Forecast Evaluation Tests - a New Approach

Ekaterina Smetanina
University of Chicago *

November 1, 2018

Abstract

Out-of-sample tests are widely used for evaluating and selecting between models' forecasts in economics and finance. Underlying these tests is often the assumption of constant relative performance between competing models, however this is invalid for many practical applications. In a world of changing relative performance, previous methodologies give rise to spurious and potentially misleading results, an example of which is the well-known "splitting point problem". We propose a new two-step methodology designed specifically for forecast evaluation 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.

1 Introduction

In a non-experimental field such as economics, an important way to judge competing models is by comparing their relative forecasting performance. Out-of-sample tests, which are broadly variations of the Diebold-Mariano test from their 1995 seminal paper, are currently the benchmark for comparing models' forecasting performance. However in an unstable environment where relative performance between models can change over time, these tests can generate spurious and

*Email: e.smetanina@chicagobooth.edu. I would like to thank my supervisor Oliver Linton for his many valuable suggestions and comments. I am indebted to Andrew Patton for our many discussions during my visit at Duke that were essential to this paper. Special thanks go to Tim Bollerslev for his many suggestions that improved this paper and for his encouragement. I would also like to thank River Chen, Dennis Kristensen, Jason Lu, Alexei Onatski, Rasmus Søndergaard Pedersen, Anders Rahbek, Barbara Rossi, George Tauchen, Andrey Temlyakov, Steve Thiele and seminar participants at Cambridge econometrics workshop for useful discussions and comments. Financial support from Cambridge-INET Institute is gratefully acknowledged.

potentially misleading results. An example of this is the well-known "splitting point problem" for out-of-sample forecast evaluation tests. The sample 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] However beyond this broad guideline, the choice of the splitting point is somewhat arbitrary and left to the discretion of the practitioner. This becomes problematic in a world of changing relative performance. In such a setting, one model may outperform its competition for some *window* of data, but underperform for a different window. Because the splitting point controls the window of data used for evaluation, different splitting points imply different evaluation windows, and the results of these out-of-sample tests may change or completely reverse depending on this arbitrary choice. Consequently, it opens up the possibility of data-mining for practitioners to select favourable splitting points that support their desired hypothesis. Despite these drawbacks, out-of-sample tests are still often preferred to their alternative, in-sample tests. In-sample tests use all available data for both estimation and evaluation, hence they do not suffer a power loss.[2] However, in-sample tests are prone to spurious results due to over-fitting. Specifically, the ability of a model to fit the data is not necessarily connected to the model's forecasting performance. In fact, Hansen (2010) shows that often, a model's in-sample fit is inversely related to its forecasting performance. See Hansen and Timmermann (2015) for a discussion on the matter.

To demonstrate the two main problems of the existing out-of-sample tests, namely low power and the arbitrary dependence on the splitting point, consider the following real-world example. We forecast the daily variance of IBM returns spanning 2006-2016 using two models: GARCH(1,1) model with Standard normal errors and GARCH(1,1) model with Student-t errors. Each point on the graph below, together with the critical values for the test statistic under the null hypothesis, represents a Diebold-Mariano-type test at that particular splitting point. Variance forecasts are produced via a standard recursive scheme, 5 minute realised volatility 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 5% critical values, across a range of splitting point choices, such that the out-of-sample data starts in December 2010, leaving at most 1500 data points for evaluation.

[1] See Diebold (2013) 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] Hansen (2008) proposes a methodology for optimal weight selection of forecasts in nested linear models, but it is not clear whether this method can be extended to nonlinear and non-nested models.

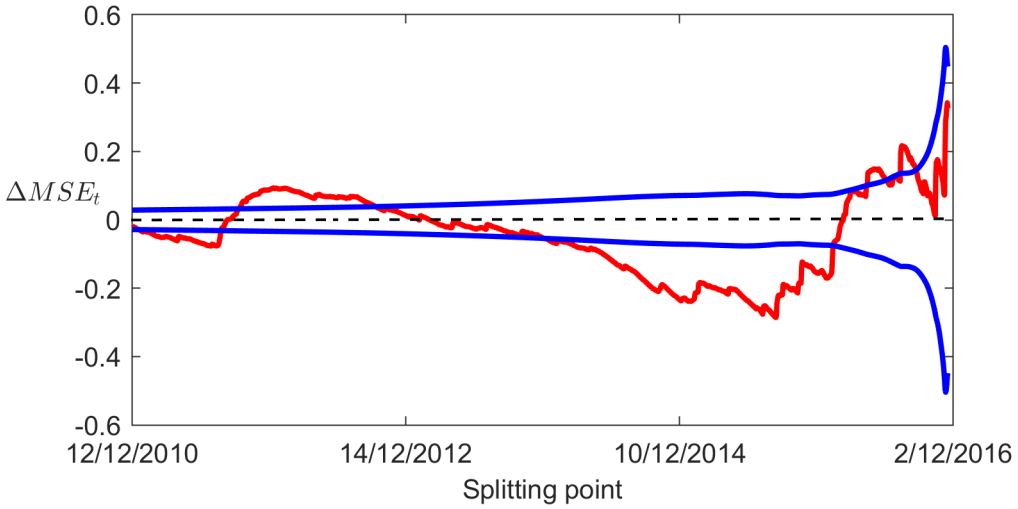


Figure 1: The figure displays the *difference* in MSE calculated for GARCH(1,1)-*N* and GARCH(1,1)-*St-t* for IBM data, 2006-2016 using recursive forecasting scheme. The MSE for each of the models is taken with respect to 5min RV calculated from the data.

This example is representative of many practical applications. For many plausible choices of the splitting point, the test is not powerful enough to reject in either direction. 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. As the practitioner is often not obliged to show results for all splitting points, in this example it is possible to select any desired outcome.

In this paper, we propose a new forecast evaluation and selection methodology that is designed explicitly for a world of changing relative performance, where constant relative performance is now a special case. In a changing world, the task of forecast evaluation versus forecast selection do not necessarily overlap. A practitioner may be interested in the question of which model performed better in the past, but it is possible that a different model will outperform for future forecasts. For the purpose of forecast selection, we propose to rank models using two alternative approaches. First, we consider ranking models based on their average past performance. Our motivation is that if past performance is indicative of future performance, then a practitioner may want to select models based on average past performance. Importantly, our methodology for evaluating average past performance is robust to the situation of unstable environments, see the discussion in Section 2. Second, we consider ranking 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. A practi-

tioner may then select a model for forecasting next period based on which model is more likely to outperform.

Our overall methodology is summarised 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 utilise these estimates to compare and rank competing models using our two proposed approaches. Our statistic measuring average performance aggregates the time-varying means normalised by its time-varying standard deviation, 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. We provide tests for Equal Predictive Ability (EPA) and for Superior Predictive Ability (SPA), which we use to compare and rank models respectively. For our second approach, we construct forecasted probabilities for how likely one model will outperform another based upon the estimates from step one. In addition, we construct *forecast intervals*, which measure the confidence interval of the forecasted probability. In general, our two approaches will often select the same model for forecasting, however this is not always the case. In some applications, a model that performed on average worse over the overall sample may suddenly outperform for a short window towards the end of the sample. In such a situation, our first approach will not select the aforementioned model, however our second approach will. As our second approach is concerned only with the next period performance, the resulting ranking is noisier and subject to change depending on the sample. If a practitioner was interested in selecting a forecast for the next period only, then our second approach should be more relevant, however it is still possible that our forecasted probability is inconclusive while past performance can be accurately compared. In general, we do not advocate one particular approach over the other. Instead we present both approaches and leave the choice to the researcher on which is more relevant for their application. We believe that it shall often be the case that both are insightful, as they address the question of forecast selection from different perspectives, and are both meaningful in their own right.

Related to this work is the paper by Giacomini and White (2006), who develop a conditional version of the unconditional EPA 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 the test on a set of covariates, enabling a test for 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 in general their test can be thought of as a test for whether we

are in a changing world or a constant world (where a rejection of their test is indicative of changing relative performance). Beyond this, and as is acknowledged by the authors, their methodology is not designed for the selection of models for forecasting. Their test informs only whether we reject the conditional null of Equal Predictive Ability, and cannot reliably indicate which model is better in the event of rejection.

Also related to this work, Giacomini and Rossi (2010) were the first to address the problem of what they call "unstable environments", i.e. a world of changing relative performance. They do this by comparing instead the *local* relative performance between two models. This acknowledges directly the possibility for the relative performance of models to change over time, and it is an important step in the literature towards addressing this issue. For the purpose of forecast selection however, the methodology of Giacomini and Rossi (2010) has two shortcomings. First, by focusing their attention on local relative performance, they can only use local data for evaluation, which likely leads to their test having low power. Second, and more importantly, their methodology can only inform the practitioner as to which model was better at a particular point in the past. It is not informative as to which model is better for future forecasts, which is the question of interest to a practitioner who is interested in model selection for future forecasts.[3]

In addition there are the papers by Inoue and Rossi (2012) and by Hansen and Timmermann (2010). They look to address the splitting point problem by bringing to attention the potential data mining of practitioners who search for favourable splitting points. They propose to explicitly mine over all splitting points for the one that is the most favourable for the null hypothesis, and they reevaluate their test statistic at this splitting point with adjusted critical values that account for the bias introduced by mining. However, underlying their test is still the assumption of constant relative performance, and they retain the need for finding the one optimal splitting point. In addition, by selecting only the most critical single splitting point they again leverage the result of their test on a particular window of evaluation. As a result, in a world of changing relative performance the mining over splitting points can lead to spurious results. Applied to an example such as the one presented earlier, it may be possible that each model is favoured over the other, where their test selects different splitting points depending on which conclusion they mine for.

The rest of the paper is organised as follows. In section 2 we further discuss the 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

[3]This is except in the situation of a clear one-time reversal in relative performance, where one model is clearly better after a sharp structural break. Giacomini and Rossi (2010) consider a version of their test which addresses this scenario. In this special case their methodology can determine which model is superior for future forecasts.

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.

Throughout this paper, the following notation is used. Let $f(x)$ be any function from $\mathbb{R}^d \to \mathbb{R}$, then $\dot{f}(x) = \partial f(x)/\partial x$ and $\ddot{f}(x) = \partial^2 f(x)/\partial x^2$ denote the first and the second derivatives with respect to the argument x respectively. Moreover, $\|g(x)\|_2 = \left(\int |g(x)|^2 dx\right)^{\frac{1}{2}}$ and $\|g(x)\|_2^2 = \int |g(x)|^2 dx$. For a generic non-singular matrix A, A^T denotes its transpose; for a square matrix B, $\mathrm{tr}\,(B)$ denotes its trace. For any given vector a, $\mathrm{diag}(a)$ creates a diagonal matrix with elements of a along the main diagonal. 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 Model ranking in unstable environments

We start with the framework of two models, although the methodology can be further generalized to many models via pairwise comparisons. Let $\mathcal{A}, \mathcal{B}$ be two models, $\{y_t\}_{t=1}^T$ be the original data, $\widehat{\beta}_t^{\mathcal{A}}, \widehat{\beta}_t^{\mathcal{B}}$ denote the parameter estimates of two models at time t, which reflect the models as well as the estimation procedures[4]. We denote the difference in forecast losses at time $t+k$ by $\Delta\mathcal{L}_{t+k}^{\mathcal{AB}} = \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{A}}\right) - \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{B}}\right)$, where $\mathcal{L}(\cdot)$ denotes the loss function chosen by the forecaster[5]. In what follows we shall refer to $\Delta\mathcal{L}_{t+k}^{\mathcal{AB}}$ as $\Delta\mathcal{L}_{t+k}$ for simplicity of notation. Note that in general the loss function will be affected by the estimation error. However, given our expanding estimation scheme for constructing the losses, in what follows it is reasonable to assume that the estimation error vanishes asymptotically. We also explicitly acknowledge that the mean and variance of $\Delta\mathcal{L}_{t+k}$ might 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 the set of possible regressors. We make our notation general, so that μ_{t+k} can potentially depend on a set of regressors, in which case μ_{t+k} denotes the conditional mean of the loss difference at time $t+k$. Note that here we refer to $\mathbb{X}_t$ as the set of possible regressors in modelling the conditional mean of the loss differences, and *not* the regressors used to construct forecasts. A natural example of $\mathbb{X}_t$ are the lags of $\Delta\mathcal{L}_t$ as in Giacomini and White (2006). However, it is often the case that we are interested in the unconditional mean which is obtained by setting $\mathbb{X}_t = \varnothing$ for all t. An example of the latter is the commonly applied Diebold-Mariano type tests.

In a world of *constant relative* forecasting performance, i.e. $\mu_{t+k} = \mu$, for all $\mathbb{X}_t$ and all $t \in$

[4]We generically refer to $\widehat{\beta}^{\mathcal{A}}$,$\widehat{\beta}^{\mathcal{B}}$ as parameter estimates, however, depending on whether a parametric, semiparametric or nonparametric model is used $\widehat{\beta}$ will be any estimator used to construct the forecasts.

[5]Throughout the paper we denote the forecast horizon by k as h will be reserved for denoting the bandwidth.

$\{1, \cdots, T\}$, the task of ranking competing models is simple. Specifically, if $\mu < 0$ we say that model $\mathcal{A}$ is better than model $\mathcal{B}$ and vice versa. In such a world, the conclusion of the standard out-of-sample tests does not depend on the evaluation window and hence neither on the choice of the splitting point, although with a too short evaluation window the test shall suffer from low power. Indeed current methodologies explicitly assume constant relative performance, including the tests by Diebold and Mariano (1995), West (1996), White (2000), Clark and McCracken (2001, 2005), McCracken (2000, 2007), Hansen (2005), Corradi and Swanson (2007), Hansen et al. (2011), Inoue and Rossi (2012) and by Hansen and Timmermann (2010) and Li and Patton (2017), among others.

However in an unstable environment, i.e. a world of changing relative forecasting performance, the task of model ranking becomes far less obvious. Note that changing relative performance can occur even when the data generating process is stationary (see the example presented in Appendix A). For the example provided in Figure 1, the two competing models often overtake each other depending on the evaluation window, and there is no clear dominant choice using any of the previous methodologies. Yet, the question of how to select a model for next period forecasting in such a changing world is of the utmost importance for practitioners.

Giacomini and Rossi (2010) is the first paper in the literature that looks to address the issue of changing relative performance. They propose to rank models at each moment in time by their *local* relative performance. Specifically, for a given forecast horizon k they propose to measure the local mean of loss differences μ_{t+k} as a sample mean centered around a window of a (fixed) size m, i.e.

$$\widehat{\mu}_{t+k} := \frac{1}{m} \sum_{s=t-m/2}^{t+m/2+1} \Delta\mathcal{L}_{s+k},$$

such that we can test the following null hypothesis:

$$\mathbb{H}_0^{GR}: \quad \mu_{t+k} < 0 \quad \forall\, t \in \{1, \cdots, T-k\} \qquad vs. \qquad \mathbb{H}_1^{GR} \quad \mu_{t+k} \geq 0 \quad \forall\, t \in \{1, \cdots, T-k\}.$$

Giacomini and Rossi's approach is insightful, however as discussed before it still has a few shortcomings if we want to decide with which model we would like to forecast with. We saw previously that each model will outperform the other at some points in time, hence their test shall reject $\mathbb{H}_0^{GR}$ for some t's and accept it for some other t's. Although very useful as an ex-post investigation of the past performance of the competing models, it does not inform the practitioner which model to select for *future* forecasts.

The purpose of this paper is to provide a methodology to inform model selection in an un-

stable environment. We do this using two approaches. Our first approach is to evaluate past performance, and select models for future forecasts based upon which model performed better in the past. In an unstable environment, the model that outperformed in the past does not necessarily outperform in the future, however it is often the case that past performance is all one can reliably use to compare models. Indeed, the previous forecast evaluation methodology explicitly assumes constant relative performance, and the rationale behind using these tests to select models for forecasting is the same as ours in this first approach. Our main contribution in this dimension is to make our evaluation methodology robust to unstable environments, because as we witnessed in the motivating example the previous out-of-sample tests can have various problems in such a situation. Our second approach is to forecast the probability that one model shall outperform the other 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 of the next period forecast loss difference as opposed to its level as levels can depend on arbitrary factors such as a factor of scaling to the loss function, and 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 indeed the sign for a particular comparison is the same across all symmetric loss functions. This latter approach is conceptually more appropriate to an unstable environment, however in general forecasting this probability is noisier and subject to change depending on the sample. Another limitation of our second approach is that a practitioner may not want to update in every period the model that they choose to forecast with. In which case, they may want to select the model that performed on average better over the entire history, versus the model that is likely to perform better in the immediate next period. As mentioned in the introduction, we believe both approaches to be informative in different situations.

Our second approach is more immediate and self evident, however for the first approach there are potentially many ways one could compare past performance. Importantly, we would like the new test to be robust to the various problems of the previous methodology. Our first innovation is to use the (near) entire series of forecast losses to construct our statistic, which extends the evaluation to (nearly) the whole sample. This makes our test more powerful, and it makes the result of our test no longer reliant on the arbitrary choice of the sample splitting point. Our metric by which we compare past performance is defined as the sum of weighted expected relative forecast losses across the entire sample, where the weighting is given by the time-varying standard deviation of the forecast losses at that point in time. We offer our weighting as the second innovation. Our metric to measure past performance belongs to a general class of metrics, which encompasses

most of the current methodologies (the general class is also formally defined below). Although our particular metric is just one of many, we argue that it has several attractive features that make it insightful to consider. With our weighting, the forecast losses at the beginning of the sample which come with the largest estimation error are naturally down weighted. Moving towards the end of the sample, more data is used for the estimation leading to lower estimation error, therefore later losses receive a larger weight. Our metric is the first to accommodate this explicitly, i.e. that different forecasting loss differences shall be weighted differently. The motivation for our weighting is to reduce the variance of the statistics, which leads to higher power.

We first define the general class of metrics and its associated ranking by the following definition:

DEFINITION 1. *Let $\mathcal{M}$ be a collection of models under consideration and $\mathcal{M} \times \mathcal{M}$ be the set of all possible model combinations from $\mathcal{M}$ and $\mathcal{A}, \mathcal{B} \in \mathcal{M}$. Let $\Delta\mathcal{L}_{t+k} = \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{A}}\right) - \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{B}}\right)$ and $\mu_{t+k} \equiv \mathbb{E}\left[\Delta\mathcal{L}_{t+k} | \mathbb{X}_t\right]$, where $\mathbb{X}_t$ denotes the set of possible regressors. Define the following binary relation on $\mathcal{M}$:*

$$\mathcal{R}_T = \left\{ (\mathcal{A}, \mathcal{B}) \middle| \sum_{t=1}^{T-k} w_{t+k}\mu_{t+k} \leq 0, \mathcal{A}, \mathcal{B} \in \mathcal{M} \right\} \subseteq \mathcal{M} \times \mathcal{M},$$

where $\sum_t w_{t+k}\mu_{t+k}$ is the metric, and $\{w_{t+k}\}_{t=1}^{T-k} \geq 0$ is a set of non-negative weights. We say that model $\mathcal{A}$ is currently superior to model $\mathcal{B}$ iff $(\mathcal{A}, \mathcal{B}) \in \mathcal{R}$.

Remark 1. Note that the above general class of rankings encompasses rankings from the following standard tests (and any variations thereof):

- *Diebold-Mariano (1995) test* if we set $w_{t+k} = 1$ for all $t \geq S$ where S is a splitting point of choice and $\mu_{t+k} = \mu$ for all $t \in T$; and $\mathbb{X}_t = \varnothing$;
- *Giacomini and White (2006) test* if we set $w_{t+k} = 1$ for all $t \geq S$ where S is a splitting point of choice and any $\mathbb{X}_t \in \mathcal{F}_{t-1}$, where $\mathcal{F}_{t-1}$ is the information set available to the forecaster at time $t-1$;
- *Giacomini and Rossi (2010) test* if we set $w_{t+k} = 1$ for all $t \in [S + k - \frac{n}{2} + j, S + k + \frac{n}{2} + j]$, where $j = 0, \cdots, T - n + 1 - S - k$, and n is a fixed size of the rolling window and S is the original splitting point of choice; and $\mathbb{X}_t = \varnothing$.

