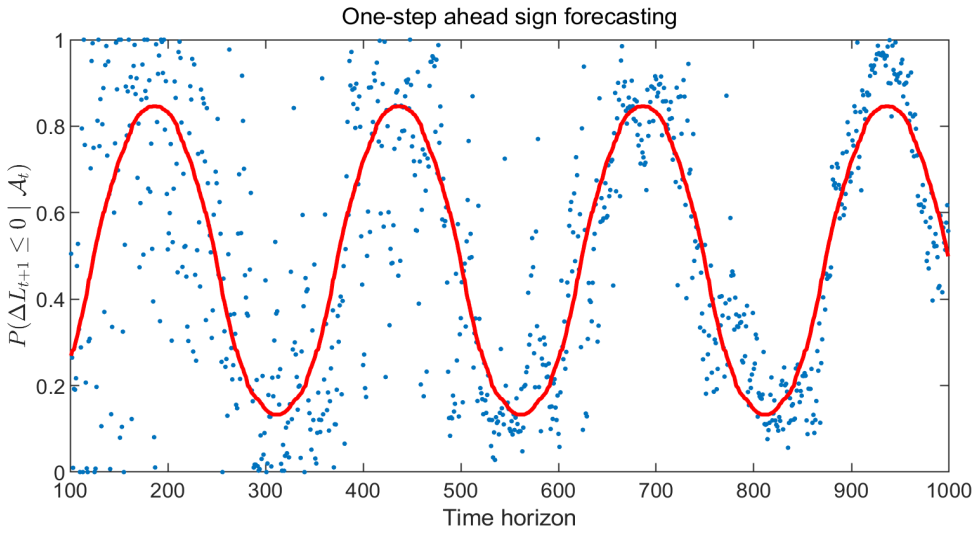

Figure 7: The red line plots the true probability $\mathbb{P}\left(\Delta\mathcal{L}_{t+1} \leq 0 | \mathcal{A}_t\right)$ and the blue dots represent the estimate $\widehat{\mathbb{P}}\left(\Delta\mathcal{L}_{T+1} \leq 0 | \mathcal{A}_T\right)$ starting at $t = \underline{T} = 100$.

Figure 7 displays the true probability $\mathbb{P}\left(\Delta\mathcal{L}_{t+1} \leq 0|\mathcal{A}_t\right)$ against its estimate $\widehat{\mathbb{P}}\left(\Delta\mathcal{L}_{t+1} \leq 0|\mathcal{A}_t\right)$, where for each point on the curve the data up to $\underline{T}$ is used, where $\underline{T} = 100, \cdots, T$. Overall, the estimated probability is quite close to its true value and becomes more precise the more data is used for the original estimation. This happens primarily due to the c.d.f. of the error term $\widehat{\varepsilon}_t$, defined in (21), being better estimated towards the end of the sample as more data is used. At the final point in the sample, we forecast a probability of 0.3829 with a corresponding forecast interval of $[0.3520,\ 0.4200]$. Finally, applying our criterion, given in eq. (24), we get $\widehat{C} = -0.012$, which points to the fact that our estimated probability $\widehat{\mathbb{P}}\left(\Delta\mathcal{L}_{t+1} \leq 0|\mathcal{A}_t\right)$ is on average overestimated by approximately by 1.2% for this sample.

7 Application

In this section we apply our proposed methodologies to the data. We first go back to the motivating example we presented in the introduction in Figure 1.

7.1 Forecasting IBM daily variance

We reconsider the example from our introduction using our new methodologies. Our data is daily IBM returns spanning 03/01/2006-03/01/2017 and we use two models to forecast daily variance: GARCH(1,1) model with Gaussian errors and GARCH(1,1) model with Student-t errors. The forecast loss is taken to be the squared error, see eq.(2) and constructed via the recursive scheme described in Section 2. We compute 5 minute realized variance series from the data and it is taken to represent the "true" daily variance. We define $\Delta\mathcal{L}_t$ to be $\Delta\mathcal{L}_t := \left(\widehat{\varepsilon}_t^{St}\right)^2 - \left(\widehat{\varepsilon}_t^{G}\right)^2$, i.e. we subtract the squared error produced by the GARCH(1,1) model with Gaussian errors from the squared error produced by the GARCH(1,1) model with Student-t errors. Once $\Delta\mathcal{L}_t$ has been constructed, we apply our proposed two-step nonparametric procedure to estimate the corresponding time-varying mean and variance. Figure 8 below depicts $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ and $\widehat{\tau}_t = \widehat{\mu}_t/\widehat{\sigma}_t$ calculated according to (14).

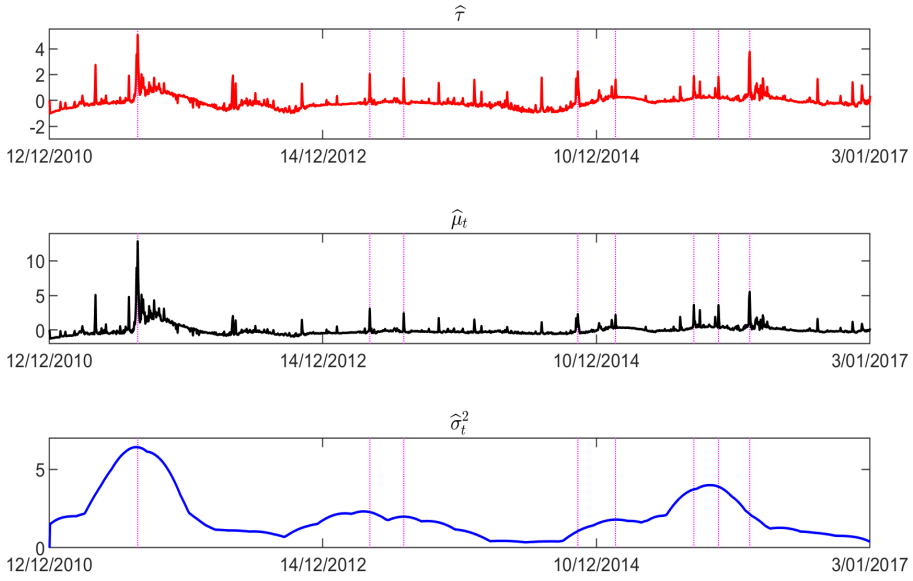

Figure 8: Plots of the estimates of $\widehat{\tau}_t$, $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ for IBM data, 2006-2017, using squared error loss and recursive forecasting scheme.

For our statistic that aggregates average past performance, recall that each corresponding $\widehat{\mu}_t$ is weighted by the inverse of the standard error $\widehat{\sigma}_t$. Hence whenever a spike occurs in the relative forecasting performance (represented by the violet dashed lines), the μ_t in those periods get down weighted. We next calculate the test statistic $\mathcal{S}_T$.

The value of the test statistic in this application example is $\mathcal{S}_T = -0.01$ with the corresponding $p-$value of 0.86. In fact, in this particular example the null of Equal Predictive Ability is not rejected at any levels of significance, indicating that the GARCH(1,1) model with normal errors is performing very similarly to the GARCH(1,1) model with Student-t errors over all sample. This is not surprising given the frequent overtaking of the relative forecasting performance of two models as is seen in Figure 1.

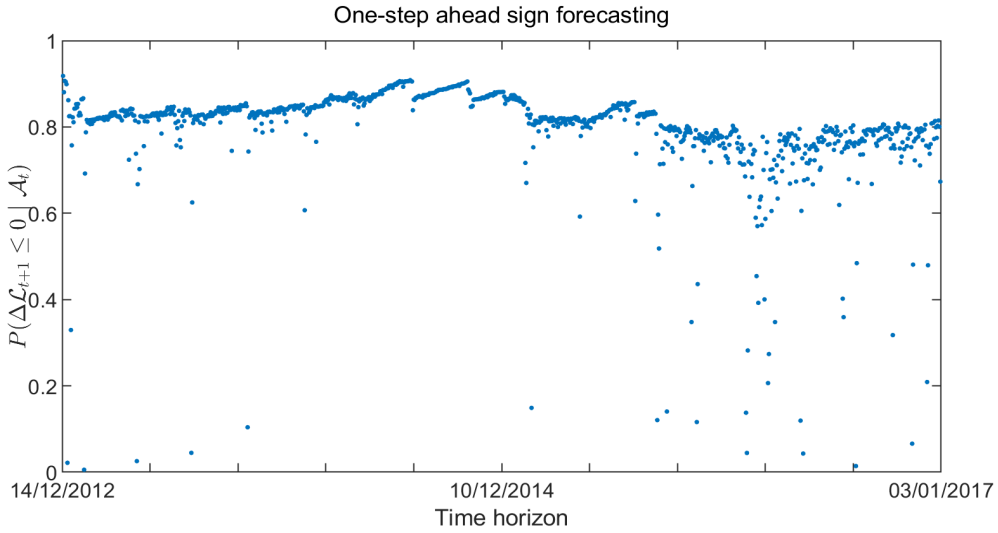

Figure 9: One-step ahead sign forecasting for the motivating example in the introduction.

Figure 9 presents the results of our pseudo out-of-sample sign forecasts. We see that primarily, the probability of the GARCH(1,1) model with Student-t errors outperforming the GARCH(1,1) model with normal errors is relatively high for most points in time with a few exceptions. Finally applying the criterion given by eq. (24) we get the value $\widehat{C} = -0.023$, indicating that our forecasted probabilities are on average overestimated by 2.3%. At the final point in the sample, we forecast a probability of 0.3129 with a corresponding forecast interval of $[0.2100,\ 0.3740]$. Interestingly, this probability does not conclude that GARCH(1,1) with Student-t errors should be selected for the next step forecasting. This highlights the randomness inherent in forecasting next period probabilities. Therefore, although our test of EPA test indicates that the two models have performed similarly for forecasting, our second approach clearly indicates that for the immediate next period the forecaster should select GARCH(1,1) with Gaussian errors if one is interested in selecting the model with the higher probability of outperforming.