The metric by which we base our statistic is a special case of the general ranking defined above. The definition for our specific metric is as follows:

DEFINITION 2. *Let $\mathcal{M}$ be a collection of models under consideration and $\mathcal{M} \times \mathcal{M}$ be the set of all possible model combinations from $\mathcal{M}$ and $\mathcal{A}, \mathcal{B} \in \mathcal{M}$. Define the following binary relation on $\mathcal{M}$:*

$$\mathcal{R}_T = \left\{ (\mathcal{A}, \mathcal{B}) \middle| \sum_{t=1}^{T-k} w_{t+k} \mu_{t+k} \leq 0, \mathcal{A}, \mathcal{B} \in \mathcal{M} \right\} \subseteq \mathcal{M} \times \mathcal{M},$$

where $\sum_t w_{t+k} \mu_{t+k}$ is the metric, and $\{w_{t+k} \propto \phi_{t+k}/\sigma_{t+k}\}_{t=1}^{T-k} > 0$ is a set of non-negative weights, where $\sigma_{t+k}^2 = var\left(\Delta \mathcal{L}_{t+k} | \mathbb{X}_t\right)$ and ϕ_{t+k} is a set of (deterministic) given weights. We say that model $\mathcal{A}$ is currently superior to model $\mathcal{B}$ iff $(\mathcal{A}, \mathcal{B}) \in \mathcal{R}$.

Remark 2. An example of ϕ_{t+k} is $\phi_{t+k} = \mathbb{1}\left(t + k \in I\right)$, where I could be a period of interest, e.g. recession times.

We now introduce the null hypothesis for our first approach formally. Firstly, we have the following null of Equal Predictive Ability (EPA$_w$):

$$\mathbb{H}_0^{(1)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k} \mu_{t+k} = 0 \qquad vs. \qquad \mathbb{H}_1^{(1)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k} \mu_{t+k} \neq 0, \tag{1}$$

and the following null of the Superior Predictive Ability (SPA$_w$):

$$\mathbb{H}_0^{(2)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k} \mu_{t+k} \leq 0 \qquad vs. \qquad \mathbb{H}_1^{(2)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k} \mu_{t+k} > 0, \tag{2}$$

where under definition 1, w_{t+k} is a set of weights s.t $\{w_{t+k}\} \geq 0$, and under definition 2, w_{t+k} is proportional to $1/\sigma_{t+k}$, and $\underline{T}$ is the point in the sample where we begin our evaluation. We discuss this further in the latter part of this section, see Figure 2. Moreover, the notation EPA$_w$ and SPA$_w$ explicitly acknowledges that these are rather a class of null hypothesis, depending on the chosen weighting scheme.

Remark 3. In practice when using the SPA null for model selection, we select model $\mathcal{A}$ as long as we do not reject the above null. If we reject the above SPA null we select model $\mathcal{B}$.

Remark 4. The new test based on the above null hypotheses is also applicable for nested models, i.e. the new test can handle cases when $\mu_{t+k} = \mathbb{E}\left[\Delta \mathcal{L}_{t+k}\right] = \mu = 0$ for all $t = 1, \cdots, T$, so long as there is non-zero variance everywhere. For example, when the practitioner wants to compare two nested models, say for example AR(1) and AR(2), although they might provide the same estimated mean at each point in time with $\mu_{t+k} = \mathbb{E}\left[\Delta \mathcal{L}_{t+k}\right] = 0$ for all $t = 1, \cdots, T$, it is likely the case that $\text{var}\left(\Delta \mathcal{L}_{t+k}\right) \neq 0$ for all $t = 1, \cdots, T$, and therefore our test shall apply. We shall stress nevertheless that there are situations when our test will not be applicable, e.g. in the situation

when variance is zero everywhere.

We next provide an intuition for how our test is related to that of Giacomini and White (2006). Recall that Giacomini and White (2006) test 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[6]: $\mathbb{H}_0 : \mathbb{E}\left[\Delta\mathcal{L}_{t+1}h_t\right] = 0$, 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:

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

It is therefore a joint test of the loss difference having zero mean and absence of serial correlation at first lag. Existence of serial correlation at the first lag would be indicative of changing relative performance, as it is no longer the case that $\mu_t = \mu$ for all t. In general, if we reject their null of $\mathbb{E}\left[\Delta\mathcal{L}_{t+1}|\mathcal{F}_t\right] = 0$ due to a dependence of the above moment condition on $\mathcal{F}_t$. This can be considered as evidence of changing relative performance due to a changing information set, $\mathcal{F}_t$. It is still possible that a rejection occurs in a world of constant relative performance, take for example a case where $\mu_t = \mu > 0$ for all t. We possibly may identify changing relative performance as when the Diebold-Mariano (1995) test does not reject, which is indicative of insufficient evidence against zero unconditional mean, however Giacomini and White (2006) test does reject, indicating the rejection is likely due to changing relative performance.

Given our modelling framework, which we shall discuss in the next section, for a particular choice of $\mathbb{X}_t$ one can test a more general null hypothesis that all time-varying coefficients in the regression of $\Delta\mathcal{L}_t$ on $\mathbb{X}_t$ are zero for all $t \in T$. This methodology can be viewed as a way of implementing the conventional Mincer-Zarnowitz (1969) regressions in unstable environments, and it shall nest the Giacomini and White (2006) test as a special case. These tests can also be interpreted as a generic version of the existing forecast rationality tests for a particular choice of $\mathbb{X}_t$. This issue is studied in Smetanina (2018b).

In what follows, we describe our method for constructing forecast losses, which we do using a standard recursive scheme. Note that contrary to the existing out-of-sample tests, for our metric we need to construct losses for the entire sample, and not just a short evaluation window towards the end of sample.

[6]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 $\mathbb{H}_0 : \mathbb{E}\left[\Delta\mathcal{L}_{t+1}|\mathcal{F}_t\right] = 0$.

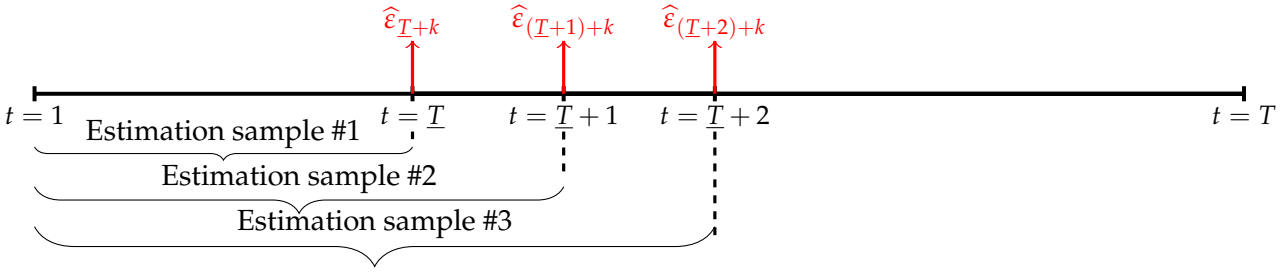


Figure 2: Construction of the time series of the forecast errors for a single model.

The pseudo-out-of-sample forecast made at time t for period $t+k$ is compared with the realised 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 initial period of length $\underline{T}$, where the initial $\underline{T}$ periods is always reserved for estimation. We recommend that practitioner uses $\underline{T} = 100$.

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

$$\Delta\mathcal{L}_{t+k} = \mathcal{L}\left(\widehat{\varepsilon}^{\mathcal{A}}_{t+k}\right) - \mathcal{L}\left(\widehat{\varepsilon}^{\mathcal{B}}_{t+k}\right), \tag{3}$$

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

$$\Delta\mathcal{L}_{t+k} = \left(\widehat{\varepsilon}^{\mathcal{A}}_{t+k}\right)^2 - \left(\widehat{\varepsilon}^{\mathcal{B}}_{t+k}\right)^2. \tag{4}$$

Following the construction of the loss differences, we may next proceed to the theoretical results, which we discuss in the next section.

3 Theoretical results

Once we have constructed the time series of $\Delta\mathcal{L}_t$, we shall from now on only 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 model $\Delta\mathcal{L}_t$ as a locally stationary process and allow its mean and variance to change smoothly over time. In particular, we model $\Delta\mathcal{L}_t$ as a function of time and its own lags only. The

rationale behind such a modelling framework is as follows. It is a well-established fact that due to estimation error, $\Delta\mathcal{L}_t$ exhibits serial correlation, see e.g. Bollerslev et. al. (2016). This motivates the autoregressive structure of our model for $\Delta\mathcal{L}_t$. However, in general we do not expect the difference in losses to depend on any other regressors. Therefore, in order to be as agnostic as possible, we do not impose any additional structure on $\Delta\mathcal{L}_t$. 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 as we will be using the infill asymptotics. So, suppose that we observe the time series of forecast loss differences $\{\Delta\mathcal{L}_{t,T}\}$, $t = 1, 2, \cdots T$. The process $\{\Delta\mathcal{L}_{t,T}\}_{t=1}^{T}$ is assumed to follow an autoregressive model with time-varying coefficients which is given by:

$$\Delta\mathcal{L}_{t,T} = \rho_{t,T}^{0} + \sum_{j=1}^{d} \rho_{t,T}^{j} \Delta\mathcal{L}_{t-j,T} + \xi_{t,T}, \qquad t = 1, \cdots, T, \tag{5}$$

where $\mathbb{E}\left[\xi_{t,T}|\mathbb{X}_{t,T}\right] = 0$ with $\mathbb{X}_{t,T} = \left(1, \Delta\mathcal{L}_{t-1,T}, \Delta\mathcal{L}_{t-2,T}, \cdots, \Delta\mathcal{L}_{t-d,T}\right)^{T}$ and $\rho_{t,T}^{j}$, $j = 1, \cdots, d$ are deterministic functions of time.[7] We use the following rescaling method. Let for each $j \in J$, $\rho^{j}(\cdot)$ be a function on $[0, 1]$ and let

$$\rho_{t,T}^{j} = \rho^{j}\left(t/T\right), \qquad t = 1, \cdots, T.$$

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. Note that model (5) is general in the sense that we do not restrict the regressors to be strictly stationary. Instead, we allow the triangular array $\mathbb{X}_{t,T}$ to be locally stationary in the following sense.

DEFINITION 3. *(Vogt, 2012). The process $\{X_{t,T}\}$ is locally stationary if for each rescaled time point $u \in [0, 1]$ there exists an associated process $\{X_t(u)\}$ with the following two properties:*

i) $\{X_t(u)\}$ is strictly stationary;

[7] In the paper we develop the theory for the general time-varying AR(d) model. However, in all applications we will restrict it to be a simple AR(1) model. We believe it is general, yet simple enough to account for serial correlation of the estimation error in the loss differences $\Delta\mathcal{L}_{t,T}$. We therefore suggest that the reader, unless for having a good reason for an alternative specification, always uses the AR(1) model.

ii) it holds that

$$\left\|X_{t,T} - X_t(u)\right\| \leq \left(\left|\frac{t}{T} - u\right| + \frac{1}{T}\right) U_{t,T}(u) \quad a.s.,$$

where $\{U_{t,T}(u)\}$ *is a process of positive variables satisfying* $\mathbb{E}\left[(U_{t,T}(u))^{\rho}\right] < C$ *for some* $\rho > 0$ *and* $C < \infty$ *independent of* u, t *and* T. $\|\cdot\|$ *denotes an arbitrary norm on* $\mathbb{R}^d$.

In addition, the error process $\{\xi_{t,T} : t = 1, \cdots, T\}$ is assumed to have the martingale difference property, i.e. for all $t = 1, \cdots, T$

$$\mathbb{E}\left[\xi_{t,T} | \{\mathbb{X}_{s,T} : s \leq t\}, \{\xi_{s,T} : s < t\}\right] = 0 \tag{6}$$

Although the above condition rules out autocorrelation in the error terms, it allows for heteroskedasticity. For example (6) is satisfied by residuals of the form:

$$\xi_{t,T} = \sigma_{t,T}\varepsilon_t = \sigma\left(\frac{t}{T}\right)\varepsilon_t,$$

where $\sigma(\cdot)$ is a time-varying volatility function and $\{\varepsilon_t\}$ is a martingale difference process with the variance normalized to 1. We impose the martingale difference structure on the regression error terms as, i) this allows to relax the stronger condition of the i.i.d. errors, yet technically convenient as one can use CLT for martingale differences in the proofs, ii) the correlation of $\Delta\mathcal{L}_{t,T}$ is already accounted for by imposing an autoregressive structure. Therefore, we can re-write our model (5) as follows:

$$\Delta\mathcal{L}_{t,T} = \mathbb{X}_{t,T}^{T}\rho\left(t/T\right) + \sigma\left(t/T\right)\varepsilon_t, \tag{7}$$

where $\rho(t/T) = (\rho_0(t/T), \rho_1(t/T), \cdots, \rho_d(t/T))^T$ and $\mathbb{X}_{t,T} = (1, \Delta\mathcal{L}_{t-1,T}, \Delta\mathcal{L}_{t-2,T}, \cdots, \Delta\mathcal{L}_{t-d,T})^T$. Finally we assume that $\rho(u) = \rho(0)$ and $\sigma(u) = (0)$ 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 the model (7) 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:

$$\widehat{\theta}(u) = \arg\min_{\theta} \sum_{t=1}^{T} K_{h_1}\left(t/T - u\right)\left(\Delta\mathcal{L}_{t,T} - \mathbb{Z}_{t,T}^{T}\theta\right)^2,$$

where $\mathbb{Z}_{t,T} = (\mathbb{X}_{t,T}, \mathbb{X}_{t,T}(t/T - u))^T$ and $\theta = \theta(t/T) = (\rho(t/T), \dot{\rho}(t/T))^T$.

Step 2 : Define the estimated error term $\widehat{\xi}_{t,T} = \Delta\mathcal{L}_{t,T} - \mathbb{Z}_{t,T}^T\widehat{\theta}(t/T)$. Then estimate the conditional variance $\sigma^2(t/T)$ by running the local constant nonparametric regression of $\widehat{\xi}_{t,T}^2$ on the rescaled time t/T, i.e.

$$\widehat{\sigma}^2(u) = \arg\min_a \sum_{t=1}^{T} K_{h_2}(t/T - u)\left(\widehat{\xi}_{t,T}^2 - a\right)^2.$$

Remark 5. In the second-step estimation we use the local constant estimator, primarily because contrary to local constant estimator $\widehat{\sigma}(u)$ (with non-negative kernel), the local linear estimator $\widehat{\sigma}(u)$ is not guaranteed to be positive.

Remark 6. Although $\widehat{\sigma}(u)$ is a second-step estimator, Fan and Yao (1998) analysed the asymptotic distribution of such an estimator and showed that its asymptotic distribution is identical to that obtained via one-step estimation based on the true errors ξ_t. This is an important result as this allows one to select the optimal bandwidths h_1, h_2 independently based on the conventional one-step procedures. We discuss the bandwidths selection procedure in section 4.

ASSUMPTION A1

(i) The function ρ is uniformly bounded below one, i.e. $\sup_{u\in[0,1]} \|\rho(u)\| \leq \overline{\rho} < 1$.

(ii) The function $\sigma(\cdot)$ is bounded from above and from below, i.e. there exist constants $C_\sigma < \infty$ and $c_\sigma > 0$ such that $0 < c_\sigma \leq \sigma(u) \leq C_\sigma < \infty$ for all $u \in [0,1]$.

(iii) The functions ρ and σ are Lipschitz continuous with respect to the rescaled time u.

(iv) The residuals $\{\varepsilon_t\}$ is a martingale difference sequence with respect to the information set $\mathcal{F}_{t-1} = \sigma(\Delta\mathcal{L}_s, \varepsilon_s | s \leq t-1)$. Moreover, ε_t satisfies $\mathbb{E}\left[|\varepsilon_t|^{4+\delta}\right] < \infty$ for some small $\delta > 0$ and are normalised such that $\mathbb{E}\left[\varepsilon_t^2 | \mathcal{F}_{t-1}\right] = 1$.

(v) The error term ε_t has an everywhere positive and continuous density f_ε. The density f_ε is bounded and Lipschitz.

Assumption $A1$ lays out the sufficient conditions for establishing that the process $\{\Delta\mathcal{L}_{t,T}\}$ can be locally approximated by $\Delta\mathcal{L}_t(u)$. Moreover one can also show that for each u the process $\{\Delta\mathcal{L}_t(u), t \in \mathbb{Z}\}$, where $\Delta\mathcal{L}_t(u) = \mathbb{X}_t^T(u)\rho(u) + \sigma(u)\varepsilon_t$ has a strictly stationary solution. Assumption $A1$ corresponds to the assumptions $(M1) - (M3)$, $(\Sigma_1) - (\Sigma_3)$ and $(E1)$ in Vogt (2012)

under which he establishes these results for a more general class of models, see Theorems 3.1 and 3.2 in Vogt (2012). However, to establish that the process $\{\Delta\mathcal{L}_{t,T}\}$ is geometrically $\beta-$mixing one extra assumption on the density of the error term ε_t is required.

ASSUMPTION A2

(i) The density f_ε fulfills the requirement:

$$\int_{\mathbb{R}} |f_\varepsilon(x) - f_\varepsilon(x+\alpha)|dx \leq C|\alpha|,$$

where C is a constant such that $C < \infty$.

Assumption $A2$ corresponds to the assumption $(E3)$ in Vogt (2012), which together with Assumption $A1$ allows one to prove that the process $\Delta\mathcal{L}_{t,T}$ is geometrically β-mixing, see Theorem 3.4 in Vogt (2012) for a proof of this result. Similar assumption can be found in e.g. Orbe et. al. (2005), who establishes the mixing property of the time-varying AR(1) process. Finally, below we introduce the rest of the assumptions that will be necessary to present the estimation theory.

ASSUMPTION A3

(i) The functions ρ and σ ate twice continuously differentiable with respect to the rescaled time u and have bounded derivatives.

(ii) The kernel K is a second-order kernel, which is bounded symmetric around zero density function that has a compact support, i.e. $K(v) = 0$ for all $|v| > C_2$ with some $C_2 < \infty$. Moreover K is Lipschitz, i.e. $|K(v) - K(v')| \leq L|v - v'|$ for some $L < \infty$ and all $v, v' \in \mathbb{R}$. In addition, K satisfies $\int K(z)dz = 1$, $\lambda_j = \int z^j K(z)dz$ and $\nu_j = \int z^j K^2(z)dz$.

(iii) The bandwidths h_1 and h_2 satisfy the following conditions: as $T \to \infty$, $h_1 \to 0$, $Th_1 \to \infty$ and $Th_1^5 \to 0$. Similarly it holds that as $T \to \infty$, $h_2 \to 0$, $Th_2 \to \infty$ and $Th_2^5 \to 0$.

Assumption $A3$(i) ensures that the resulting estimators $\widehat{\rho}(\cdot)$ and $\widehat{\sigma}^2(\cdot)$ are well-behaved which will allow me to apply the kernel methods as well as the bootstrap methods later on. Assumptions $A3$(ii)-(iii) are standard assumptions on the kernel function and bandwidths, where we take the kernel K to be the second-order kernel. Note that we work with equally spaced time periods, however this is not strictly necessary. For instance, the theory will hold with slight modifications,

which we do not present here, for t_i, $i = 1, \cdots, n$ such that $\{t_i/T,\ i = 1, \cdots, n\}$ is dense on a unit interval.