7.2 Forecasting macroeconomic indicators

In this section we consider an application where several models are compared against each other in their ability to forecast multiple macroeconomic indicators. We consider the "balanced panel" of the dataset FRED-MD, consisting of 128 monthly economic time series measured over January, 1959 - December, 2019, and apply the transformations to the original series, as documented in

Appendix to the dataset[7]. We then use several forecasting models, described below, to construct 1-month ahead forecasts of four US macroeconomic variables: two real variables - industrial production (abbreviated IP) and real personal income less transfers (abbreviated RPI); and two price indices: consumer price index (abbreviated CPI) and producer price index (abbreviated PPI).

In what follows we denote by $\mathcal{Y}_t$ the variable being forecasted. The first forecasting method considers the full model $\mathcal{Y}_{t+1} = \alpha + \beta' \mathcal{X}_t + \varepsilon_{t+1}$, where $\mathcal{X}_t$ contains all 128 predictors from the FRED-MD dataset. To overcome multicollinearity in $\mathcal{X}_t$, we follow Giacomini and White (2006) and replace the groups of variables in $\mathcal{X}_t$ whose correlation is greater that 0.98 with their average. The new $\mathcal{X}_t$ contains 113 predictors. We next apply Elastic Net (abbreviated E-Net) as a more stabilized version of lasso which also allows for grouping effects, see Zou and Hastie (2005), to select the relevant predictors. Let $Z_t = (1, \mathcal{X}_t')'$, then Elastic Net estimator $\widehat{\nu}$ of the parameter vector ν is the solution of the following minimization problem:

$$\widehat{\nu} = \underset{\nu \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^{T} (\mathcal{Y}_{t+1} - \nu' Z_t)^2 + \lambda_1 \sum_{j=1}^{p} |\nu_j| + \lambda_2 \sum_{j=1}^{p} \nu_j^2 \right\},$$

where p is the dimension of the parameter vector ν and both tuning parameters $\lambda_1 > 0$ and $\lambda_2 > 0$ are selected by cross-validation.

To account for potential nonlinearity in the evolution of the forecasting performance we next consider nonlinear model for the conditional mean of $\mathcal{Y}_{t+1}$. More precisely, we let

$$\mathcal{Y}_{t+1} = m(\mathcal{Y}_t) + \varepsilon_t,$$

where the smooth function $m(\cdot)$ is estimated via local linear nonparametric regression. In particular, denote by $\mathcal{W}_t^y = (1, \mathcal{Y}_t - y)'$ and further let $K_h(x) := K(x/h)/h$ denote the kernel function with h denoting the corresponding bandwidth, which is subsequently chosen via multi-fold cross-validation. The local linear estimator is then obtained as the solution to the following minimization objective:

$$\widehat{\varpi}(y) = (\widehat{m}(y), \widehat{\partial_y m}(y))' = \underset{\varpi \in \mathbb{R}^2}{\arg\min} \sum_{t=1}^{T-1} K_h(\mathcal{Y}_t - y) \left(\mathcal{Y}_{t+1} - \varpi' \mathcal{W}_t^y \right)^2,$$

[7]The FRED-MD dataset is collected and constantly updated by the Federal Reserve Bank of St. Louise and can be found online with the following link. For the variables we consider in this paper, the transformations are as follows: the first log difference for RPI and IP variables; and the second log difference for CPI and PPI variables.

with the explicit solution given by

$$\widehat{\varpi}(y) = \left(\sum_{t=1}^{T-1} K_h \left(\mathcal{Y}_t - y \right) \mathcal{W}_t^y (\mathcal{W}_t^y)' \right)^{-1} \left(\sum_{t=1}^{T-1} K_h \left(\mathcal{Y}_t - y \right) \mathcal{W}_t^y \mathcal{Y}_{t+1} \right). \tag{31}$$

Next, we consider random forests model (abbreviated RF) proposed by Breiman (2001), which has recently been found to perform well in the context of the predicting macroeconomic indicators due to its ability to capture complex interaction structures in the data, see e.g. Medeiros et al. (2019). Random forests model is essentially a collection of several regression trees, each grown on a bootstrap sample of the original data. For a given tree the feature space is partitioned into a set of rectangles and within each rectangle a simple unconditional mean model is estimated. To avoid overfitting we first reduce the overall number of features $\mathcal{X}_t$ consisting of 128 predictors by extracting the factors $\widehat{F}_t$ from $\mathcal{X}_t$ by implementing principal component analysis. The number of factors j is chosen by applying Onatski (2009) test which results in $j = 8$ factors. The construction of the forecast from the random forests model then proceeds as follows. For each bootstrap sample $b = 1, \ldots, B$ a tree with M_b regions denoted by $R_{i,b}$, $i = 1, \ldots, M_b$ is grown by randomly selecting a subset of the original factors $\widehat{F}_t$ and determining the splits by minimizing the sum of the squared errors from a regression