Before stating the first main results, we need to introduce some further notation. We define the following two matrices:

$$\Omega_{t,T} = \mathbb{E}\left[\mathbb{X}_{t,T}\mathbb{X}_{t,T}^T\right], \quad \text{and} \quad H = \begin{bmatrix} I_{d+1} & 0 \\ 0 & h_1 I_{d+1} \end{bmatrix},$$

where I_{d+1} is the identity matrix of dimension $(d+1) \times (d+1)$.

THEOREM 1. *Let Assumptions (A1)-(A3) hold. Then for any $u \in (0,1)$ it holds that*

$$\sqrt{Th_1}\left(H\{\widehat{\theta}(u) - \theta(u)\} - h_1^2 \mathbb{B}_1(u))\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_\theta(u)\right),$$

where

$$\mathbb{B}_1(u) = \frac{1}{2}\begin{pmatrix} \lambda_2 \ddot{\rho}(u) \\ 0 \end{pmatrix} \qquad \textit{and} \qquad \mathbb{V}_\theta(u) = \begin{pmatrix} \nu_0 \sigma^2(u)\Omega^{-1}(u) & 0 \\ 0 & \lambda_2^{-2}\nu_2 \sigma^2(u)\Omega^{-1}(u) \end{pmatrix}.$$

In step two we estimate the variance $\sigma^2(u)$ by running local constant nonparametric regression of squared residuals

$$\widehat{\xi}_{t,T}^2 = \left[\Delta\mathcal{L}_{t,T} - \mathbb{Z}_{t,T}^T \widehat{\theta}(t/T)\right]^2, \qquad t = 1, \cdots, T$$

on the rescaled time t/T. Given that our test statistics based on eq. (1) or eq. (2) aggregates $\widehat{\mu}_t(u)$ over $u \in [0,1]$ weighted by its standard deviation, it becomes necessary to establish the uniform convergence of $\widehat{\sigma}^2(u)$ over the whole support of u rather than just establishing pointwise consistency of $\sigma^2(u)$. The next theorem states the uniform convergence rate for the second-step estimator $\widehat{\sigma}^2(u)$.

THEOREM 2. *Let Assumptions (A1)-(A3) hold. Denote by $I_{h_2} = [C_1 h_2, 1 - C_1 h_2]$, where $C_1 > 0$ such that $C_1 h_2 \to 0$ and $1/C_1 \to 0$. Then*

$$\sup_{u \in I_{h_2}} \left|\widehat{\sigma}^2(u) - \sigma^2(u)\right| = O_p\left(\sqrt{\frac{\log T}{Th_2}} + h_2^2\right),$$

and with probability tending to one it also holds that

$$\sup_{u \in I_{h_2}} |\widehat{\sigma}^2(u)| \leq C_\sigma \leq \infty,$$

where $C_\sigma > 0$.

3.1 Test Statistics

Provided with our new definition of ranking, it is then straightforward to state the null and alternative hypotheses. In particular, consider the SPA_w:

$$\mathbb{H}_0 : \quad \sum_{t=1}^{T} w_t \mu_t \leq 0 \qquad \text{vs} \qquad \mathbb{H}_1 : \quad \sum_{t=1}^{T} w_t \mu_t > 0. \tag{8}$$

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 $w_t \propto \phi_t / se_t$, see section 2 for the detailed motivation of such a weighting. We form the test statistic corresponding to the above null by replacing the unknown quantities with the respective estimators. We first define the local t-statistic, which we denote by $\widehat{\tau}(u)$:

$$\widehat{\tau}(u) = \frac{\widehat{\mu}(u)}{\widehat{se}(u)} = \frac{\sqrt{Th_1}\mathbb{X}_t^T(u)\widehat{\rho}(u)}{\widehat{\sigma}(u)\sqrt{\nu_0 \mathbb{X}_t^T(u)\widehat{\Omega}^{-1}(u)\mathbb{X}_t(u)}}, \tag{9}$$

and then the integrated t-statistic is given by:

$$\mathcal{S}_T = \int_0^1 \widehat{\tau}(u)du,$$

or a slightly extended version with an extra (given) weighting $\phi(u)$:

$$\mathcal{S}'_T = \frac{1}{\sqrt{\Phi}} \int_0^1 \phi(u)\widehat{\tau}(u)du, \qquad \text{with} \qquad \Phi = \int_0^1 \phi^2(u)du.$$

An example of $\phi(u)$ is $\phi(u) = \mathbb{1}(u \in I)$, where I could be a period of time that forecaster is interested in, for example, recession times. In what follows we analyse the asymptotic behaviour of $\mathcal{S}_T$ under the null as well as under fixed and local alternatives. The fixed alternative hypothesis

is given by

$$\mathbb{H}_1 : \qquad \int_0^1 \tau(u)du > 0.$$

In addition, to get a rough 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 $\tau_T(u)$ given by:

$$\tau_T(u) = \tau(u) + c_T \triangle (u),$$

where $c_T \to 0$ as $T \to \infty$, the function $\triangle$ is continuous and the quantity $\int_0^1 \tau(u)du$ satisfies the null hypothesis $\mathbb{H}_0$. Under these local alternatives the process $\Delta\mathcal{L}_{t,T}$ is given by

$$\Delta\mathcal{L}_{t,T} = \mathbb{X}_{t,T}^T \rho\left(t/T\right) + c_T \triangle\left(t/T\right)\sigma\left(t/T\right)\sqrt{\nu_0 \mathbb{X}_{t,T}^T \Omega^{-1}(u)\mathbb{X}_{t,T}/Th_1} + \xi_{t,T}, \tag{10}$$

for $t = 1, \cdots, T$, and therefore under (10), we move along the following sequence of local alternatives:

$$\mathbb{H}_{1,T} : \qquad \int_0^1 \tau_T(u)du = c_T \int_0^1 \triangle(u)du.$$

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 $T \to \infty$.

THEOREM 3. *Let Assumptions (A1)-(A3) hold. Then conditional on the sample* $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^T$ *under* $\mathbb{H}_0$,

$$\sqrt{T}\left(S_T - h_1^2 \mathbb{B}_T\right) \xrightarrow{d} \mathcal{N}\left(0,1\right),$$

where

$$\mathbb{B}_T = \frac{1}{2}\int_0^1 \frac{\lambda_2 \mathbb{X}_t^T(u)\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0 \mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}}du. \tag{11}$$

The next theorem states the asymptotic distribution of the modified statistic $\mathcal{S}'_T$.

THEOREM 4. *Let Assumptions (A1)-(A3) hold. Then conditional on the sample* $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^T$ *under* $\mathbb{H}_0$,

$$\sqrt{T}\left(S'_T - h_1^2 \mathbb{B}'_T\right) \xrightarrow{d} \mathcal{N}\left(0,1\right),$$

where

$$\mathbb{B}'_T = \frac{1}{2\sqrt{\Phi}} \int_0^1 \frac{\phi(u)\lambda_2 \mathbb{X}_t^T(u)\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0 \mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}} du \qquad \textit{and} \qquad \Phi = \int_0^1 \phi^2(u)du.$$

We now turn to the theoretical results for the fixed and local alternatives. The next theorem states that the bias-corrected statistic $\mathcal{S}_T$ diverges in probability to infinity under $\mathbb{H}_1$. This allows me to establish consistency of the test against fixed alternatives.[8]

THEOREM 5. *Let Assumptions (A1)-(A3) hold. Then under* $\mathbb{H}_1$,

$$S_T - h_1^2 \mathbb{B}_T \xrightarrow{p} \int_0^1 \triangle(u)du > 0,$$

where $\mathbb{B}_T$ *is given by eq.(11).*

We next examine the behaviour of $\mathcal{S}_T$ under local alternatives. The theorem 6 below states that the asymptotic power of the test against local alternatives of the form $\tau_T(u) = \tau(u) + c_T \triangle(u)$ with $c_T = 1/\sqrt{T}$ and $\int_0^1 \tau(u)du$ satisfying $\mathbb{H}_0$, is constant for all functions $\triangle$ and is determined by $\int_0^1 \triangle(u)du$.

THEOREM 6. *Let Assumptions (A1)-(A3) hold. Let* $c_T = 1/\sqrt{T}$, *then conditional on the sample* $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^T$ *under* $\mathbb{H}_{1,T}$,

$$\sqrt{T}\left(S_T - h_1^2 \mathbb{B}_T\right) \xrightarrow{d} \mathcal{N}\left(\int_0^1 \triangle(u)du, 1\right),$$

where $\mathbb{B}_T$ *is given by eq.(11).*

3.2 Sign Forecasting

We now present the theory for sign forecasting. We could also forecast the level of forecast losses, however for the reasons outlined in section 2 we believe that the sign is more informative for model selection. Given that our model (5) 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 $\mathcal{F}_{t,T} =$

[8] This result also holds for the modified statistic $\mathcal{S}'_T$.

$\sigma\left(\Delta\mathcal{L}_{s,T}, \varepsilon_{s,T} | s \leq t\right)$ to be the sigma-algebra generated by the history of $\Delta\mathcal{L}_{t,T}$ and $\varepsilon_{t,T}$, and recall that the model for $\Delta\mathcal{L}_t$ is given by

$$\Delta\mathcal{L}_t = \mathbb{X}_t^T \rho(t/T) + \sigma(t/T)\varepsilon_t, \quad t = 1, \cdots, T \tag{12}$$

where with some abuse of notation due to the meaning of T, x^T denotes the transpose of x and $\mathbb{X}_t = (1, \Delta\mathcal{L}_{t-1}, \cdots, \Delta\mathcal{L}_{t-d})^T$ and ε_t is a m.d.s. At the final point in the sample T we would like to predict the sign of $\Delta\mathcal{L}_{T+1}$, i.e. we would like to know:

$$\begin{aligned} Pr\left(\Delta\mathcal{L}_{T+1} \leq 0 | \mathcal{F}_T\right) = Pr\left(\mathbb{X}_{T+1}^T \rho\left(\frac{T+1}{T}\right) + \sigma\left(\frac{T+1}{T}\right)\varepsilon_{T+1} \leq 0 | \mathcal{F}_T\right) = \\ = Pr\left(\varepsilon_{T+1} \leq \frac{-\mathbb{X}_{T+1}^T \rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\right) = F(\varepsilon). \end{aligned} \tag{13}$$

Here, given the autoregressive structure of the difference in losses, $\mathbb{X}_{T+1} \in \mathcal{F}_T$. Recall also that $\rho(u) = \rho(1)$ and $\sigma(u) = \sigma(1)$ for $u \geq 1$. It then holds from (13) that

$$Pr\left(\Delta\mathcal{L}_{T+1} \leq 0 | \mathcal{F}_T\right) =: F\left(\varepsilon^\star(1)\right),$$

where $\varepsilon^\star(1) := -\mathbb{X}_{T+1}^T \rho\left(1\right)/\sigma(1)$ is the standardised residual from eq.(12) at the last point $u \approx T/T = 1$. Conditional on the sample $\{\Delta\mathcal{L}_t\}_{t=1}^T$, we can estimate the conditional probability as follows:

$$\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = \widehat{F}\left(\widehat{\varepsilon}^\star(1)\right) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(\widehat{\varepsilon}_t \leq \frac{-\mathbb{X}_{T+1}^T \widehat{\rho}\left(1\right)}{\widehat{\sigma}\left(1\right)}\right),$$

where $\widehat{\varepsilon}^\star(1)$ is an estimate of $\varepsilon^\star(1)$. 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 7 below, which allows the practitioner to calculate the probability as well as the confidence intervals for this probability, which we call **forecast intervals**.

THEOREM 7. *Let Assumptions (A1)-(A3) hold. Let $\mathcal{F}_{t,T} = \sigma\left(\Delta\mathcal{L}_{s,T}, \varepsilon_{s,T} | s \leq t\right)$ to be the sigma-algebra generated by the history of $\Delta\mathcal{L}_{t,T}$ and $\varepsilon_{t,T}$. Then the forecast of the sign of $\Delta\mathcal{L}_{T+1}$ made at time T is given by*

$$\sqrt{T}\left[\widehat{F}(\widehat{\varepsilon}^\star(1)) - F(\varepsilon^\star(1)) - \mathbb{B}_3(1)\right] \xrightarrow{d} \mathcal{N}\left(0, F\left(\varepsilon^\star(1)\right)\left(1 - F\left(\varepsilon^\star(1)\right)\right)\right), \tag{14}$$

where

$$\mathbb{B}_3(1) = \frac{f\left(\varepsilon^\star(1)\right)}{2\sigma^2(1)} \mathbb{X}_{T+1}(1)\left\{h_1^2 \lambda_2 \ddot{\rho}(1)\sigma(1) + h_2^2 \lambda_2 \ddot{\sigma}(1)\right\},$$

and

$$F(\varepsilon^{\star}(1)) = Pr(\Delta\mathcal{L}_{T+1} \leq 0) \quad \textit{and} \quad \varepsilon^{\star}(1) := \frac{-\mathbb{X}_{T+1}^{T}\rho\,(1)}{\widehat{\sigma}(1)}.$$

Remark 7. The bias term $\mathbb{B}_3(1)$ is due to the estimation error of $\widehat{\varepsilon}(1)$, which itself involves the estimates of $\widehat{\rho}(1)$ and $\widehat{\sigma}(1)$, leading to a particular form of the bias in Theorem 7.

In the simulations (see Figure 8) we see that the sign forecasts perform quite well, forecasting near to 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 ε^* threshold is small, the probability will not respond much to inaccuracies in ρ. One way to improve forecasts even further would be 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, in one of our simulations and in all of our applications we will be producing 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 our applications where the true probabilities are unknown, we use the following criterion to assess the quality of our forecasts:

$$\widehat{C} := \frac{1}{T-\underline{T}} \sum_{t=\underline{T}}^{T} \left[\mathbb{1}\left(\Delta\mathcal{L}_{t+1} \leq 0\right) - \widehat{Pr}^{bc}\left(\Delta\mathcal{L}_{t+1} \leq 0 | \Delta\mathcal{L}_{t}\right) \right], \tag{15}$$

where $\underline{T} = 100$ is the first splitting point where we begin our evaluation, and $\widehat{Pr}^{bc}(\cdot)$ 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 the optimal bandwidths h_1 and h_2. We start with the optimal 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:

$$\text{AMSE}(h_1) = \frac{h_1^4}{4}\mu_2^2\|\ddot{\theta}(u)\|_2^2 + \frac{\text{tr}\left(\mathbb{V}_\theta(u)\right)}{Th_1},$$

where $V_\theta(u)$ is given in Theorem 1. Then minimising $\text{AMSE}(h_1)$ with respect to h_1 provides the optimal bandwidth h_1^{opt} given by:

$$h_1^{opt} = \left\{\text{tr}\left(\mathbb{V}_\theta(u)\right)\mu_2^{-2}\|\ddot{\theta}(u)\|_2^{-2}\right\}^{-1/5}T^{-1/5} \tag{16}$$

However, note that (16) involves the unknown quantity $\ddot{\theta}(u)$ 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, Fan and Yao (2000), Cai, Fan and Li (2000) 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 $\widehat{\theta}^{-j}$ using the remaining data after deleting the jth group. Now denote by Y_{-d_j} the fitted values of Y_t using the data with the jth group deleted. Then the cross-validation criterion has the following form:

$$\text{CV}(h_1) = \sum_{j=0}^{Q-1}\sum_{i\in d_j}\left[Y_i - \widehat{Y}_{-d_j}\right]^2.$$

Alternatively, one can form variants of the cross-validation criteria based on the Pearson's residuals:

$$\text{CV1}(h_1) = \sum_{j=0}^{Q-1}\sum_{i\in d_j}\left[Y_i \log\left\{\frac{Y_i}{\widehat{Y}_{-d_j}}\right\} - \{Y_i - \widehat{Y}_{-d_j}\}\right],$$

where in the above one would need to set $0 \log 0 = 0$ to account for the cases when $Y_i = 0$. Finally, another cross-validation criterion can be

$$\text{CV2}(h_1) = \sum_{j=0}^{Q-1} \sum_{i \in d_j} \left[\frac{Y_i - \widehat{Y}_{-d_j}}{\sqrt{\widehat{Y}_{-d_j}}} \right].$$

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

5 Bootstrapping $\mathcal{S}_T$

Theorems 3-6 allow one to conduct inference for $\mathcal{S}_T$, as the distribution of the test statistics is the simple standard normal distribution. Note also, that the test statistic $\mathcal{S}_T$ is a nonparametric statistic, however through aggregation it converges to $\mathcal{N}(0,1)$ with the standard parametric $\sqrt{T}$ rate. The bias term in Theorems 3-6, however, contains unknown quantities, such as $\ddot{\rho}(u)$. Although it is possible to estimate these unknown quantities, replacing them with the consistent estimates will further result in approximation errors. We therefore choose to bootstrap the statistics, which will automatically allow me to estimate the bias without estimating the unknown quantities. In what follows, we discuss the bootstrap procedure in the context of Theorems 3-6, however the same methodology will be applied to obtain the bias and the forecast intervals in Theorem 7. We set up the fixed regressor wild bootstrap procedure to account for the time series structure of the data. In particular, the wild bootstrap sample, which we denote by $\{\Delta\mathcal{L}^{\star}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$, where

$$\Delta\mathcal{L}^{\star}_{t,T} = \mathbb{X}^{T}_{t,T}\widetilde{\rho}\,(t/T) + \xi^{\star}_{t,T}, \tag{17}$$

and the bootstrap residuals are constructed as follows:

$$\xi^{\star}_{t,T} = \widehat{\xi}_{t,T}\eta_t,$$

where $\widehat{\xi}_{t,T} := \Delta\mathcal{L}_{t,T} - \widehat{\mu}_t(t/T) = \Delta\mathcal{L}_{t,T} - \mathbb{X}^{T}_{t,T}\widehat{\rho}(t/T)$ are the estimated residuals and $\{\eta_t\}_{t=1}^{T}$ is a sequence of i.i.d. variables normalized such that it has zero mean and unit variance. We further

choose η_t to have a Rademacher distribution. Finally $\widetilde{\rho}(\cdot)$ is given by:

$$\widetilde{\rho}(u) := \widehat{\rho}(u) - \bar{\rho}, \qquad \bar{\rho} = \int_0^1 \widehat{\rho}(u)du, \tag{18}$$

The intuition behind construction of the mean function of $\Delta\mathcal{L}^{\star}_{t,T}$ given by (17)-(18) is such that the bootstrapped sample $\{\Delta\mathcal{L}^{\star}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$ imitates the model under the null hypothesis whether the alternative hypothesis is true or not. Therefore the distribution of the bootstrapped statistic $S^{\star}_T$, stated below, mimics the distribution of S_T under the null hypothesis regardless whether the null holds or not. Given the bootstrap sample $\{\Delta\mathcal{L}^{\star}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$, we define the bootstrapped test statistics $\mathcal{S}^{\star}_T$

$$\mathcal{S}^{\star}_T = \int_0^1 \frac{\widehat{\mu}^{\star}_t(u)}{\widehat{se}^{\star}_t(u)}du, \qquad \widehat{se}^{\star}_t(u) = \widehat{\sigma}^{\star}(u)\sqrt{\nu_0 \mathbb{X}^T_t(u)\widehat{\Omega}^{-1}(u)\mathbb{X}_t(u)/Th_1},$$

where

$$\widehat{\mu}^{\star}_t(u) = \widehat{\rho}^{\star}(u)\mathbb{X}^T_t(u), \quad \text{where} \quad \widehat{\rho}^{\star}(u) := \arg\min_a \sum_{t=1}^{T} K_{h_1}(t/T - u)\left(\Delta\mathcal{L}^{\star}_{t,T} - \mathbb{X}^T_{t,T}a\right)^2,$$

and defining $\widehat{\xi}^{\star}_{t,T} = \Delta\mathcal{L}^{\star}_{t,T} - \widehat{\mu}^{\star}_t$, we further have that $\widehat{\sigma}^{\star}(u)$ is given by

$$\widehat{\sigma}^{\star}(u) = \frac{\sum_{t=1}^{T} K_{h_2}(t/T - u)\left(\widehat{\xi}^{\star}_{t,T}\right)^2}{\sum_{t=1}^{T} K_{h_2}(t/T - u)}.$$

The next theorem states that the wild bootstrap described above is consistent.

THEOREM 8. *Let Assumptions (A1)-(A3) hold. Then conditional on the sample* $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$ *with probability tending to one*

$$\sqrt{T}\left(\mathcal{S}^{\star}_T - \mathcal{S}^{\star} - h_1^2\mathbb{B}_T\right) \xrightarrow{d} \mathcal{N}(0,\, 1),$$

where $\mathbb{B}_T$ *is given by eq.(11). In other words,* $\mathbb{P}^{\star}\left(\mathcal{S}^{\star}_T - \mathcal{S}^{\star} - h_1^2\mathbb{B}_T \leq x\right) \xrightarrow{p} \Phi(x)$*, where* $\Phi(x)$ *is a Gaussian distribution function with zero mean and variance 1.*

Once the wild bootstrap is set up, the size and the power of the test in the next section will then be 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 $\mathcal{S}^{\star}_{T,n,b}$ be the value of the bootstrap statistics $\mathcal{S}^{\star}_T$ 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 $\{\mathcal{S}_{T,n,b}^\star\}_{b=1}^B$. Then the actual size of the test statistics can be calculated as follows. Given a fixed nominal size α, for each simulated sample $n \in N$, 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

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

6 Simulations

In this section we provide the simulations' results for the size and power of our test statistic $\mathcal{S}_T$ as well as demonstrating the sign forecasting methodology. We start by investigating the size of our test $\mathcal{S}_T$.

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:

$$\mathbb{H}_0^{(1)}: \qquad \Delta\mathcal{L}_t = \mu(1-\rho) + \rho\Delta\mathcal{L}_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d.\mathcal{N}(0,1) \tag{19}$$

For each of $n \in N$ simulations we generate a sequence of loss differences $\Delta\mathcal{L}_t$ of length $T = 150$ according to (19), 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 $\rho(0, 0.05, \cdots, 0.9)$.[9] Tables 1-2 show the simulated actual size for different levels of the nominal size $\alpha = 0.01,\ 0.05,\ 0.10,\ 0.15$.

[9] We present the results for $\rho = 0.2$ only as varying ρ virtually leaves the results unchanged.

Table 1: Actual size versus nominal size of two-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(1)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.012	0.072	0.117	0.162
$B = 750$	0.009	0.062	0.107	0.156
$B = 1000$	0.009	0.057	0.104	0.151

Table 2: Actual size versus nominal size of one-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(1)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.012	0.056	0.101	0.160
$B = 750$	0.011	0.050	0.099	0.162
$B = 1000$	0.011	0.051	0.097	0.155

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

$$\mathbb{H}_0^{(2)}: \qquad \Delta\mathcal{L}_t = \frac{\mu}{p(1-p)}(S_t - p) + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.\mathcal{N}(0,1), \tag{20}$$

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 $\mathbb{E}[\Delta\mathcal{L}_t] = 0$, however

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

Tables 3-4 show the simulated actual size for different levels of the nominal size $\alpha = 0.01,\ 0.05,\ 0.10,\ 0.15$.

Table 3: Actual size versus nominal size of two-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(2)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.020	0.048	0.108	0.160
$B = 750$	0.018	0.052	0.107	0.154
$B = 1000$	0.015	0.050	0.103	0.150

Table 4: Actual size versus nominal size of one-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(2)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.012	0.058	0.104	0.150
$B = 750$	0.011	0.050	0.099	0.148
$B = 1000$	0.010	0.050	0.099	0.148

We next 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:

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

and

$$\rho^0(u) = \sin(8\pi u), \quad \rho^1(u) = 0 \quad \forall\, u \quad \text{and} \quad \sigma(u) = 1 \quad \forall\, u.$$

Under $\mathbb{H}_0^{(3)}$ the mean of $\Delta\mathcal{L}_t$ is time-varying. The mean performs four full sine cycles over the course of our sample, so that over the whole sample the overall mean is also zero by symmetry. For simplicity we set the variance to be constant throughout.

Table 5: Actual size versus nominal size of two-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(3)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.015	0.065	0.107	0.150
$B = 750$	0.014	0.061	0.105	0.145
$B = 1000$	0.012	0.060	0.105	0.145

Table 6: Actual size versus nominal size of one-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(3)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.010	0.043	0.087	0.141
$B = 750$	0.009	0.043	0.088	0.143
$B = 1000$	0.010	0.045	0.090	0.143

The above tables show that the actual size is very close to the nominal size for all levels and for all nulls under consideration. The results are stable regardless of the number of bootstrap

replications B.

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 simulates the loss differences $\Delta\mathcal{L}_t$ according to (19) such that $\rho = 0$ and $\mu = (0, 0.05, \cdots, 1)$. We fix the nominal size of the test to be 5%. Below we show the power curves when applying our one-sided and two-sided tests as well as Giacomini and White (2006) test.

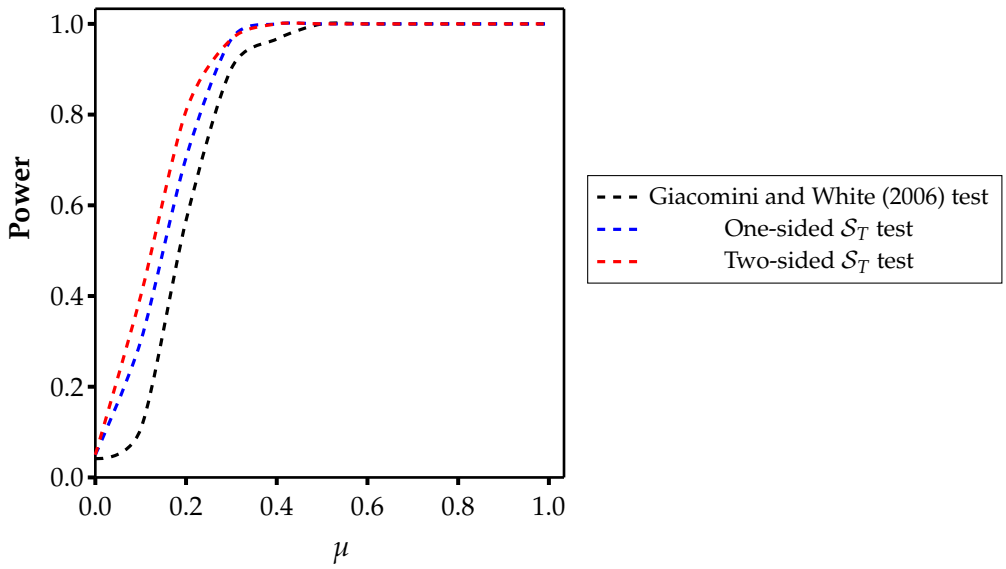


Figure 3: Power curves under alternative $\mathbb{H}_1^{(1)}$.

We next consider another alternative that we borrow from Giacomini and White (2006) paper. In particular, we again generate the loss differences $\Delta\mathcal{L}_t$ according to (20), 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 choice weighting functions to be the states of the world, i.e. we set $\phi_t = S_t$, conditional on the states of the world S_t. In this case (20) constitutes an alternative for our null as well. We plot the power curves below.

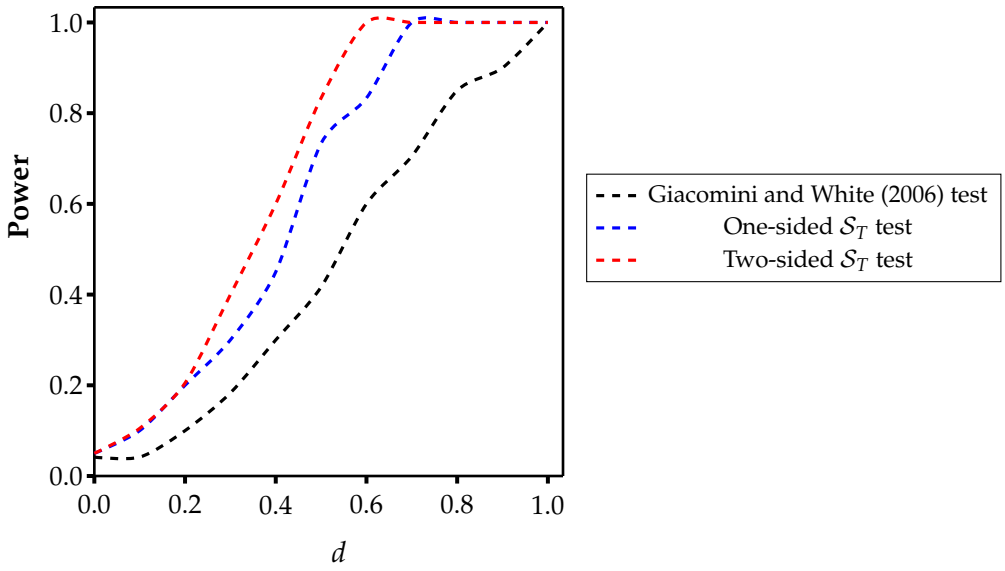


Figure 4: Power curves under alternative $\mathbb{H}_1^{(2)}$.

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:

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

where

$$\rho^0(u) = \sin(8\pi u) + 0.1, \quad \rho^1(u) = 0 \ \ \forall\, u \quad \text{and} \quad \sigma(u) = 1 \ \ \forall\, u.$$

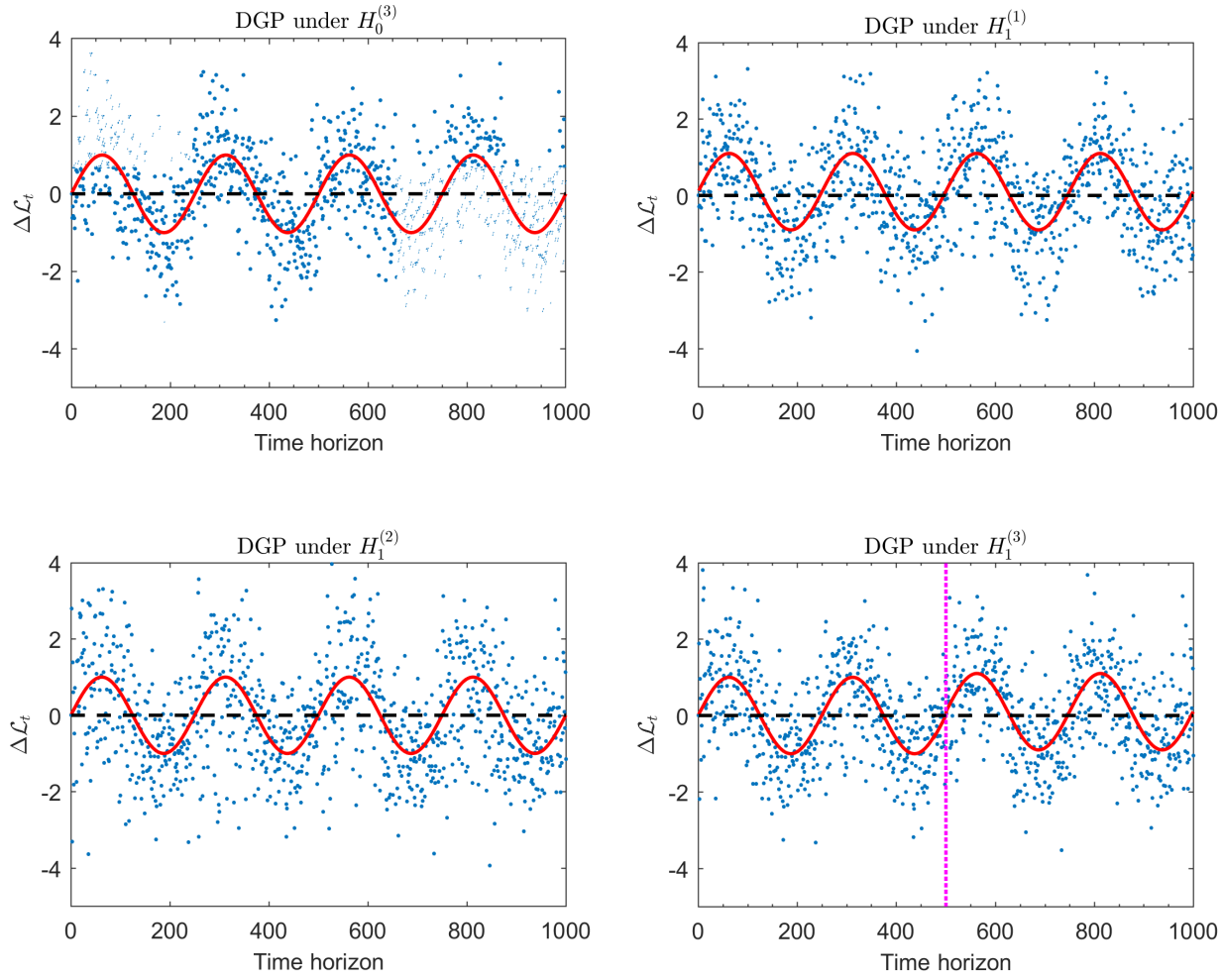


Figure 6: Data generating processes (DGP) under the null $\mathbb{H}_0^{(3)}$ as the corresponding alternatives $\mathbb{H}_1^{(2)}$, $\mathbb{H}_1^{(3)}$ and $\mathbb{H}_1^{(3)}$. The red lines represents the true mean function μ_t.

Under $\mathbb{H}_1^{(1)}$, we add a small intercept to the curve of the mean from the null. 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 volatile and downswings are more volatile, more precisely:

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

and where

$$\rho^0(u) = \sin(8\pi u), \qquad \rho^1(u) = 0 \quad \forall\, u,$$

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

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

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 $\mathcal{B}$ versus the model $\mathcal{A}$.

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

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

where $\rho^1(u) = 0 \quad \forall\, u \quad$ and $\quad \sigma(u) = 1 \quad \forall\, u$, and

$$\rho^0(u) = \begin{cases} \sin(8\pi u) & \text{for } u \in [1, T/2], \\ \sin(8\pi u) + 0.1 & \text{for } u \in [T/2+1, T]. \end{cases}$$

This alternative highlights the ability for our statistic to 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 7: Mean of $\mathcal{S}_T$.

Alternative	$\mathbb{E}(\mathcal{S}_T)$
$\mathbb{H}_1^{(1)}$	3.08
$\mathbb{H}_1^{(2)}$	2.82
$\mathbb{H}_1^{(3)}$	1.40

Table 8: Power for different alternatives with two-sided null.

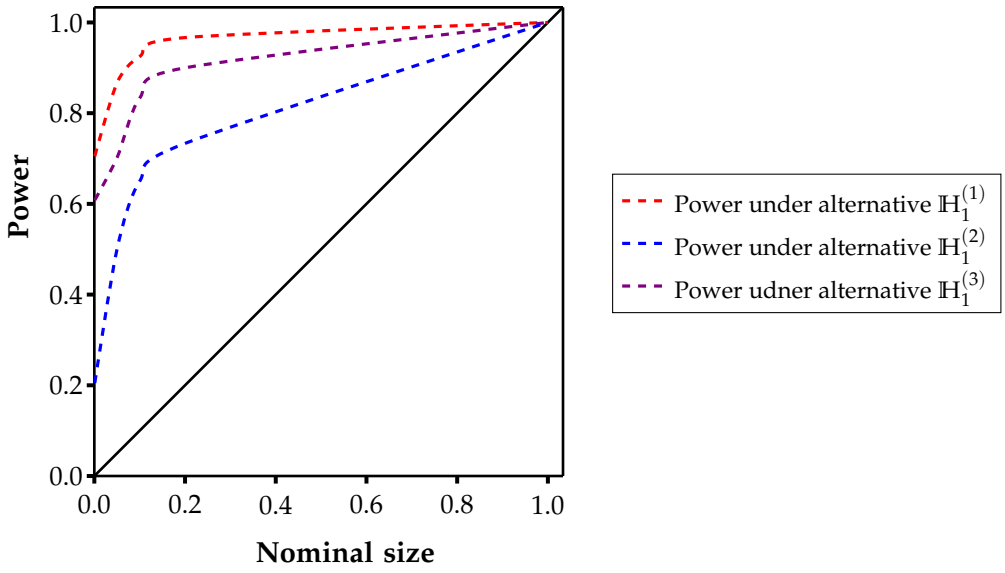
Nominal size	$\mathbb{H}_1^{(1)}$	$\mathbb{H}_1^{(2)}$	$\mathbb{H}_1^{(3)}$
$\alpha = 0.01$	0.75	0.60	0.24
$\alpha = 0.05$	0.88	0.64	0.38
$\alpha = 0.10$	0.94	0.75	0.50
$\alpha = 0.15$	0.97	0.85	0.57

Table 9: Mean of $\mathcal{S}_T$.

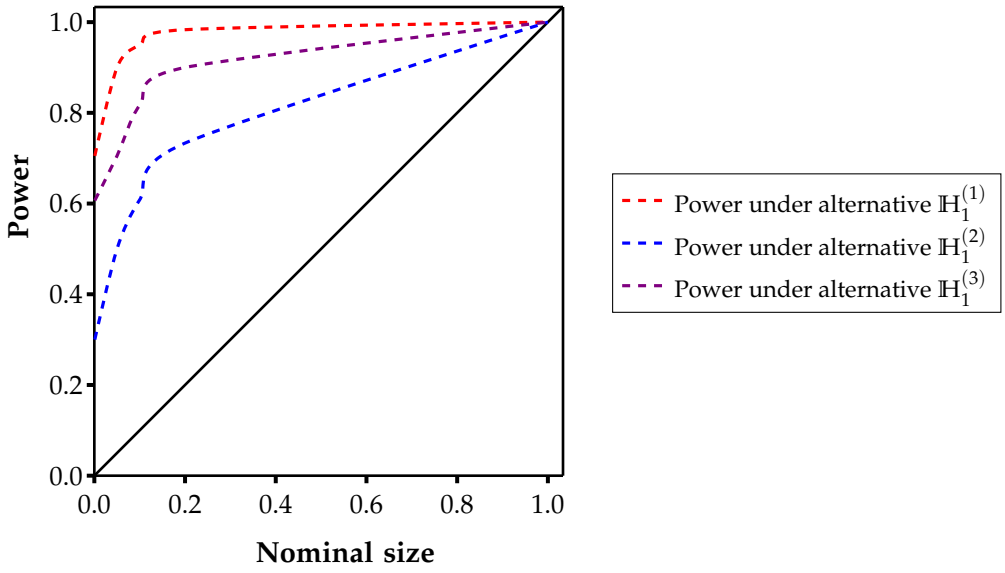
Alternative	$\mathbb{E}(\mathcal{S}_T)$
$\mathbb{H}_1^{(1)}$	3.08
$\mathbb{H}_1^{(2)}$	2.82
$\mathbb{H}_1^{(3)}$	1.40

Table 10: Power for different alternatives with one-sided null.

Nominal size	$\mathbb{H}_1^{(1)}$	$\mathbb{H}_1^{(2)}$	$\mathbb{H}_1^{(3)}$
$\alpha = 0.01$	0.76	0.66	0.28
$\alpha = 0.05$	0.90	0.74	0.48
$\alpha = 0.10$	0.96	0.83	0.62
$\alpha = 0.15$	0.98	0.90	0.72



(a) Two-sided test $\mathcal{S}_T$.



(b) One-sided test $\mathcal{S}_T$.