$$\mathcal{Y}_{t+1} = \sum_{i=1}^{M_b} c_{i,b} \mathcal{I}_{i,b} \left(\widehat{F}_t; \theta_{i,b} \right), \quad \text{where} \quad \mathcal{I}_{i,b} \left(\widehat{F}_t; \theta_{i,b} \right) = \begin{cases} 1 & \text{if} \quad \widehat{F}_t \in R_{i,b} \left(\theta_{i,b} \right) \\ 0 & \text{otherwise} \end{cases},$$

where in the above $\theta_{i,b}$ characterizes the bth random forest and ith tree in terms of the splits that result in region $R_{i,b}$. We let $B = 1000$ and use block bootstrap of Politis and Romano (1994) with the block size chosen according to Politis and White (2004). Once all B random forests are constructed, the forecast of $\mathcal{Y}_{t+1}$ made at time t is obtained as follows:

$$\widehat{Y}_{t+1|t} = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{M_b} \widehat{c}_{i,b} \mathcal{I}_{i,b} \left(\widehat{F}_t; \widehat{\theta}_{i,b} \right).$$

As one of the benchmark models frequently used for forecasting macroeconomic indicators we consider a simple autoregressive model (denoted by AR) given by:

$$\mathcal{Y}_{t+1} = \gamma_0 + \gamma_1 \mathcal{Y}_t + \gamma_2 \mathcal{Y}_{t-1} + \cdots + \gamma_{p+1} \mathcal{Y}_{t-p} + \varepsilon_{t+1},$$

where the lag length p is selected by BIC. Finally, motivated by a successful use of the forecast

combinations, see e.g. Stock and Watson (2004), Aiolfi et al. (2010), we consider equal weight forecast (abbreviated EW) which is constructed as a simple average of the forecasts from the models described above. We use the squared error as our loss function for evaluating the forecasts and construct the time series of losses for $k = 1$ month according to the recursive scheme, described in Figure 2 and setting $\underline{T} = 100$. The loss differences are constructed as the loss of the column model minus the loss of the row model in all our tables.

For forecast evaluation we perform our equal predictive ability test $\mathcal{S}_T$ given in (15), results for which are presented in Table 5. A negative test statistic indicates that the column model had lower aggregated losses when compared to the row model, whereas a positive test statistic is indicative of higher aggregated losses of the column model versus the row model. We report the value of the test statistic $\mathcal{S}_T$ together with the associated $p-$value calculated via bootstrap described in section 5. The cases when we reject the null of EPA (13) are highlighted in bold. There are several observations that we can make. Firstly, for two real variables, industrial production and real personal income, random forests and equally weighted forecasts seem to perform the best as seen by the the frequency of a rejection of test when any given model is compared with the random forest model or equally weighted forecast. Moreover, in all of these rejection cases the positive value of the test statistic indicates that the random forests/equally weighted forecast has smaller value of the aggregated losses. This is perhaps not surprising since both models use averaging and in the context of (potentially) unstable environments this proves to be especially useful since it reduces the variance (and thus uncertainty about the forecasts themselves) and therefore *on average* they perform better. For the price indices however all models seem to perform on average quite similarly with the weak preference for random forests model for industrial production.

To proceed to the task of forecast selection, we now apply the sign forecasting methodology, described in Section 3.3, results for which are presented in Table 6. For each of the variables we report the forecasted probability $\mathbb{P}\left(\Delta\mathcal{L}_{T+1} \leq 0 | \mathcal{A}_T\right) =: \widehat{\mathcal{P}}\left(\mathbb{X}_{T+1}\right)$, where $\mathbb{X}_{T+1} := (1, \Delta\mathcal{L}_T)'$ at the end of the sample as well as the associated forecast interval which we simply denote by $[\widehat{FI}_l,\ \widehat{FI}_u]$. We say that the column model weakly outperforms the row model if the forecasted probability $\mathcal{P}_T(\mathbb{X}_{T+1}) > 0.5$ but $0.5 \in [\widehat{FI}_l, \widehat{FI}_u]$. Similarly, we say that that the column model weakly underperforms the row model if the forecasted probability $\mathcal{P}_T(\mathbb{X}_{T+1}) < 0.5$ but $0.5 \in [\widehat{FI}_l, \widehat{FI}_u]$.

Furthermore, it is possible to generate the entire ranking of models by aggregating the bilateral comparisons. Such a ranking will help the practitioner in selecting the best model (among those under comparison) to use for forecasting at time $T + 1$. In each case, we begin by identifying the strict bilateral out-performances and constructing a partial ranking of models based on only strict out-performances. The remainder of the ranking is then filled based on weak bilateral

out-performances. For example, in the case of industrial production variable, all but the random forest model comparisons are strict with the random forest model comparisons being all weak, which indicates that there is great uncertainty over the performance of the random forest model in period $T+1$. An important question is whether the overall constructed probability ranking is always transitive. In Appendix A2 we give a detailed discussion that this is indeed the case under fairly general conditions. In short, the ranking is transitive so long as the errors generating each of the losses are jointly Gaussian and are symmetric around zero. Under these assumptions the practitioner can construct the overall ranking by aggregating bilateral comparisons. In practice, however, there might be cases where bilateral weak out-performances generates a non-transitive ranking. This can happens due to a random estimation error. In such case the weak comparisons are considered equal. However for our example there are no cases of non-transitive loops for the strict out performance bilateral comparisons as one would expect.

The rankings for each of the forecasted variables are therefore as follows:

- Personal income: AR $>$ EW $\geq$ RF $\geq$ NP $\geq$ E-Net;
- Industrial production: EW $>$ AR $>$ E-Net $>$ NP $\geq$ RF;
- Producer price index: AR $=$ E-Net $\geq$ NP $\geq$ EW $>$ RF;
- Consumer price index: NP $>$ AR $\geq$ EW $\geq$ E-Net $>$ RF;

Specifically, autoregressive model dominates for personal income while for industrial production the equally weighted forecast is projected to perform the best next month. For the producer price index autoregressive model together with Elastic Net seem to be equally good and both dominate the rest of the models in terms of the projected performance next month. Finally, for the consumer price index the nonparametric regression is projected to perform much better than the rest of the models under comparisons.

Interestingly, the random forests model is no longer the best performing model. Thinking further about why our sign forecasting approach to ranking indicates that the random forests model is never the *best* model for next period forecasting, we note that this is contrast to our statistic measuring average performance where the forecast from random forests model together with equally weighted forecast had the best historical performance. In this particular period of $T+1$, we see that the forecast from the random forests model across all four series is always dominated by forecasts from other models. However across all periods historically, the model that outperforms the random forests model is most likely not the same for any of the four series. Our average performance metric, due in part to the weighting we employ, favors models that perform well consistently and

with low variance over models that perform very well some of the time but not so well the other times. Since random forests is essentially an improved modification of bagging and, as with any averaging, reduces the noise leading to reduction in variance, it is therefore perhaps not surprising that random forests and equally weighted forecasts are favored by our average performance metric. However forecasting one period ahead, one can expect to find a model that is likely to perform better. The emphasis on conservatism based on historical performance versus projected performance next period again emphasizes the difference between our two approaches.

Autoregressive model, on the other hand, performs quite well for all series. This is again perhaps not surprising since it incorporates time series information, which in the context of slowly changing environments is especially beneficial. Recall that another model, namely nonparametric regression, also incorporates time series information, however apart from consumer price index, it does not seem to be the favoured choice. One potential reason for this is that forecast from a nonparametric regression uses estimated regression function $\widehat{m}(\cdot)$ which is time-invariant. However, in slowly changing environments, depending on the period the forecast is made for, it is reasonable to assume that the regression function is also changing over time. Therefore a time-varying nonparametric regression might be an alternative model one might consider. That being said however, for consumer price index the forecast from a nonparametric regression is the dominant choice. This further highlights that in the context of changing environments the best model in terms of the forecasting performance is constantly changing and the average historical performance might not be the best predictor of the future forecasting performance in such environments.

8 Concluding remarks

In this paper we address the issue of forecast evaluation and forecast selection in unstable environments. Existing out-of-sample tests often suffer from low power, and in unstable environments they can generate spurious and potentially misleading results. We address the possibility of unstable environments explicitly and provide two tests in such a context. The first compares the overall historical performance of models and the second directly forecasts which model is more likely to outperform the next period. We demonstrate that our methodology performs well across a variety of applications, and our test has high power against a range of fixed and local alternatives.

Table 5: Results for the test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 1$ month.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		AR	E-Net	NP	RF	EW	AR	E-Net	NP	RF	EW	AR	E-Net	NP	RF	EW	AR	E-Net	NP	RF	EW
E-Net	$\mathcal{S}_T$	0.297					-0.258					-0.007					-0.029				
	p	**0.049**					**0.001**					0.851					0.454				
NP	$\mathcal{S}_T$	0.073	-0.028				0.042	0.120				-0.100	-0.023				-0.009	0.009			
	p	0.168	0.576				0.601	0.154				0.065	0.505				1.000	0.797			
RF	$\mathcal{S}_T$	0.159	0.029	0.051			0.226	0.277	0.121			0.033	0.116	0.058			-0.004	-0.002	0.001		
	p	**0.025**	0.671	0.344			**0.000**	**0.000**	0.074			0.341	**0.003**	0.114			0.928	0.963	1.000		
EW	$\mathcal{S}_T$	0.251	-0.105	0.014	-0.081	–	0.233	0.230	-0.006	-0.196	–	-0.013	0.010	0.052	-0.058	–	-0.013	-0.003	0.002	-0.006	–
	p	**0.048**	0.164	1.000	0.244	–	**0.000**	**0.001**	0.882	**0.002**	–	0.758	0.794	0.142	0.102	–	0.753	0.936	0.983	1.000	–

Note: Table reports the value of the test statistic $\mathcal{S}_T$ that corresponds to the null of equal predictive ability, see eq. (13), for horizon $k = 1$ month. The p-values are obtained via bootstrap procedure described in section 5. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, for Personal Income $\Delta\mathcal{L}_t^{\text{AR,E-Net}} = \mathcal{L}_t^{\text{AR}} - \mathcal{L}_t^{\text{E-Net}}$ for which the test statistics $S_T = 0.297$ (the positive sign indicating that E-Net is better) with the associated p-value of 0.049. The p-values in bold indicate rejection of the null (13) at the 5% level of significance.