Figure 7: The figure plots 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 7 shows that our test has a 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 $\Delta\mathcal{L}_t$ 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 11 displays the results of applying these tests as function of the cutoff point ρ, which is a fraction of the sample length used for forecast evaluation to the sample length used for the model estimation.

Table 11: Results of applying standard tests under different alternatives.

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

p-value/Cutoff ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
DM (1995)	0.419	0.170	0.624	**0.002**	**0.042**	**0.033**	0.310	**0.040**	0.207	**0.026**
GW (2006)	**0.011**	**0.010**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

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

p-value/Cutoff ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
DM (1995)	0.524	0.205	0.098	**0.031**	**0.012**	0.219	0.146	0.924	0.609	0.057
GW (2006)	0.586	0.206	0.100	**0.010**	**0.024**	**0.035**	**0.010**	**0**	**0**	**0**

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

p-value/Cutoff ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
DM (1995)	0.484	0.194	0.474	**0.014**	0.408	0.062	0.065	0.136	0.9496	**0.017**
GW (2006)	0.090	0.100	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Note: The cutoff point ρ is defined as a fraction of the evaluation to estimation samples, i.e. $\rho = T_2/T_1$, where T_2 is the length of the sample used for forecast evaluation and T_1 is the length of the sample used for estimation. The values in the table present the p-values from the test at the nominal level $\alpha = 5\%$. The p-values in bold indicate rejection at the nominal size $\alpha = 5\%$. DM abbreviates Diebold and Mariano (1995) test of equal predictive ability and GW abbreviates Giacomini and White (2006) test of conditional predictive ability with $h_t = [1, \Delta\mathcal{L}_{t-1}]'$.

Table 11 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[10]. This is indicative of changing relative performance as we knew ex-ante, hence 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 inappropriate assumption.

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 $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ is known. For simplicity, we choose the $\mathbb{H}_0^{(3)}$ as our true data generating process for $\Delta\mathcal{L}_t$ and forecast the probability $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, starting from $\underline{T} = 100$.

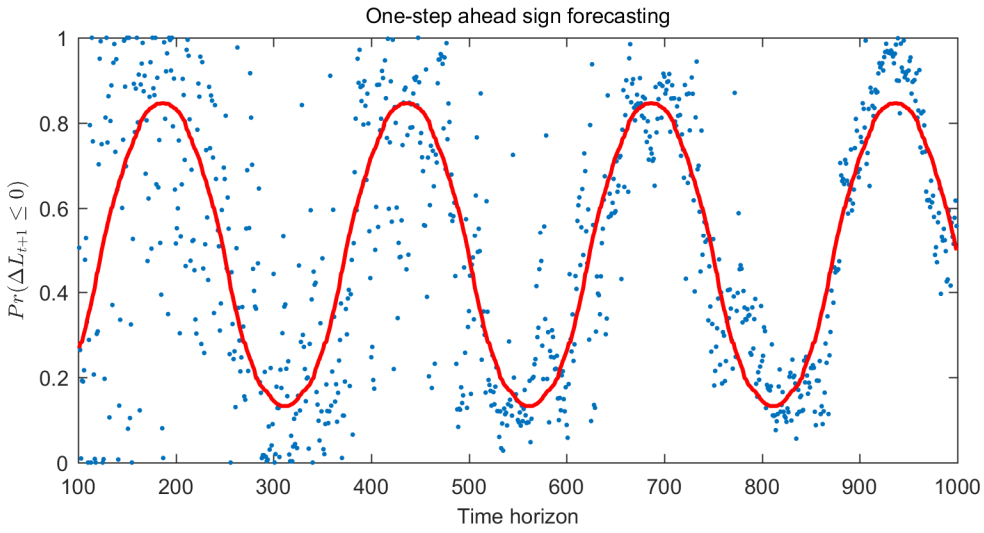


Figure 8: The red line plots the true probability $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ and the blue dots represent the estimate $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$.

[10]Note that this result is not specific to Diebold and Mariano (1995) test, but in fact to all existing tests that derive from it, including superior predictive ability tests, see White (2000), Hansen (2005) as well as Model Confidence Set test by Hansen et.al.(2011).

Figure 8 plots the true probability $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ against its estimate $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, 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 $\widehat{\varepsilon}_t$ 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. (15), we get $\widehat{C} = -0.052$, which points to the fact that on average our estimated probability $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ is on average overestimated by approximately 5.2%.

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 Motivating example in the Introduction

We consider the daily IBM returns spanning 03/01/2006-29/12/2016 and 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.(4) and constructed via the recursive scheme described in Section 2. We compute the 5 minute realized volatility 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 the $\{\Delta\mathcal{L}_t\}$ has been constructed, we apply our proposed two-step nonparametric procedure using AR(1) time-varying coefficient model (5) to estimate the corresponding time-varying mean and variance. Figure 9 depicts $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ and $\widehat{\tau}(t/T)$ calculated via eq.(9).

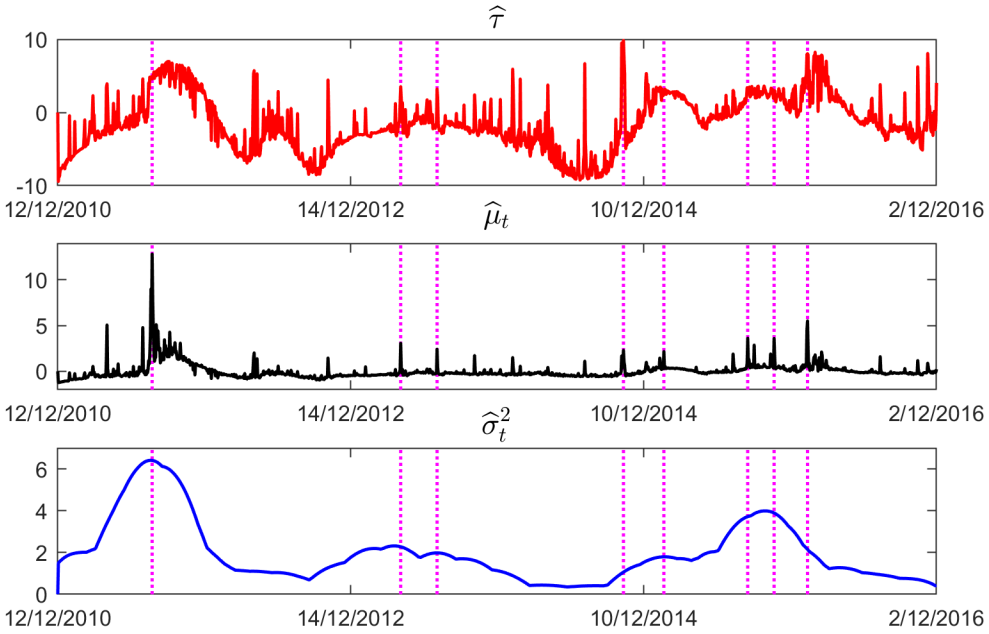


Figure 9: Plots of the estimates of $\widehat{\tau}$, $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ for IBM data, 2006-2016, using squared error loss and recursive forecasting scheme.

Recall that each corresponding $\widehat{\mu}_t$ is weighted by the inverse of the standard error of $\widehat{\mu}_t(u)$. One can approximately take the weight to be $1/\widehat{\sigma}_t$. 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 $\mathcal{S}_T$. Tables below show the critical values of two-sided and one-sided $\mathcal{S}_T$ test, specific to this application.

Table 12: Critical values for $\mathcal{S}_T$.

Quantiles	0.005	0.025	0.05	0.075	0.85	0.90	0.925	0.95	0.975	0.99	0.995
Cr. values	-2.48	-1.90	-1.64	-1.37	1.00	1.43	1.61	1.72	2.10	2.60	2.77

Note: The critical values are calculated via the wild bootstrap and are specific to the application at hand.

The value of the test statistic in this application example is $\mathcal{S}_T = -26.33$. Provided the critical values in Table 12, when the null of Equal Predictive Ability is tested, it is rejected at all levels of significance. Under the one-sided null of Superior Predictive Ability, the null is not rejected for all significance levels, indicating that there is no evidence that the GARCH(1,1) model with normal errors is superior to the GARCH(1,1) model with Student-t. Under remark 4 the practitioner

should default to choosing GARCH(1,1) with Student-t errors for forecasts. If supposing we test the opposite one-sided null, we find evidence that GARCH(1,1) with Student-t errors is superior to GARCH(1,1) with normal errors at all levels.

Below we present the results of our pseudo out-of-sample sign forecasts.

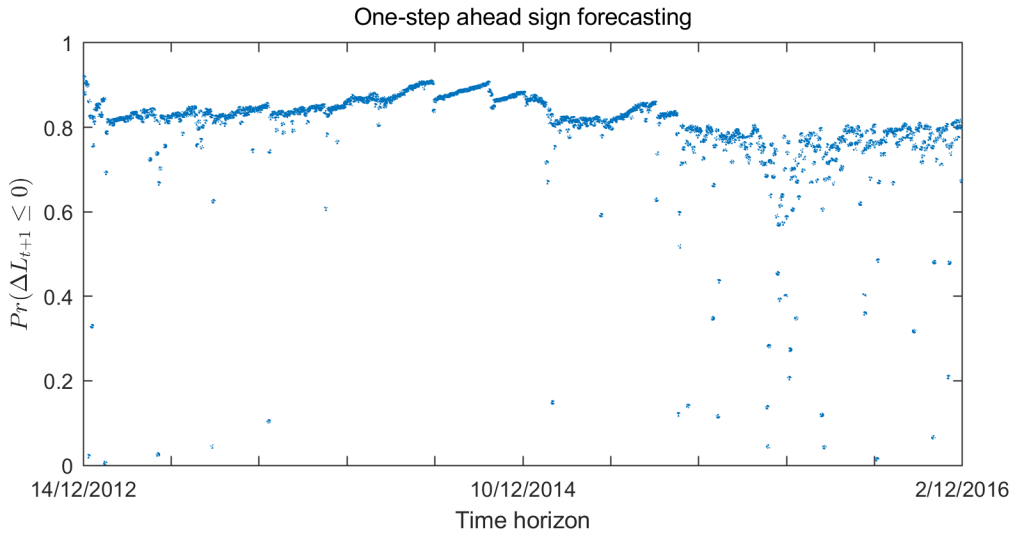


Figure 10: One-step ahead sign forecasting for the motivating example in the introduction.

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. (15) we get the value $\widehat{C} = -0.033$, indicating that our forecasted probabilities are on average overestimated by 3.3%. At the final point in the sample, we forecast a probability of 0.3129 with a corresponding forecast interval of $[0.2700,\ 0.3540]$. Interestingly, this probability does not conclude that GARCH(1,1) with Student-t errors should be selected. This highlights the randomness inherent in forecasting next period probabilities. In this case, our two approaches would select different models for forecasting.

7.2 Comparing parameter-reduction methods

In this section we condider an application similar to that considered in Giacomini and White (2006). We consider the “balanced panel" of the dataset FRED-MD, consisting of 128 monthly economic time series measured over January, 1959 - August, 2017, and apply the same transformations to the

original series, as documented in Appendix to the dataset[11]. In particular, compared to Giacomini and White (2006), we extend their dataset to the August of 2017. We replace the sequential model examined in Giacomini and White (2006) with lasso to avoid multiple sequential testing. We then use several parameter-reduction methods, 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).

All forecasting models project the $k-$step ahead variable of interest $\mathcal{Y}_{t+k}$ onto time t predictors $\mathcal{X}_t$ and lags of the variable of interest $\mathcal{Y}_t$, $\mathcal{Y}_{t-1}, \cdots$. We next describe the forecasting methods.

The full model for the k-step ahead forecast of the variable of interest $\mathcal{Y}_t$ is as follows:

$$\mathcal{Y}_{t+k} = \alpha + \beta \mathcal{X}_t + \gamma_1 \mathcal{Y}_t + \gamma_2 \mathcal{Y}_{t-1} + \cdots + \gamma_6 \mathcal{Y}_{t-5} + \varepsilon_{t+k}, \tag{21}$$

where $\mathcal{X}_t$ contains all 135 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 120 predictors.

The first method considers the full model (21) and applies lasso to determine the relevant predictors. Denote by $Z_t = (X_t', \mathcal{Y}_t, \mathcal{Y}_{t-1}, \cdots, \mathcal{Y}_{t-5})'$, then lasso estimates the parameter vector $\theta = (\alpha, \beta', \gamma_1, \gamma_2, \cdots, \gamma_6)'$ by solving

$$\widehat{\theta} := \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^{T} \|\mathcal{Y}_t - \theta' Z_t\|_2^2 + \lambda \|\theta\|_1 \right\},$$

where d is the dimension of the parameter vector θ and $\|\cdot\|_2$ and $\|\cdot\|_1$ denotes the L_2- and L_1-norms respectively.

The next model we consider is the diffusion index method (abbreviated DI) that first uses principal component analysis to estimate j factors $\widehat{F}_t$ from the predictors $\mathcal{X}_t$ and then considers the reduced model given by:

$$\mathcal{Y}_{t+k} = \alpha + \beta \widehat{F}_t + \gamma_1 \mathcal{Y}_t + \cdots + \gamma_p \mathcal{Y}_{t-p} + \varepsilon_{t+k},$$

where the lag length p is selected by BIC and the number of factors j is chosen by applying Onatski's (2009) test.

[11] The FRED-MD dataset is collected and constantly updated by the Federal Reserve Bank of St. Louse 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.

The Bayesian shrinkage method (abbreviated Bay) considers the full model (21) and applies Bayesian estimation with Normal-Gamma priors for the coefficients. Moreover, the Bayesian estimation is coupled with the use of the Elastic Net as a more stabilized version of lasso, see Zou and Hastie (2005), that also allows grouping effects. In particular, the Elastic Net estimator $\widehat{\theta}$ of the parameter vector $\theta = (\alpha, \beta', \gamma_1, \gamma_2, \cdots, \gamma_6)'$ is the solution of the following minimization problem:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(\mathcal{Y}_t - \theta' Z_t\right)^2 + \lambda_1 \sum_{j=1}^{d} |\theta_j| + \lambda_2 \sum_{j=1}^{d} \theta_j^2,$$

where d is the dimension of the parameter vector θ and $Z_t = (X_t', \mathcal{Y}_t, \mathcal{Y}_{t-1}, \cdots, \mathcal{Y}_{t-5})'$. We follow Korobilis (2013) for setting the priors. In particular, the Bayesian prior for θ in the above penalized regression is

$$\pi(\theta|\sigma^2) \sim e^{-\frac{\lambda_1}{\sqrt{\sigma^2}} \sum_{j=1}^{d} |\theta_j| - \frac{\lambda_2}{2\sigma^2} \sum_{j=1}^{d} \theta_j^2},$$

and for the shrinkage parameter $\tau_j, j = 1, \cdots, d$ the hyperprior on τ_j^2 is given by

$$\pi(\tau_j^2|\lambda_1^2) \sim \text{Exponential}\left(\frac{\lambda_1^2}{2}\right), \qquad \text{for} \qquad j = 1, \cdots, d,$$

which leads to the prior of the parameter vector have the following diagonal covariance matrix:

$$V = \begin{pmatrix} \left(\tau_1^{-2} + \lambda_2\right)^{-1} & & & & \\ & \left(\tau_2^{-2} + \lambda_2\right)^{-1} & & & \\ & & \ddots & & \\ & & & \left(\tau_{d-1}^{-2} + \lambda_2\right)^{-1} & \\ & & & & \left(\tau_d^{-2} + \lambda_2\right)^{-1} \end{pmatrix}.$$

The benchmark methods are the autoregressive model (denoted by AR) given by:

$$\mathcal{Y}_{t+k} = \alpha + \gamma_1 \mathcal{Y}_t + \gamma_2 \mathcal{Y}_{t-1} + \cdots + \gamma_6 \mathcal{Y}_{t-5} + \varepsilon_{t+k},$$

where p is selected by BIC and $0 \leq p \leq 6$, and the random walk model (denoted by RW) in levels, corresponding to the forecasting model in differences $\mathcal{Y}_{t+k} = \alpha + \varepsilon_{t+k}$, which therefore captures just the unconditional mean of the variable of interest. 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 the recursive scheme, described in Figure 2 with $\underline{T} = 100$.

Our one-sided test $\mathcal{S}_T$ tests the null of SPA, against an alternative of inferior predictive ability.

The loss differences are constructed from the loss of the model from the column model of our table, minus the loss of the model from the row model of our table. Therefore, a negative test statistic is indicative of SPA of the column model versus the row model, whereas a positive test statistic is indicative of the inferior predictive ability. Our decision rule is to select the column model whenever there is not significant evidence to reject the null of SPA. Otherwise we select the row model. We highlight the cases when we reject the null of SPA in bold. We also identify that in general, the Bayesian model performs consistently poorly for all four variables. Conversely, the Random Walk model performs in general the best, except for forecasting personal income where it is insignificantly worse than the AR model.

Our two-sided test allows us to construct our overall ranking for models. The results are presented in Tables 13-14. Significant rejections in either direction of our null is highlighted again in bold. For each of our variables of interest, we obtain the following rankings:

- Personal income: AR $\geq$ RW$\geq$ DI $\geq$ Lasso $\geq$ Bay;
- Industrial production: RW $\geq$ Lasso $\geq$ AR $\geq$ DI $>$ Bay;
- Producer price index: RW$>$Lasso $\geq$ DI $\geq$ AR $\geq$ Bay;
- Consumer price index: RW $\geq$ Lasso $>$ DI $\geq$ AR $>$ Bay.

In the above, $\geq$ indicates an insignificant superior ranking and $>$ indicates a significant superior ranking. We remark that for all four variables, the Bayesian shrinkage method is consistently the worst in terms of its forecasting ability for $k = 1$ month, and significantly so for industrial production and the consumer price index. Conversely, the random walk model performs the best (followed by lasso), for all variables except personal income where the AR model is insignificantly ranked higher.

We now proceed to applying the sign forecasting methodology, described in Section 3.3. For each of the variables we report the forecasted probability $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ at the end of the sample as well as the associated forecast interval $[\widehat{FI}_l,\ \widehat{FI}_u]$. Also, for all variables we perform our pseudo out-of-sample sign forecasting exercise, starting with $\underline{T} = 100$, and report the value of our criterion $\widehat{C}$. Results are presented in Table 15. From the results in Table 15 we can infer our ranking of models based on forecasted next period next period performance. We do so in the following way. We say that model $\mathcal{A}$ outperforms model $\mathcal{B}$, denoted as $\mathcal{A} > \mathcal{B}$, if $\widehat{FI}_l > 0.5$ and $\mathcal{A} \geq \mathcal{B}$ if $0.5 \in [\widehat{FI}_l, \widehat{FI}_u]$ and $Pr\left(\Delta\mathcal{L}^{\mathcal{AB}}_{T+1} < 0\right) > 0.5$. Using the above described notation the ranking is as follows:

- Personal income: Lasso > AR > RW > DI > Bay;

- Industrial production: Lasso > AR > RW > DI > Bay;
- Producer price index: AR > Lasso > RW > DI > Bay;
- Consumer price index: AR > Lasso$\geq$ RW > DI > Bay,

This ranking is to some degree similar to the ranking based on average past performance. In particular, at the bottom of our ranking we see that the Bayesian shrinkage method performs consistently the worst out of all models for all four variables. The diffusion index method (DI) performs also generally poorly, which is again consistent with our metric for past performance. One likely explanation of the poor performance of the DI method is the potential for overfitting, which translates into poor out-of-sample forecasting. The reason for the latter is that, in addition to the lags of the forecasted variable, the DI model also includes the k common factors extracted from the whole dataset, which might be irrelevant for forecasting in any particular period.