Table 6: Sign forecasting for $\Delta\mathcal{L}_{T+1}$ for horizon $k = 1$ month.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		AR	E-Net	NP	RF	EW	AR	E-Net	NP	RF	EW	AR	E-Net	NP	RF	EW	AR	E-Net	NP	RF	EW
E-Net	$\widehat{\mathcal{P}}_T(\mathbb{X}_{T+1})$	0.620					0.514					0.500					0.526				
	$\widehat{FI}_l$	0.552					0.448					0.462					0.460				
	$\widehat{FI}_u$	0.719					0.587					0.537					0.585				
NP	$\widehat{\mathcal{P}}_T(\mathbb{X}_{T+1})$	0.812	0.497				0.806	0.546				0.513	0.513				0.099	0.398			
	$\widehat{FI}_l$	0.751	0.330				0.759	0.501				0.451	0.480				0.059	0.350			
	$\widehat{FI}_u$	0.937	0.683				0.916	0.607				0.573	0.549				0.198	0.452			
RF	$\widehat{\mathcal{P}}_T(\mathbb{X}_{T+1})$	0.577	0.450	0.476			0.542	0.521	0.533			0.542	0.604	0.540			0.562	0.645	0.615		
	$\widehat{FI}_l$	0.544	0.421	0.415			0.513	0.495	0.467			0.498	0.591	0.504			0.511	0.598	0.578		
	$\widehat{FI}_u$	0.625	0.494	0.553			0.570	0.544	0.582			0.583	0.615	0.578			0.607	0.686	0.647		
EW	$\widehat{\mathcal{P}}_T(\mathbb{X}_{T+1})$	0.717	0.364	0.404	0.495	–	0.463	0.510	0.401	0.466	–	0.502	0.516	0.505	0.457	–	0.519	0.447	0.637	0.395	–
	$\widehat{FI}_l$	0.682	0.299	0.309	0.446	–	0.450	0.491	0.319	0.447	–	0.485	0.494	0.478	0.443	–	0.457	0.380	0.593	0.364	–
	$\widehat{FI}_u$	0.765	0.493	0.509	0.528	–	0.480	0.527	0.449	0.485	–	0.518	0.540	0.531	0.472	–	0.578	0.516	0.677	0.427	–

Note: Table reports the results of the sign forecasting for $\Delta\mathcal{L}_{T+1}$ for the forecast horizon $k = 1$ month. Recall $\mathbb{X}_{T+1} = (1, \Delta\mathcal{L}_T)'$ and $\widehat{\mathcal{P}}_T(\mathbb{X}_{T+1}) := \widehat{\mathbb{P}}(\Delta\mathcal{L}_{T+1} \leq 0|\mathcal{A}_T)$, i.e. the forecasted probability at the very end of the sample given past loss differences. $\widehat{FI}_u$ and $\widehat{FI}_l$ denote the upper and the lower bounds of the forecast interval respectively, such that $\widehat{\mathcal{P}}_T(\mathbb{X}_{T+1}) \in [\widehat{FI}_l, \widehat{FI}_u]$. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, for Personal Income $\Delta\mathcal{L}_t^{\text{AR,E-Net}} = \mathcal{L}_t^{\text{AR}} - \mathcal{L}_t^{\text{E-Net}}$, for which $\widehat{\mathcal{P}}_T(\mathbb{X}_{T+1}) = 0.620$ with the corresponding forecast interval [0.552, 0.719].

References

Aiolfi, M., Capistrán, C., and Timmermann, A. (2010). Forecast combinations. *CREATES research paper*, (2010-21).

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217.

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

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brockwell, P. J. and Davis, R. A. (2006). *Time series: theory and methods*. Springer Series in Statistics. Springer, New York. Reprint of the second (1991) edition.

Cai, Z., Fan, J., and Li, R. (2000a). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902.

Cai, Z., Fan, J., and Yao, Q. (2000b). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1):85–110.

Clark, T. E. and McCracken, M. W. (2005). Evaluating direct multistep forecasts. *Econometric Reviews*, 24(4):369–404.

Corradi, V. and Swanson, N. R. (2007). Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *International Economic Review*, 48(1):67–109.

Dahlhaus, R. and Polonik, W. (2009). Empirical spectral processes for locally stationary time series. *Bernoulli*, 15(1):1–39.