Interestingly, our sign forecasting approach to ranking indicates that the random walk model is not the best model for next period forecasting, as it is always dominated by either the lasso or the autoregressive model. In the case of the autoregressive model, it is likely because the autoregressive model can account for serial correlation in loss differences, which sometimes is the dominant feature in the data. Likewise, it appears that some of the time the lasso model is better able to capture next period performance than both the autoregressive model and the random walk model. We can also infer that our average performance metric, especially due 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. Hence the random walk model, as the conservative choice out of our selection of models, is often the best by our average performance metric. However forecasting one period ahead, we see it is the case that either or the autoregressive model will outperform. For the sake of brevity, we present the results for longer horizons, $k = 6$ and $k = 12$ months in Appendix C.

Table 13: Results for one-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.54					−2.13					−1.62					−2.92				
	p	0.936					0.980					0.948					0.998				
DI	$\mathcal{S}_T$	0.95	1.51				−0.64	2.11				−1.44	1.65				−4.21	2.94			
	p	0.172	0.086				0.768	**0.022**				0.938	**0.042**				1.00	**0.004**			
AR	$\mathcal{S}_T$	0.98	1.70	1.78			−0.46	2.17	0.08			−3.10	1.52	−1.79			−2.35	2.99	−1.46		
	p	0.164	**0.042**	**0.040**			0.666	**0.022**	0.468			0.998	0.082	0.958			0.992	**0.002**	0.936		
RW	$\mathcal{S}_T$	1.35	1.61	0.52	−1.76	−	0.91	2.06	0.41	1.48	−	4.75	1.50	2.58	2.96	−	1.70	3.02	3.65	2.90	−
	p	0.100	0.453	0.519	0.948	−	0.170	**0.028**	0.358	0.072	−	**0.000**	0.062	**0.008**	**0.004**	−	**0.046**	**0.000**	**0.000**	**0.005**	−

Note: Table reports the value of the one-sided test statistic $\mathcal{S}_T$ that corresponds to the null of superior predictive ability, see eq. (2), for horizon $k = 1$ month. The p-values are obtained via the wild 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.54$ (indicating that Lasso is better) with the p-value of 0.936. The p-values in bold indicate rejection of the null (2) at the 5% level of significance.

Table 14: Results for two-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.49					−2.04					−1.64					−3.07				
	p	0.152					**0.044**					0.172					**0.028**				
DI	$\mathcal{S}_T$	0.96	1.50				−0.64	2.20				−1.46	1.47				−4.32	2.99			
	p	0.304	0.124				0.524	**0.048**				0.148	0.124				**0.000**	**0.006**			
AR	$\mathcal{S}_T$	0.98	1.64	1.78			−0.46	2.13	0.08			−3.26	1.49	−1.47			−2.37	2.96	−1.40		
	p	0.296	0.132	0.052			0.628	**0.047**	0.92			**0.000**	0.108	0.068			**0.012**	**0.010**	0.152		
RW	$\mathcal{S}_T$	1.35	1.55	0.52	−1.80	−	0.86	2.02	0.58	1.75	−	4.96	1.54	2.59	3.08	−	1.63	2.98	3.52	3.02	−
	p	0.196	0.092	0.600	0.088	−	0.388	**0.048**	0.584	0.092	−	**0.000**	0.136	**0.008**	**0.000**	−	0.088	**0.007**	**0.000**	**0.000**	−

Note: Table reports the value of the two-sided test statistic $\mathcal{S}_T$ that corresponds to the null of equal predictive ability, see eq. (1), for horizon $k = 1$ month. The p-values are obtained via the wild 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.49$ (indicating that Lasso is better) with the p-value of 0.152. The p-values in bold indicate rejection of the null (1) at the 5% level of significance.

Table 15: Sign forecasting for $\Delta\mathcal{L}_{T+1}$ for horizon $k = 1$ month.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\widehat{Pr}_{T+1}$	0.983					1.000					0.999					0.999				
	$\widehat{FI}_l$	0.976					1.000					0.997					0.998				
	$\widehat{FI}_u$	0.990					1.000					1.000					1.000				
	$\widehat{C}$	0.060					0.014					0.056					0.045				
DI	$\widehat{Pr}_{T+1}$	0.884	0.001				0.836	0.000				0.550	0.001				0.755	0.002			
	$\widehat{FI}_l$	0.839	0.000				0.809	0.000				0.520	0.000				0.716	0.001			
	$\widehat{FI}_u$	0.905	0.005				0.849	0.000				0.578	0.006				0.771	0.003			
	$\widehat{C}$	-0.024	-0.078				-0.032	-0.014				0.067	-0.068				0.071	-0.032			
AR	$\widehat{Pr}_{T+1}$	0.954	0.001	0.089			0.537	0.000	0.212			0.319	0.001	0.570			0.436	0.001	0.412		
	$\widehat{FI}_l$	0.912	0.000	0.064			0.510	0.000	0.195			0.281	0.000	0.550			0.386	0.000	0.385		
	$\widehat{FI}_u$	0.986	0.002	0.126			0.560	0.000	0.246			0.344	0.004	0.600			0.460	0.002	0.456		
	$\widehat{C}$	-0.018	-0.105	-0.049			-0.090	-0.014	-0.017			0.037	-0.049	-0.034			0.058	-0.039	-0.025		
RW	$\widehat{Pr}_{T+1}$	0.829	0.002	0.078	0.816		0.660	0.000	0.181	0.624		0.944	0.002	0.380	0.669		0.978	0.001	0.162	0.523	
	$\widehat{FI}_l$	0.802	0.001	0.066	0.790	–	0.628	0.000	0.160	0.597	–	0.917	0.001	0.348	0.620	–	0.920	0.000	0.141	0.482	–
	$\widehat{FI}_u$	0.861	0.004	0.115	0.839	–	0.690	0.000	0.214	0.654	–	0.954	0.004	0.413	0.706	–	0.997	0.001	0.201	0.572	–
	$\widehat{C}$	-0.026	-0.085	-0.023	0.041		-0.015	-0.015	-0.011	0.001		-0.048	-0.069	-0.053	-0.034		-0.037	-0.043	-0.061	-0.036	

Note: Table reports the results of the sign forecasting for $\Delta\mathcal{L}_{T+1}$ for the forecast horizon $k = 1$ month. $\widehat{Pr}_{T+1}$ is an abbreviation of $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, i.e. the forecasted probability at the very end of the sample. $\widehat{FI}_u$ and $\widehat{FI}_l$ denotes the upper and lower bounds of the forecast interval, such that $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) \in [\widehat{FI}_l,\ \widehat{FI}_u]$. Finally, $\widehat{C}$ denotes the value of the criterion in eq.(15). 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$, for which $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = 0.983$ with the corresponding forecast interval $[0.976,\ 0.990]$.

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 methods by which to inform the selection of models for future forecasts. Importantly, our new methodology is no longer reliant on a sample splitting point, which is directly connected to the two limitations of the existing out-of-sample tests. 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.

References

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

Bradley, R.C., 2005, Basic properties of strong mixing conditions. A survey and some open questions, *Probability Surveys*, 2, 107-144.

Cai, Z., 2007, Trending time-varying coefficient time series models with serially correlated errors, *Journal of Econometrics*, 136, 163-288.

Cai, Z., Fan, J., and Li, R.Z., 2000, Efficient estimation and inferences for varying-coefficient models, *Journal of American Statistical Association*, 95, 888-902.

Cai, Z., Fan, J., and Yao, Q., 2000, 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, *Econometrics Reviews*, 24, 369-404.

Diebold, F.X., 2013, Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of diebold-mariano tests, *Journal of Business and Economic Statistics*.

Diebold, F.X., and Mariano R.S., 1995, Comparing predictive accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

Fan J., and Yao Q., 1998, Efficient estimation of conditional variance functions in stochastic regression, *Biometrika*, 85(3), 645-660.

Giacomini R., and Rossi B., 2009, Detecting and predicting forecast breakdowns, *The Review of Economic Studies* 76(2), 669-705.

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, Test of conditional predictive ability, *Econometrica* 74(6), 1545-1578.

Corradi V., and Swanson, N.R., 2007, Nonparametric bootstrap procedures for predictive inference based on the recursive estimation schemes, *International Economic Review*, 48, 67-109.

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

Hansen B.E., 2008, Least squares forecast averaging," *Journal of Econometrics*, 146, 342-350.

Hansen B.E., 2007, Least squares model averaging," *Econometrica*, 75, 1175-1189.

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, *Stanford University Working paper.*

Hansen P.R., and Timmermann, A., 2015, Discussion of comparing predictive accuracy, twenty years later by Francis X. Diebold, *Journal of Business and Economics Statistics*, 33, 17-21.

Hansen P. R., and Timmermann, A., 2010, Choice of sample split in out-of-sample forecast evaluation, Working paper.

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

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., 2004, In-sample or out-of-sample tests of predictability: which one should we use?, *Econometric Reviews*, 23(4), 371-402.

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

Korobilis, D., 2013, Hierarchical shrinkage priors for dynamic regressions with many predictors, *International Journal of Forecasting*, 29, 43-59.

Kristensen, D., 2009, Uniform convergence rates of kernel estimators with heterogeneous dependent data, *Econometric Theory*, 25(5), 1433-1445.

Kristensen, D., 2012, Non-parametric detection and estimation of structural change, *Econometrics Journal*, 15, 420-461.

Li, J., and Patton, A.J., 2017, Asymptotic inference about predictive accuracy using high-frequency, *Duke University Working paper*.

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, 195-223.

McCracken, M.W., 2007, Asymptotics for out-of-sample tests of granger causality, *Journal of Econometrics*, 140, 719-752.

Mincer, J., and Zarnowitz, V., 1969, The evaluation of economic forecasts, in *Economic Forecasts and Expectations*, ed. J. Mincer, New York: National Bureau of Economic Research.

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

Orbe, S., Ferreira, E., and Rodriguez-Poo, J., 2005, Nonparametric estimation of time varying parameters under the shape restrictions, *Journal of Econometrics*, 126, 53-77.

Pollard, D., 1984, *Convergence of stochastic processes*, Springer, New York.

Stinchcombe, M.B., and White, H., 1998, Consistent specification testing with nuisance parameters present only under the alternative, *Econometric Theory*, 14, 295-325.

Vogt, M., 2012, Nonparametric regression for locally stationary time series, *Annals of Statistics*, 40, 2601-2633.

West, K.D., 1996, Asymptotic inference about predictive ability, *Econometrica*, 64, 1067-1084.

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

Zou, H., and Hastie, T., 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, 67, 301-320.

9 Appendix A.

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

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

where ε_t is a m.d.s. Assume one uses two simple models to forecast y_t one-step ahead:

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

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

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

and

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

and therefore

$$\Delta\mathcal{L}_t^{\mathcal{AB}} = \mathcal{L}_t^{\mathcal{A}} - \mathcal{L}_t^{\mathcal{B}} = 0.01 - 0.2\rho y_t. \tag{22}$$

From eq.(22) it then follows that

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

10 Appendix B.

This Appendix presents proofs of the theoretical results. Throughout the proofs for brevity of notation we will drop the second subscript in the triangular array notation, i.e. for any variable $X_{t,T}$ we will just write X_t. Moreover, throughout the Appendix, the symbol C denotes a universal real constant which may take a different value on each occurrence.

Proof of Theorem 1.

The proof is based on that of Kristensen (2012) by extending the notation to accommodate the locally linear estimator of $\theta(t/T)$. We first lay out the notation used in establishing Theorem 1, since the rest of subsequent theory will use the same notation. Recall that we model $\Delta\mathcal{L}_t$ as the following time-varying coefficient model:

$$\Delta\mathcal{L}_t = \rho_0(t/T) + \sum_{t=1}^{d} \rho_j(t/T)\Delta\mathcal{L}_{t-j} + \xi_t, \qquad \xi_t = \sigma(t/T)\varepsilon_t.$$

We further make use of the following notation: $\mathbb{X}_t = (1, \Delta\mathcal{L}_{t-1}, \cdots, \Delta\mathcal{L}_{t-d})^T$ and $\rho(t/T) = (\rho_0(t/T), \rho_1(t/T), \cdots, \rho_d(t/T))^T$. In addition, using the first-order Taylor approximation we have:

$$\rho_j(t/T) = a_j + b_j\,(t/T - u) + o\left(|t/T - u|\right), \qquad 0 \leq j \leq d,$$

where $a_j = \rho_j(u)$ and $b_j = \dot{\rho}_j(u)$. Denote further by $\mathbb{Z}_t = \left(\mathbb{X}_t^T, \mathbb{X}_t^T\,(t/T - u)\right)^T$ and $\theta = \theta(u) = \left(\rho^T(u), \dot{\rho}^T(u)\right)^T$. Then the locally weighted least squares is

$$\widehat{\theta}(u) = \arg\min_{\theta} \sum_{t=1}^{T} K_{h_1}\left(t/T - u\right)\left(\Delta\mathcal{L}_t - \mathbb{Z}_t^T\theta\right)^2. \tag{23}$$

Minimising (23) with respect to θ provides the local linear estimator of $\rho(u)$, denoted by $\widehat{\rho}(u)$, which are the first $(d+1)$ elements of $\widehat{\theta}$ and the local linear estimator of the derivatives of $\rho(u)$, denoted by $\widehat{\dot{\rho}}(u)$, which are the last $(d+1)$ elements of $\widehat{\theta}$. It is straightforward to show that

$$\widehat{\theta}(u) = \begin{bmatrix} \Sigma_{T,0}(u) & \Sigma_{T,1}(u) \\ \Sigma_{T,1}(u) & \Sigma_{T,2}(u) \end{bmatrix}^{-1} \begin{pmatrix} W_{T,0}(u) \\ W_{T,1}(u) \end{pmatrix} = \Sigma_T^{-1}(u) W_T(u),$$

where

$$\Sigma_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T - u\right)\left(t/T - u\right)^m \mathbb{X}_t\mathbb{X}_t^T, \qquad \text{for} \qquad m = 0, 1, 2,$$

and

$$W_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^m \mathbb{X}_t \Delta\mathcal{L}_t, \qquad \text{for} \qquad m=0,1.$$

Recall that $\Omega_{t,T} \equiv \mathbb{E}\left[\mathbb{X}_t\mathbb{X}_t^T\right] = \Omega(t/T) + o(1)$ and $H = \text{diag}\left(I_{d+1}, h_1 I_{d+1}\right)$ with I_{d+1} being the $(d+1)\times(d+1)$ identity matrix. In addition, we define the following quantities:

$$V(t/T,u) = \rho(t/T) - \left\{\rho(u) + \dot{\rho}(u)(t/T-u) + \frac{1}{2}\ddot{\rho}(u)(t/T-u)^2\right\} \tag{24}$$

$$\widetilde{W}_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^m \mathbb{X}_t\xi_t, \quad \mathbb{B}_{T,m}(u) = \frac{1}{2}\Sigma_{T,m+2}(u)\ddot{\rho}(u),$$

and

$$\mathbb{R}_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^m \mathbb{X}_t\mathbb{X}_t^T V(t/T,u). \tag{25}$$

With the above notation we can rewrite $W_{T,m}$ for $m=0,1$ as follows:

$$W_{T,m} = \Sigma_{T,m}(u)\rho(u) + \Sigma_{T,m+1}(u)\dot{\rho}(u) + \widetilde{W}_{T,m}(u) + \mathbb{B}_{T,m}(u) + \mathbb{R}_{T,m}(u).$$

Then it holds that

$$\widehat{\theta}(u) - \theta(u) - \Sigma_T^{-1}(u)\mathbb{B}_T(u) - \Sigma_T^{-1}(u)\mathbb{R}_T(u) = \Sigma_T^{-1}(u)\widetilde{W}_T(u). \tag{26}$$

To establish the proof of Theorem 1 it remains to show the following results:

(C1) $\sup_{\theta\in I_{h_1}} \|\widehat{\theta}(u) - \theta(u)\| = O_p\left(\sqrt{\frac{\log T}{Th_1}} + h_1^2\right)$, where $I_{h_1} := [Ch_1, 1-Ch_1]$ and $C>0$ such that $Ch_1 \to 0$ and $1/C \to 0$,

(C2) $H\{\widehat{\theta}(u) - \theta(u)\} = h_1^2\mathbb{B}_1(u) + o_p(1)$, and

(C3) $\sqrt{Th_1}H^{-1}\widetilde{W}_T(u) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_\theta(\theta)\right)$, where $\mathbb{B}_1(u)$ and $\mathbb{V}_\theta(u)$ are given in Theorem 1.

Proof of (C1) and (C2). In light of Assumptions $A1$ and $A2$ the process $\Delta\mathcal{L}_t$ is strongly mixing and $\sup_{t,T}\mathbb{E}\left[|\Delta\mathcal{L}_{t,T}|^s\right] < \infty$ for $s>4$, see Orbe et. al.(2005) Lemma A4 for a proof of this result. Once the mixing condition of $\Delta\mathcal{L}_t$ is established as well as the finiteness of its moments, in conjunction with Assumptions $A1-A3$ we now satisfy e.g. Assumptions $A1-A6$ for Theorem 1 in Kristensen

(2009) or Assumptions $K1-K3$ for Theorem 4.1 in Vogt (2012) to conclude that

$$\sup_{\theta\in I_{h_1}}\|\widehat{\theta}(u)-\theta(u)\|=O_p\left(\sqrt{\frac{\log T}{Th_1}}+h_1^2\right),$$

where $I_{h_1}:=[Ch_1,1-Ch_1]$ and $C>0$ such that $Ch_1\to 0$ and $1/C\to 0$. To establish (C2) note that $V(t/T,u)=o_p(h_1^2)$ and we therefore can ignore this term and focus on $\mathbb{B}_{T.m}(u)$ for $m=0,1$. First consider $\Sigma_{T,m}(u)$ terms. Using Riemann sum approximation of an integral the following holds:

$$\begin{aligned}h_1^{-m}\mathbb{E}\left[\Sigma_{T,m}(u)\right]&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{X}_t\mathbb{X}_t^T K_{h_1}\left(t/T-u\right)\left(\frac{t/T-u}{h_1}\right)^m\right]=\\&=\frac{1}{T}\sum_{t=1}^{T}K_{h_1}\left(t/T-u\right)\left(\frac{t/T-u}{h_1}\right)^m\Omega(t/T)+o(1)=\int_{-1}^{1}y^mK(y)\Omega(u+yh_1)dy+o(1)=\\&=\lambda_m\Omega(u)+o(1).\end{aligned}$$

Therefore it holds that

$$h_1^{-m}\Sigma_{T,m}(u)=\lambda_m\Omega(u)\{1+o_p\},\quad\text{i.e.}\quad h_1^{-m}\Sigma_{T,m}(u)\xrightarrow{p}\lambda_m\Omega(u),\tag{27}$$

and

$$H^{-1}\mathbb{B}_T(u)=\frac{h_1^2}{2}\begin{pmatrix}\lambda_2\Omega(u)\\0\end{pmatrix}\otimes\ddot{\rho}(u)+o_p(h_1^2),\tag{28}$$