Dahlhaus, R., Richter, S., and Wu, W. B. (2019). Towards a general theory for nonlinear locally stationary processes. *Bernoulli*, 25(2):1013–1044.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 31(2):393–427.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Hall, P. and Heyde, C. (1980). Martingale limit theory and its applications,. *Academic Press, New York*.

Hall, P. and Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, 86(415):665–672.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Hansen, P. R. (2010). A winner's curse for econometric models: on the joint distribution of in-sample fit and out-of-sample fit and its implications for model selection. *Research Paper*, pages 1–39.

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

Hansen, P. R. and Timmermann, A. (2012). Choice of sample split in out-of-sample forecast evaluation.

Hirano, K. and Wright, J. H. (2017). Forecasting with model uncertainty: Representations and risk reduction. *Econometrica*, 85(2):617–643.

Inoue, A. and Kilian, L. (2005). In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews*, 23(4):371–402.

Li, J. and Patton, A. J. (2018). Asymptotic inference about predictive accuracy using high frequency data. *Journal of Econometrics*, 203(2):223–240.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

McCracken, M. W. (2000). Robust out-of-sample inference. *Journal of Econometrics*, 99(2):195–223.

Medeiros, M. C., Vasconcelos, G. F., Veiga, A., and Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, pages 1–22.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313.

Politis, D. N. and White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric reviews*, 23(1):53–70.

Pourahmadi, M. (1992). Arma approximations and representations of a stationary time series. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 54(2):235–241.

Richter, S. and Dahlhaus, R. (2019). Cross validation for locally stationary processes. *Ann. Statist.*, 47(4):2145–2173.

Richter, S. and Phandoidaen, N. (2020). Empirical processes for locally stationary processes using the functional dependence measure. *technical report*.

Rossi, B. and Inoue, A. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economic Statistics*, 30(3):432–453.

Stinchcombe, M. B. and White, H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric theory*, 14(3):295–325.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of forecasting*, 23(6):405–430.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, pages 1067–1084.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154.

Wu, W. B. and Zhou, Z. (2011). Gaussian approximations for non-stationary multiple time series. *Statistica Sinica*, 21(3):1397–1413.

Zhang, D. and Biao Wu, W. (2015). Gaussian Approximation for High Dimensional Time Series. *arXiv e-prints*, page arXiv:1508.07036.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Appendix A1.

Consider the following hypothetical example. Assume the true data generating process for $\{y_t\}_{t=1}^{T}$ follows an AR(1) process:

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.(0, \sigma^2), \qquad |\rho| < 1.$$

Furthermore, assume one uses two simple models to forecast y_t one-step ahead:

- Model A uses $\widehat{y}_{t+1|t} = 0 \quad \text{for all} \;\; t = 1, \cdots, T$ as a forecast for y_{t+1};
- Model B uses $\widehat{y}_{t+1|t} = 0.1 \quad \text{for all} \;\; t = 1, \cdots, T$ as a forecast for y_{t+1};

Suppose that the forecaster uses the mean squared error (MSE) as the loss to assess the quality of the forecasts, i.e.

$$\mathcal{L}_t^A = \mathbb{E}\left[\left(y_{t+1} - \widehat{y}_{t+1|t}\right)^2 | \mathcal{F}_t\right] = \rho^2 y_t^2 + \sigma^2,$$

and

$$\mathcal{L}_t^B = \mathbb{E}\left[\left(y_{t+1} - \widehat{y}_{t+1|t}\right)^2 | \mathcal{F}_t\right] = \rho^2 y_t^2 + \sigma^2 - 0.2\rho y_t + 0.01,$$

and therefore

$$\Delta\mathcal{L}_t^{AB} = \mathcal{L}_t^A - \mathcal{L}_t^B = 0.01 - 0.2\rho y_t. \tag{32}$$

From (32) it then follows that

$$\begin{cases} \Delta\mathcal{L}_t^{AB} \leq 0 & \text{if} \quad y_t > 0.05/\rho, \\ \Delta\mathcal{L}_t^{AB} > 0 & \text{if} \quad y_t < 0.05/\rho. \end{cases}$$

This example highlights the possibility of relative performance to be changing over time. Even in a world where the data generating process is stationary, and the models used to forecast are likewise stable, it is still possible for relative performance between models to not be constant.

Appendix A2.

In this Appendix we show that when applying our forecasted probability in Theorem 5 for forecast selection among models $1, \cdots, n$, true probability dominance is indeed transitive under conditions we specify below. It is sufficient to show the argument for 3 generic models. Therefore let A, B, C be 3 generic models used for forecasting. Let $\mathcal{L}_t^i$ be forecast losses of models $i = A, B, C$, and let

$$\Delta\mathcal{L}_t^{AB} =: \mathcal{L}_t^A - \mathcal{L}_t^B, \qquad \Delta\mathcal{L}_t^{BC} =: \mathcal{L}_t^B - \mathcal{L}_t^C, \qquad \Delta\mathcal{L}_t^{AC} =: \mathcal{L}_t^A - \mathcal{L}_t^C.$$

Assume also that all three time series $\mathcal{L}_t^i$, $i \in \{A, B, C\}$ are generated with different mean-zero innovations ε_t^i, $t \in \mathbb{Z}$ and follow a time-varying autoregressive model (tvAR) model. Finally, we assume that ε_t^i, $i = A, B, C$ are either independent or jointly Gaussian distributed, i.e.

$$\varepsilon_{T+1}^{ABC} := \begin{pmatrix} \varepsilon_{T+1}^A \\ \varepsilon_{T+1}^B \\ \varepsilon_{T+1}^C \end{pmatrix} \sim N(0, \Sigma), \tag{33}$$

where Σ is an arbitrary positive semi-definite matrix. Let $\mathcal{A}_T = \sigma(\varepsilon_T^A, \varepsilon_T^B, \varepsilon_T^C, \varepsilon_{T-1}^A, \varepsilon_{T-1}^B, \varepsilon_{T-1}^C, ...)$ and denote further

$$p_{AB} := \mathbb{P}\left(\Delta\mathcal{L}_{T+1}^{AB} \leq 0 | \mathcal{A}_T\right), \tag{34}$$

$$p_{BC} := \mathbb{P}\left(\Delta\mathcal{L}_{T+1}^{BC} \leq 0 | \mathcal{A}_T\right) \tag{35}$$

and

$$p_{AC} := \mathbb{P}\left(\Delta\mathcal{L}_{T+1}^{AC} \leq 0 | \mathcal{A}_T\right). \tag{36}$$

To establish transitivity we then need to show that the following holds:

$$\textit{Statement}: \quad \text{if} \quad p_{AB} > 1/2 \quad \text{and} \quad p_{BC} > 1/2 \quad \Longrightarrow \quad p_{AC} > 1/2. \tag{37}$$

The statement in (37) implies that the probability ranking is indeed transitive.

Proof of (37). Consider models A and B first. For model A we can write

$$\mathcal{L}_{T+1}^A = \mu^A + Z^A,$$

where $\mu^A = \mathbb{E}[\mathcal{L}^A_{T+1}|\mathcal{A}_T]$ and $Z_A = \mathcal{L}^A_{T+1} - \mathbb{E}[\mathcal{L}^A_{T+1}|\mathcal{A}_T]$. More precisely, the following decomposition holds:

$$\mathcal{L}^A_{T+1} = \sum_{k=0}^{\infty} W^A_k \varepsilon^A_{T+1-k} + \mu^A = \underbrace{W^A_0 \varepsilon^A_{T+1}}_{=Z^A} + \underbrace{\sum_{k=1}^{\infty} W^A_k \varepsilon^A_{T+1-k} + b^A}_{=\mu^A} = \mu^A + Z^A, \tag{38}$$

where W^A_k, b^A depend on the corresponding parameters $\{\rho^A_s\}^T_{s=T+1-k}$, $\{\sigma^A_s\}^T_{s=T+1-k}$ and for b^A we take $k = \infty$. Similar decomposition holds for $\mathcal{L}^B_{T+1}$ and $\mathcal{L}^C_{T+1}$. For models A, B consider the difference between their stochastic parts:

$$Z^A - Z^B = W^A_0 \varepsilon^A_{T+1} - W^B_0 \varepsilon^B_{T+1} = (W^A_0, -W^B_0, 0)\varepsilon^{ABC}_{T+1} \sim N\Big(0, (W^A_0, -W^B_0, 0)\Sigma(W^A_0, -W^B_0, 0)'\Big). \tag{39}$$

From the representation (38) it follows that:

$$\frac{1}{2} < p_{AB} = \mathbb{P}(\Delta\mathcal{L}^{AB}_{T+1} \leq 0|\mathcal{A}_T) = \mathbb{P}(\mu^A - \mu^B \leq Z^A - Z^B|\mathcal{A}_T), \tag{40}$$

and from (39) it further follows that $Z^A - Z^B$ has median 0 conditionally on $\mathcal{A}_T$. Then from (40) we further get that

$$\mu^A - \mu^B < 0$$

Similarly, if $Z^B - Z^C$ has median 0 conditionally on $\mathcal{A}_T$, we get from $\frac{1}{2} < p_{BC}$ that

$$\mu^B - \mu^C < 0.$$

We therefore have $\mu_A - \mu_C = (\mu^A - \mu^B) + (\mu^B - \mu^C) < 0$. Finally, if $Z^A - Z^C$ has median 0 conditionally on $\mathcal{A}_T$, then

$$p_{AC} = \mathbb{P}(\mu^A - \mu^C \leq Z^A - Z^C|\mathcal{A}_T) > \frac{1}{2}.$$

Combining all of the above we therefore get the implication:

$$p_{AB} > \frac{1}{2}, \quad p_{BC} > \frac{1}{2} \quad \Rightarrow \quad p_{AC} > \frac{1}{2} \qquad ,$$

which establishes the proof of (37). $\square$