We can therefore re-write (26) as follows:

$$\begin{aligned}&\sqrt{Th_1}\left(H\{\widehat{\theta}(u)-\theta(u)\}-\frac{h_1^2}{2}\begin{pmatrix}\lambda_2\Omega(u)\\0\end{pmatrix}\otimes\ddot{\rho}(u)+o_p(h_1^2)\right)=\\&=\left(H^{-1}\Sigma_T(u)H^{-1}\right)^{-1}\sqrt{Th_1}H^{-1}\widetilde{W}_T(u)=\Sigma(u)^{-1}\sqrt{Th_1}H^{-1}\widetilde{W}_T(u)\{1+o_p(1)\}=O_p(1),\end{aligned}\tag{29}$$

where in light of (27)

$$\Sigma(u)=\begin{pmatrix}\Omega(u)&0\\0&\lambda_2\Omega(u)\end{pmatrix}.$$

Proof of (C3). Define $Y_{t,T}^m:=h_1^{-m-1/2}K\left(\frac{t/T-u}{h_1}\right)(t/T-u)^m\mathbb{X}_t\xi_t$, which is a martingale difference sequence w.r.t. $\mathcal{F}_t=\sigma(\Delta\mathcal{L}_t,\varepsilon_t,\Delta\mathcal{L}_{t-1},\varepsilon_{t-1},\cdots)$. To complete the proof of (C3) it suffices to verify Lemma B.13 in Kristensen (2012) for $Y_{t,T}^m$, $m=0,1$. In particular, for $m=0,1$ as $Th_1\to\infty$ it holds

that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[Y_{t,T}^{m}(Y_{t,T}^{m})'\right]=\frac{1}{Th_1}\sum_{t=1}^{T}K\left(\frac{t/T-u}{h_1}\right)^{2}\left(\frac{t/T-u}{h_1}\right)^{2m}\Omega(t/T)\sigma(t/T)^2+o(1)=$$

$$=\frac{1}{h_1}\int_{-1}^{1}K^2\left(\frac{y-u}{h_1}\right)\left(\frac{y-u}{h_1}\right)^{2m}\sigma^2(y)\Omega(y)dy+o(1)=\nu_{2m}\sigma^2(u)\Omega(u)+o(1),$$

and

$$\frac{1}{T^{1+\delta/2}}\sum_{t=1}^{T}\mathbb{E}\left[\|Y_{t,T}^{m}\|^{2+\delta}\right]=\frac{1}{(Th_1)^{1+\delta/2}}\sum_{t=1}^{T}K^{2+\delta}\left(\frac{t/T-u}{h_1}\right)\left(\frac{t/T-u}{h_1}\right)^{(2+\delta)m}\mathbb{E}\left[\|\mathbb{X}_t\|^{2+\delta}|\varepsilon_t|^{2+\delta}\right]=$$

$$=\frac{C}{(Th_1)^{\delta/2}}\sigma^{2+\delta}(u)\int_{-1}^{1}K^{2+\delta}(y)y^{m(2+\delta)}dy=o(1).$$

Therefore $\sqrt{Th_1}H^{-1}\widetilde{W}_T\xrightarrow{d}\mathcal{N}\left(0,\,\Xi(u)\right)$, where

$$\Xi(u)=\begin{pmatrix}\nu_0\sigma^2(u)\Omega(u) & 0\\ 0 & \nu_2\sigma^2(u)\Omega(u)\end{pmatrix},$$

Combining all of the above with (29), it then follows that

$$\sqrt{Th_1}\left(H\{\widehat{\theta}(u)-\theta(u)\}-\frac{h_1^2}{2}\begin{pmatrix}\lambda_2\Omega(u)\\ 0\end{pmatrix}\otimes\ddot{\rho}(u)+o_p(h_1^2)\right)\xrightarrow{d}\mathcal{N}\left(0,\,\Sigma(u)^{-1}\Xi(u)\Sigma(u)^{-1}\right),$$

which completes the proof of Theorem 1. ■.

Proof of Theorem 2.

Define the estimated errors $\widehat{\xi}_t$:

$$\widehat{\xi}_t=\Delta\mathcal{L}_t-\mathbb{Z}_t^T\widehat{\theta}(t/T)=\mathbb{Z}_t^T\theta(t/T)+\xi_t-\mathbb{Z}_t^T\widehat{\theta}(t/T)=\mathbb{Z}_t^T\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}+\xi_t.$$

Running the local constant nonparametric regression of $\widehat{\xi}_t^2$ on rescaled time we get:

$$\widehat{\sigma}^2(u) = \frac{\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\widehat{\xi}_t^2}{\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)}.$$

Note also that

$$\widehat{\xi}_t^2 = \mathbb{Z}_t^T\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t + \xi_t^2 + 2\mathbb{Z}_t^T\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}\xi_t,$$

and therefore

$$\begin{aligned}\widehat{\xi}_t^2 - \xi_t^2 = \mathbb{Z}_t^T\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t + 2\mathbb{Z}_t^T\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}\xi_t = \\ = \mathbb{Z}_t^T\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t - 2\mathbb{Z}_t^T\left\{\widehat{\theta}(t/T) - \theta(t/T)\right\}\xi_t.\end{aligned}$$

We therefore can write:

$$\begin{aligned}\widehat{\sigma}^2(u) - \sigma^2(u) &= \frac{1}{T\widehat{f}(u)}\sum_{t=1}^{T} K_{h_2}\left(t/T - h\right)\left[\widehat{\xi}_t^2 - \sigma^2(u)\right] = \\ &= \frac{1}{T\widehat{f}(u)}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\left[\left(\mathbb{Z}_t^T\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t + \sigma^2(t/T)\varepsilon_t^2 + \right.\right. \\ &\qquad \left.\left. - 2\mathbb{Z}_t^T\left\{\widehat{\theta}(t/T) - \theta(t/T)\right\}\sigma(t/T)\varepsilon_t\right) - \sigma^2(u)\right] = \\ &= \frac{1}{T\widehat{f}(u)}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\left[\sigma^2(t/T) - \sigma^2(u)\right] + \frac{1}{T\widehat{f}(u)}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\sigma^2(t/T)\left\{\varepsilon_t^2 - 1\right\} + \\ &\qquad - \frac{2}{T\widehat{f}(u)}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\mathbb{Z}_t^T\{\widehat{\theta}(t/T) - \theta(t/T)\}\sigma(t/T)\varepsilon_t + \\ &\qquad + \frac{1}{T\widehat{f}(u)}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\mathbb{Z}_t^T\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}\left\{\theta(t/T) - \widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t.\end{aligned}$$

To make it easier to work with the above expression, we write it as follows:

$$\widehat{\sigma}^2(u) - \sigma^2(u) = \frac{1}{\widehat{f}(u)}\left(\widehat{\mathcal{I}}_1(u) + \widehat{\mathcal{I}}_2(u) + \widehat{\mathcal{I}}_3(u) + \widehat{\mathcal{I}}_4(u) - \sigma^2(u)\widehat{f}(u)\right),$$

with

$$\widehat{f}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right),$$

$$\widehat{\mathcal{I}}_1(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\sigma^2(t/T),$$

$$\widehat{\mathcal{I}}_2(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\sigma^2(t/T)\{\varepsilon_t^2 - 1\},$$

$$\widehat{\mathcal{I}}_3(u) = -\frac{2}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\mathbb{Z}_t^T\{\widehat{\theta}(t/T) - \theta(t/T)\}\sigma(t/T)\varepsilon_t$$

and

$$\widehat{\mathcal{I}}_4(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\mathbb{Z}_t^T\{\widehat{\theta}(t/T) - \theta(t/T)\}\{\widehat{\theta}(t/T) - \theta(t/T)\}^T\mathbb{Z}_t.$$

We start by deriving some preliminary results for the expressions above:

i) By applying Theorem 4.1 in Vogt (2012) by setting $d = 0$ and $W_{t,T} = 1$ we arrive at the following result:

$$\sup_{u\in I_{h_2}}\left|\widehat{f}(u) - f(u)\right| = o_p(1). \tag{30}$$

Moreover, (30) together with an extra condition that $\inf_{u\in[0,1]} f(u) > 0$ implies that

$$\sup_{u\in(0,1)}\widehat{f}(u)^{-1} = O_p(1).$$

ii) By applying Theorem 4.1 in Vogt (2012) by setting $d = 0$ and $W_{t,T} = \widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u)$ we arrive at the following result:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u) - \mathbb{E}\left[\widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u)\right]\right| = O_p\left(\sqrt{\frac{\log T}{Th_2}}\right).$$

iii)

$$\sup_{u\in I_{h_2}}\left|\mathbb{E}\left[\widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u)\right]\right| = O_p\left(h_2^2\right).$$

(iv) Applying Theorem 4.1 of Vogt (2012) to $\widehat{I}_2(u)$ yields:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_2(u)\right| = O_p\left(\sqrt{\frac{\log T}{Th_2}}\right).$$

(v)

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_3(u)\right| = O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right).$$

(vi)

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_4(u) - \mathcal{I}_4(u)\right| = O_p\left(\frac{1}{T^2h_1^2}\sqrt{\frac{\log T}{Th_2}}\right).$$

Proof of (v). We next turn to $\widehat{\mathcal{I}}_3$:

$$\widehat{\mathcal{I}}_3(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\mathcal{Z}_t^T\{\widehat{\theta}(t/T)-\theta(t/T)\}\xi_t =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t\mathcal{Z}_t^T\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u)+\Sigma_T^{-1}(u)\widetilde{W}_T(u)\right] = \widehat{\mathcal{I}}_{31}(u)+\widehat{\mathcal{I}}_{32}(u).$$

In addition, for the simplicity of exposition of further results, we partition $\Sigma(u)^{-1}$ into 4 submatrices each of dimension $(d+1)\times(d+1)$:

$$\Sigma(u)^{-1} = \begin{bmatrix}\widetilde{\Sigma}_{11}(u) & \widetilde{\Sigma}_{12}(u)\\ \\ \widetilde{\Sigma}_{21}(u) & \widetilde{\Sigma}_{22}(u)\end{bmatrix}.$$

In addition we denote $\{\ddot{\rho}(u)\}_{1:(d+1)}$ denotes the first $d+1$ elements of $\ddot{\rho}(u)$ and $\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}$

denotes the last $d+1$ elements of $\ddot{\rho}(u)$. We start with $\widehat{\mathcal{I}}_{31}(u)$:

$$\widehat{\mathcal{I}}_{31}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathbb{Z}_t^T \Sigma_T^{-1}(u)\mathbb{B}_T(u) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathbb{Z}_t^T H^{-1}\left[H^{-1}\Sigma_T(u)H^{-1}\right]^{-1} H^{-1}\mathbb{B}_T(u) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathbb{Z}_t^T H^{-1}\left[\Sigma^{-1}(u)+o_p(1)\right] H^{-1}\mathbb{B}_T(u) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{B}_T(u) + o_p(1) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathrm{tr}\left[\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{B}_T(u)\right] + o_p(1) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathrm{tr}\left[\Sigma^{-1}(u)H^{-1}\mathbb{B}_T(u)\mathbb{Z}_t^T H^{-1}\right] + o_p(1) = \frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \times$$

$$\times\left\{\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^2 \mathrm{tr}\left[\widetilde{\Sigma}_{11}(u)H^{-1}\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T H^{-1}\right] + \right.$$

$$\left. + \sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^4 \mathrm{tr}\left[\widetilde{\Sigma}_{22}(u)H^{-1}\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_t^T H^{-1}\right]\right\} + o_p(1) =$$

$$= \frac{h_1^2}{T^2 h_1 h_2}\sum_{t=1}^{T}\sum_{s=1}^{T} K\left(\frac{t/T-u}{h_2}\right)\xi_t K\left(\frac{s/T-u}{h_1}\right)\times$$

$$\times\left\{\left(\frac{s/T-u}{h_1}\right)^2 \mathrm{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_s^T\right] + \right.$$

$$\left. + \left(\frac{s/T-u}{h_1}\right)^4 \mathrm{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_s^T\right]\right\} + o_p(1).$$

Therefore, provided that $|K(x)| \leq 1$:

$$\left|\widehat{\mathcal{I}}_{31}(u)\right| \leq \frac{h_1^2}{Th_1h_2}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} K\left(\frac{t/T-u}{h_2}\right)\xi_t K\left(\frac{s/T-u}{h_1}\right)\times\right.$$
$$\times\left\{\left(\frac{s/T-u}{h_1}\right)^2 \operatorname{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_s^T\right]+\right.$$
$$\left.\left.+\left(\frac{s/T-u}{h_1}\right)^4 \operatorname{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_s^T\right]\right\}\right| \leq$$

$$\leq \frac{h_1^2}{Th_1h_2}\left|\frac{1}{T}\sum_{t=1}^{T} K\left(\frac{t/T-u}{h_2}\right)K\left(\frac{t/T-u}{h_1}\right)\xi_t\left\{\left(\frac{t/T-u}{h_1}\right)^2 \operatorname{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T\right]+\right.\right.$$
$$\left.\left.+\left(\frac{t/T-u}{h_1}\right)^4 \operatorname{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_t^T\right]\right\}\right| \leq$$
$$\leq \frac{h_1^2}{Th_1}\left|\frac{1}{Th_2}\sum_{t=1}^{T} K\left(\frac{t/T-u}{h_2}\right)\xi_t\left\{\left(\frac{t/T-u}{h_1}\right)^2 \operatorname{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T\right]+\right.\right.$$
$$\left.\left.+\left(\frac{t/T-u}{h_1}\right)^4 \operatorname{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_t^T\right]\right\}\right|$$

Finally applying Theorem 4.1 in Vogt (2012) by setting $d = 0$ and

$$W_{t,T} = \xi_t\left(\frac{t/T-u}{h_1}\right)^j \operatorname{tr}\left[\widetilde{\Sigma}_{ii}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T\right], \quad \text{for} \quad i = 1,2 \quad \text{and} \quad j = 2,4.$$

yields the final result:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_{31}(u)\right| = O_p\left(\frac{h_1}{T}\sqrt{\frac{\log T}{Th_2}}\right). \tag{31}$$

We now consider $\widehat{\mathcal{I}}_{32}(u)$:

$$\begin{aligned}\widehat{\mathcal{I}}_{32}(u) &= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathbb{Z}_t^T H^{-1}\left(H^{-1}\Sigma_T(u)H^{-1}\right)^{-1}H^{-1}\widetilde{W}_T(u) = \\ &= \frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)H^{-1}\mathbb{Z}_t\xi_t + o_p(1) = \\ &= \frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)K_{h_1}\left(t/T-u\right)\xi_t^2 \mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_t + \\ &+ \frac{2}{T^2}\sum_{j=1}^{T-1} K_{h_2}\left(t/T-u\right)K_{h_1}\left((t-j)/T-u\right)\xi_t\xi_{t-j}\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_{t-j} + o_p(1).\end{aligned}$$

In addition, it holds that

$$\begin{aligned}\left|\widehat{\mathcal{I}}_{32}(u)\right| &= \left|\frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)K_{h_1}\left(t/T-u\right)\xi_t^2 \mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_t + \right. \\ &\left. + \frac{2}{T^2}\sum_{j=1}^{T-1} K_{h_2}\left(t/T-u\right)K_{h_1}\left((t-j)/T-u\right)\xi_t\xi_{t-j}\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_{t-j} + o_p(1)\right| \leq \\ &\leq \frac{1}{Th_1}\left|\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t^2 \mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_t\right| + \\ &+ \frac{2}{Th_1}\left|\frac{1}{T}\sum_{j=1}^{T-1} K_{h_2}\left(t/T-u\right)\xi_t\xi_{t-j}\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_{t-j}\right|.\end{aligned}$$

Therefore applying Theorem 4.1 of Vogt (2012) to the appropriate terms in the above expression we arrive at the following results:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_{32} - \mathbb{E}\left[\widehat{\mathcal{I}}_{32}\right]\right| = O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right), \tag{32}$$

and

$$\sup_{u\in I_{h_2}}\left|\mathbb{E}\left[\widehat{\mathcal{I}}_{32} - \mathcal{I}_{32}\right]\right| = O_p\left(\frac{1}{Th_1}\right) \tag{33}$$

and

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_{32}(u) - \mathcal{I}_{32}\right| = O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right). \tag{34}$$

Combining (31)-(34) yields:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_3(u)-\mathcal{I}_3(u)\right|=O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right).$$

Proof of (vi). Finally we address $\widehat{\mathcal{I}}_4(u)$ term:

$$\begin{aligned}
&\widehat{\mathcal{I}}_4(u)=\\
&=\frac{1}{T}\sum_t^T K_{h_2}(t/T-u)\,\mathbb{Z}_t\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u)+\Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u)+\Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]^T\mathbb{Z}_t=\\
&\qquad=\frac{1}{T}\sum_t^T K_{h_2}(t/T-u)\,\mathrm{tr}\Bigg\{\mathbb{Z}_t\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u)+\Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]\times\\
&\times\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u)+\Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]^T\mathbb{Z}_t\Bigg\}=\frac{1}{T}\sum_t^T K_{h_2}(t/T-u)\,\mathrm{tr}\Bigg\{\Bigg[\Sigma_T^{-1}(u)\Big(\mathbb{B}_T(u)\mathbb{B}_T^T(u)+\\
&\qquad+2\mathbb{B}_T(u)\widetilde{W}_T^T(u)+\widetilde{W}_T(u)\widetilde{W}_T^T(u)\Big)\Sigma_T^{-1}(u)\Bigg]\mathbb{Z}_t\mathbb{Z}_t^T\Bigg\}.
\end{aligned}$$

Taking the expectation of the above expression, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{I}}_4(u)\right]&=\frac{1}{T}\mathbb{E}\sum_t^T K_{h_2}(t/T-u)\,\mathrm{tr}\Bigg\{\Bigg[\Sigma_T^{-1}(u)\Big(\mathbb{B}_T(u)\mathbb{B}_T^T(u)+\\
&\qquad+2\mathbb{B}_T(u)\widetilde{W}_T^T(u)+\widetilde{W}_T(u)\widetilde{W}_T^T(u)\Big)\Sigma_T^{-1}(u)\Bigg]\mathbb{Z}_t\mathbb{Z}_t^T\Bigg\}=\\
&=\frac{1}{T}\mathbb{E}\sum_t^T K_{h_2}(t/T-u)\,\mathrm{tr}\Bigg\{\Bigg[\left(H^{-1}\Sigma_T(u)H^{-1}\right)^{-1}\Big(H^{-1}\mathbb{B}_T(u)\mathbb{B}_T^T(u)H^{-1}+\\
&\qquad+H^{-1}\widetilde{W}_T(u)\widetilde{W}_T^T(u)H^{-1}\Big)\left(H^{-1}\Sigma_T(u)H^{-1}\right)\Bigg]\mathbb{Z}_t\mathbb{Z}_t^T\Bigg\}=\\
&=\frac{1}{T}\mathbb{E}\Bigg\{K_{h_2}(t/T-u)\,\mathrm{tr}\Bigg[\Sigma^{-1}(u)\Big(H^{-1}\mathbb{B}_T(u)\mathbb{B}_T^T(u)H^{-1}+\\
&\qquad+H^{-1}\widetilde{W}_T(u)\widetilde{W}_T^T(u)H^{-1}\Big)\Sigma^{-1}(u)\mathbb{Z}_t\mathbb{Z}_t^T\Bigg]\Bigg\}+o(1).
\end{aligned}$$

Using similar steps as for proving v) we arrive at the following result:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_4(u)-\mathcal{I}_4(u)\right|=O_p\left(\frac{1}{T^2h_1^2}\sqrt{\frac{\log T}{Th_2}}\right).$$

And finally combining results from (i)-(vi) yields:

$$\sup_{u\in I_{h_2}}\left|\widehat{\sigma}^2(u)-\sigma^2(u)\right|=O_p\left(\sqrt{\frac{\log T}{Th_2}}+h_2^2\right), \tag{35}$$

which completes the proof. ■

Proofs of Theorems 3-6.

In what follows we show the proof of Theorem 6 for the extended statistics $\mathcal{S}'_T$ since the basic statistics $\mathcal{S}_T$ (without an extra weighting $\phi(u)$) is nested in $\mathcal{S}'_T$ and can be obtained by setting $\phi(u) = 1 \ \forall \ u \in [0,1]$. Proofs of Theorems 3 and 4 can be obtained as a special case of the proof of Theorem 6, i.e. by setting the function $\triangle$ to zero in the proofs. Proof of Theorem 5 is straightforward provided the proof of Theorems 3 and therefore is omitted. We start by rewriting the test statistics $\mathcal{S}'_T$, normalised by its rate as follows:

$$\sqrt{T}\mathcal{S}'_T=\frac{\sqrt{T}}{\sqrt{\Phi}}\int_0^1\phi(u)\widehat{\tau}(u)du=\sqrt{T}\int_0^1\left[V_T(u)+\mathcal{B}_T(u)\right]du,$$

where $\Phi=\int_0^1\phi^2(u)du$ and with the notation from (26) we can write

$$V_T(u)=\frac{\sqrt{h_1}\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\widetilde{W}_{T,0}(u)}{\sqrt{\Phi}\sigma(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}},$$

and

$$\mathcal{B}_T(u)=\frac{1}{2}\frac{\phi(u)h_1^2X_t^T(u)\Sigma_{T,0}^{-1}(u)\left(\Sigma_{T,2}(u)\ddot{\rho}(u)+\mathbb{R}_{T,0}(u)\right)}{\sqrt{\Phi}\sigma(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}+$$
$$+c_T\triangle(u)+c_T\Big\{\triangle(t/T)-\triangle(u)\Big\},$$

where we used Theorem 2 to substitute $\widehat{\sigma}(u)$ with $\sigma(u)$ and where $\mathbb{R}_{T,0}(u)$ is the residual part, defined in (24)-(25) and

$$\Sigma_{T,m}(u)=\frac{1}{T}\sum_{t=1}^{T}K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^m\mathbb{X}_t\mathbb{X}_t^T,$$

$$\widetilde{W}_{T,0}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T - u\right)\mathbb{X}_t\xi_t,$$

where again recall that for brevity we write $\mathbb{X}_t$ to abbreviate $\mathbb{X}_{t,T}$. Proof of the Theorem 6 follows from the following two lemmas.

Lemma 1. *Under the assumptions (A1)-(A3), it holds that*

$$\sqrt{T}\int_0^1 V_T(u)du \xrightarrow{d} \mathcal{N}\left(0,1\right).$$

Lemma 2. *Under the assumptions (A1)-(A3), it holds that*

$$\sqrt{T}\int_0^1 \mathcal{B}_T(u)du = \int_0^1 \triangle(u)du + h_1^2\sqrt{T}\mathbb{B}_T + o_p(1),$$

where

$$\mathbb{B}_T = \frac{1}{2\sqrt{\Phi}}\int_0^1 \frac{\phi(u)\lambda_2\mathbb{X}_t^T(u)\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}}du.$$

Proof of Lemma 1.

For the ease of exposition we need to introduce some further notation. First, denote by $\delta(u) = \sigma(u)\sqrt{\nu_0\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}$ and denote by $Y_{t,T}$ the following quantity:

$$Y_{t,T} = \frac{\sqrt{h_1}}{\sqrt{T\Phi}}\int_0^1 \delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)K_{h_1}\left(t/T - u\right)\mathbb{X}_t\xi_t du.$$

It is then straightforward to verify that $Y_{t,T}$ is a martingale difference array since conditional on the $X_{t,T}$:

$$\mathbb{E}\left[Y_{t,T}|\mathcal{F}_{t-1,T},\mathbb{X}_{t,T}\right] = \frac{\sqrt{h_1}}{\sqrt{T\Phi}}\int_0^1 \delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)K_{h_1}\left(t/T - u\right)\mathbb{E}\left[\mathbb{X}_t\xi_t|\mathcal{F}_{t-1,T}\right]du + o(1) = 0,$$

where $\mathcal{F}_{t-1,T} := \sigma\left(\mathbb{X}_{s,T},\xi_{s,T} : s \leq t-1\right)$ denotes the sigma-algebra induced by the history of $\mathbb{X}_{t,T}$

and $\xi_{t,T}$. We can therefore apply the central limit theorem for the martingale difference arrays (e.g. Theorem 3.2 in Hall and Heyde (1980)) to establish that $\sum_{t=1}^{T} Y_{t,T}$ is asymptotically normal. For applying Theorem 3.2 in Hall and Heyde (1980), one needs to verify the following conditions:

(C1) $\sum_{t=1}^{T} \mathbb{E}\left[Y_{t,T}^2 | \mathcal{F}_{t-1,T}, \mathbb{X}_{t,T}\right] \xrightarrow{p} V_1$,

(C2) for every $\epsilon > 0$, it holds that $\sum_{t=1}^{T} \mathbb{E}\left[Y_{t,T}^2 \{|Y_{t,T} > \epsilon|\} | \mathcal{F}_{t-1,T}, \mathbb{X}_{t,T}\right] \xrightarrow{p} 0$.

To establish (C1) and (C2) it suffices to verify Lemma B.13 in Kristensen (2012). In particular, conditional on $\{\mathbb{X}_{t,T}\}$, the following two conditions should hold:

(D1) $\sum_{t=1}^{T} \mathbb{E}\left[Y_{t,T}^2\right] \xrightarrow{p} V_1$,

(D2) for some $\delta > 0$, it holds that $\sum_{t=1}^{T} \mathbb{E}\left[|Y_{t,T}|^{2+\delta}\right] \xrightarrow{p} 0$.

Proof of (D1).

We first calculate $Y_{t,T}^2$, which is given by the following expression:

$$Y_{t,T}^2 = \frac{h_1}{T\Phi}\left(\int_0^1 \delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)K_{h_1}\left(t/T-u\right)\mathbb{X}_t\xi_t\right)^2 =$$

$$= \frac{h_1}{T\Phi}\int_0^1 \delta^{-2}(u)\phi^2(u)K_{h_1}^2\left(t/T-u\right)\xi_t^2\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t\mathbb{X}_t^T\Omega^{-1}(u)\mathbb{X}_t(u)du +$$

$$+\frac{h_1}{T\Phi}\int_0^1\int_0^1 \delta^{-2}(u)\phi^2(u)K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\xi_t^2\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t\mathbb{X}_t^T\Omega^{-1}(u')\mathbb{X}_t(u')dudu' + o_p(1).$$

Taking conditional expectation of the above expression

$$\sum_{t=1}^{T}\mathbb{E}\left[Y_{t,T}^2|X_t(u)\right]=$$

$$=\frac{h_1}{T\Phi}\sum_{t=1}^{T}\int_0^1\delta^{-2}(u)\phi^2(u)K_{h_1}^2\left(t/T-u\right)\sigma^2(t/T)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{E}\left[\varepsilon_t^2\mathbb{X}_t\mathbb{X}_t^T\right]\Omega^{-1}(u)\mathbb{X}_t(u)du+$$

$$+\frac{h_1}{T\Phi}\sum_{0}^{1}\int_0^1\int_0^1\delta^{-2}(u)\phi^2(u)K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\sigma^2(t/T)\mathbb{E}\left[\mathbb{X}_t^T(u)\varepsilon_t^2\mathbb{X}_t\mathbb{X}_t^T\Omega^{-1}(u')\mathbb{X}_t(u')\right]dudu'+$$

$$+o(1)=\frac{1}{h_1\Phi}\int_0^1\int_0^1\delta^{-2}(u)\phi^2(u)K^2\left(\frac{y-u}{h_1}\right)\sigma^2(y)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)dudy+o(1)=$$

$$=\frac{1}{\Phi}\int_0^1\delta^{-2}(u)\phi^2(u)\nu_0\sigma^2(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)du+o(1)=\frac{1}{\Phi}\int_0^1\phi^2(u)du+o(1)=1+o(1),$$

where we used the fact that $\delta^{-2}(u)=\sigma^2(u)\nu_0\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t$. We now establish the following intermediate result. We calculate the covariance between $\sqrt{Th_1}\widetilde{W}_{T,0}(u)$ and $\sqrt{Th_1}\widetilde{W}_{T,0}(u')$ for generic rescaled time points $u,u'\in[0,1]$. Without loss of generality, assume that $u'<u$ for $(u-h_1)T\leq t\leq(u+h_1)T$ and $(u'-h_1)T\leq t\leq(u'+h_1)T$, then it holds that

$$\mathbb{E}\left(\frac{h_1}{T}\sum_{t=1}^{T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=1}^{T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right)=$$

$$=\mathbb{E}\left(\frac{h_1}{T}\sum_{t=(u-h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u'-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right)=$$

$$=\mathbb{E}\left\{\frac{h_1}{T}\left[\sum_{t=(u-h)T}^{(u'+h)T-1}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t+\sum_{t=(u'+h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\right]\times\right.$$

$$\left.\times\left[\sum_{t=(u'-h)T}^{(u-h)T-1}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right]\right\}=$$

$$=\mathbb{E}\left\{\frac{h_1}{T}\sum_{t=(u-h)T}^{(u'+h)T-1}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u'-h)T}^{(u-h)T-1}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+\right.$$

$$+\frac{h_1}{T}\sum_{t=(u-h)T}^{(u'+h)T-1}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+$$

$$+\frac{h_1}{T}\sum_{t=(u'+h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u'-h)T}^{(u-h)T-1}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+$$

$$\left.+\frac{h_1}{T}\sum_{t=(u'+h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right\}=$$

$$=\frac{h_1}{T}\mathbb{E}\left(\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right)=$$

$$\frac{h_1}{T}\mathbb{E}\left\{\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\mathbb{X}_t(u)\mathbb{X}_t^T(u')\sigma^2(t/T)\varepsilon_t^2+\right.$$

$$\left.+\underbrace{\sum_{t=(u-h)T}^{(u'+h)T}\sum_{s=(u-h)T}^{(u'+h)T}}_{t\neq s}K_{h_1}\left(t/T-u\right)K_{h_1}\left(s/T-u'\right)\mathbb{X}_t(u)\mathbb{X}_s^T(u')\sigma(t/T)\sigma(s/T)\varepsilon_t\varepsilon_s\right\}=$$

$$=h_1\mathbb{E}\left(K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\mathbb{X}_t(u)\mathbb{X}_t^T(u')\sigma^2(t/T)\varepsilon_t^2\right)=$$

$$=\frac{h_1}{T}\sum_{t=1}^{T}K_{h_1}(t/T-u)K_{h_1}(t/T-u')\Omega(t/T)\sigma^2(t/T)+o(1)=$$

$$=\int_0^1K\left(y\right)K\left(y+\frac{u-u'}{h_1}\right)y\sigma^2(yh_1+u)\Omega(yh_1+u)dy=\nu_0\sigma^2(u)\Omega(u)+o(1).\qquad\blacksquare$$

Proof of (D2).

Provided that $\mathbb{X}_t = \mathbb{X}(u) + O_p\left(|t/T - u| + 1/T\right)$ and $\forall\ t$ we have that $\mathbb{E}|\varepsilon_t|^{2+\delta} < \infty$, then for a large enough $C < \infty$, and conditional on $\mathbb{X}_t(u)$, it holds:

$$\sum_{t=1}^{T} \mathbb{E}\left[|Y_{t,T}|^{2+\delta}\right] = \frac{h_1^{1+\delta/2}}{(T\Phi)^{1+\delta/2}} \sum_{t=1}^{T} \mathbb{E}\left[\int_0^1 \left|\delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)K_{h_1}\left(t/T-u\right)\mathbb{X}_t\xi_t\right|^{2+\delta} du\right] + o(1)$$

$$\leq \frac{h_1^{1+\delta/2}}{(T\Phi)^{1+\delta/2}} \mathbb{E}\left[\int_0^1 \sum_{t=1}^{T} K_{h_1}^{2+\delta}(t/T-u)\sigma^{2+\delta}(t/T)\left|\delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)\varepsilon_t\right|^{2+\delta} du\right] + o(1) =$$

$$\leq \frac{h_1^{1+\delta/2}}{(T\Phi)^{1+\delta/2}} \left[\int_0^1 \sum_{t=1}^{T} K_{h_1}^{2+\delta}(t/T-u)\sigma^{2+\delta}(t/T)\left|\delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)\right|^{2+\delta} du\right] \mathbb{E}\left[|\varepsilon_t|^{2+\delta}\right]$$

$$\leq \frac{C}{(Th_1)^{\delta/2}\Phi^{1+\delta/2}} \left[\int_0^1 \left|\frac{\delta(u)\phi(u)}{\sigma(u)\nu_0}\right|^{2+\delta} du\right] \int_0^1 K^{2+\delta}(z)dz = o(1).$$

Combining all of the above derivations proves Lemma 1. ■

Proof of Lemma 2. First recall from (27) that

$$h_1^{-m}\Sigma_{T,m} = \lambda_m\Omega(u)\{1 + o_p(1)\} \qquad \text{and} \qquad h_1^{-m}\mathbb{R}_{T,m} = o_p(h_1^2).$$

With the notation introduced in the proof of Lemma 1, we can write

$$\mathcal{B}_T(u) = \frac{1}{2\sqrt{\Phi}} h_1^2\phi(u)\delta^{-1}(u)X_t^T(u)\ddot{\rho}(u)\frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^2\mathbb{X}_t\mathbb{X}_t^T + c_T \triangle\left(u\right) + o_p(h_1^2).$$

Setting $c_T = 1/\sqrt{T}$, normalising $\mathcal{B}_T(u)$ by $\sqrt{T}$ and taking expectation yields:

$$\mathbb{E}\left[\sqrt{T}\int_0^1 \mathcal{B}_T(u)du\right] = \frac{\sqrt{T}h_1^2}{2\sqrt{\Phi}}\int_0^1 \frac{\phi(u)X_t^T(u)\lambda_2\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}du + \int_0^1 \triangle(u)du + o(1),$$

which concludes the proof of Lemma 2. ■

Proof of Theorem 7.

We start by deriving $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ using our model (12):

$$Pr\left(\Delta\mathcal{L}_{T+1} \leq 0|\mathcal{F}_T\right) = Pr\left(\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right) + \sigma\left(\frac{T+1}{T}\right)\varepsilon_{T+1} \leq 0|\mathcal{F}_T\right) = \\ = Pr\left(\varepsilon_{T+1} \leq \frac{-\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\right).$$

Recall that we stipulated that $\rho(t/T) = \rho(1)$ for any $t \geq T$ and similarly $\sigma(t/T) = \sigma(1)$ for any $t \geq T$. We therefore are looking to estimate

$$Pr\left(\varepsilon_{T+1} \leq \frac{-\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\right) = Pr\left(\varepsilon_{T+1} \leq \frac{-\mathbb{X}_{T+1}^T\rho\left(1\right)}{\sigma\left(1\right)}\right) =: F\left(\varepsilon^\star(1)\right),$$

where $\varepsilon^\star(1) := \frac{-\mathbb{X}_{T+1}^T\rho(1)}{\sigma(1)}$. Now, conditional on the sample $\{\Delta\mathcal{L}_t\}_{t=1}^T$:

$$\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = \widehat{F}\left(\widehat{\varepsilon}^\star(1)\right) = \frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\widehat{\varepsilon}_t \leq \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}\left(1\right)}{\widehat{\sigma}\left(1\right)}\right), \tag{36}$$

where $\widehat{\varepsilon}^\star(1) := \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}(1)}{\widehat{\sigma}(1)}$. Taking expectation of the above and using a first-order Taylor expansion it holds that:

$$\mathbb{E}\left[\widehat{F}\left(\widehat{\varepsilon}^\star(1)\right)\right] = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\widehat{\varepsilon}_t \leq \widehat{\varepsilon}^\star(1)\middle|\widehat{\varepsilon}^\star\right)\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\widehat{\varepsilon}_t \leq \widehat{\varepsilon}^\star(1)\middle|\widehat{\varepsilon}^\star\right)\right]\right] = \\ = \mathbb{E}\left[F\left(\varepsilon^\star(1)\right) + F'\left(\varepsilon^\star(1)\right)\left(\widehat{\varepsilon}^\star(1) - \varepsilon^\star(1)\right) + o(T^{-1/2})\right] = \\ = F(\varepsilon^\star(1)) + f(\varepsilon^\star(1))\mathbb{E}\left\{\left(\widehat{\varepsilon}^\star(1) - \varepsilon^\star(1)\right)\right\} + o(T^{-1/2}).$$

Using a first-order Taylor expansion of $\widehat{\varepsilon}^\star$ around $\varepsilon^\star$, it holds that:

$$\widehat{\varepsilon}^\star(1) = \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}\left(1\right)}{\widehat{\sigma}(1)} = \frac{-\mathbb{X}_{T+1}^T\rho\left(1\right)}{\sigma(1)} - \frac{\mathbb{X}_{T+1}^T\left[\widehat{\rho}(1) - \rho\left(1\right)\right]}{\widehat{\sigma}(1)} + \mathbb{X}_{T+1}^T\frac{\rho\left(1\right)\left[\widehat{\sigma}(1) - \sigma(1)\right]}{\widehat{\sigma}^2(1)} = \\ = \varepsilon^\star(1) - \frac{\mathbb{X}_{T+1}^T\left[\widehat{\rho}(1) - \rho\left(1\right)\right]}{\sigma(1)} + \frac{\mathbb{X}_{T+1}^T\rho\left(1\right)\left[\widehat{\sigma}(1) - \sigma(1)\right]}{\sigma^2(1)},$$

and therefore

$$\mathbb{E}\left[\widehat{\varepsilon}^{\star}(1)-\varepsilon(1)\right]=\frac{1}{2\sigma^{2}(1)}\mathbb{X}_{T+1}(1)\left\{h_{1}^{2}\lambda_{2}\ddot{\rho}(1)\sigma(1)+h_{2}^{2}\lambda_{2}\ddot{\sigma}(1)\right\}=:\mathbb{B}_{3}(1).$$

Given that $\mathbb{1}\left(\widehat{\varepsilon}_{t}\leq\frac{-\mathbb{X}_{T+1}^{T}\widehat{\rho}(1)}{\widehat{\sigma}(1)}\right)$ is a Bernulli random variable and therefore

$$\widehat{F}\left(\widehat{\varepsilon}^{\star}(1)\right)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\left(\widehat{\varepsilon}_{t}\leq\frac{-\mathbb{X}_{T+1}^{T}\widehat{\rho}\left(1\right)}{\widehat{\sigma}\left(1\right)}\right)$$

has a binomial distribution which as $T\rightarrow\infty$ becomes a normal distribution with the mean and variance given by

$$\begin{aligned}\text{var}\left(\sqrt{T}\widehat{F}\left(\widehat{\varepsilon}^{\star}(1)\right)\right)=\mathbb{E}\left[T\widehat{F}\left(\widehat{\varepsilon}^{\star}(1)\right)^{2}\right]-\left(T\mathbb{E}\left[\widehat{F}\left(\widehat{\varepsilon}^{\star}(1)\right)\right]\right)^{2}=\\ =F(\varepsilon^{\star}(1))\left(1-F\left(\varepsilon^{\star}(1)\right)\right)+o(1).\end{aligned}$$

Therefore, it holds that

$$\sqrt{T}\left[\widehat{F}(\widehat{\varepsilon}^{\star}(1))-F(\varepsilon^{\star}(1))-\mathbb{B}_{3}(1)\right]\xrightarrow{d}\mathcal{N}\left(0,F\left(\varepsilon^{\star}(1)\right)\left(1-F\left(\varepsilon^{\star}(1)\right)\right)\right),\tag{37}$$

where

$$\mathbb{B}_{3}(1)=\frac{f\left(\varepsilon^{\star}(1)\right)}{2\sigma^{2}(1)}\mathbb{X}_{T+1}(1)\left\{h_{1}^{2}\lambda_{2}\ddot{\rho}(1)\sigma(1)+h_{2}^{2}\lambda_{2}\ddot{\sigma}(1)\right\},$$

which completes the proof. ■

Proof of Theorem 8.

The proof of the Theorem 8 closely follows original proof of Theorems 3-6 with the bootstrapped quantities, denoted by $\star$. In particular, $\mathbb{E}^{\star}(\cdot)$, $\text{var}^{\star}(\cdot)$ and $\mathbb{P}^{\star}(\cdot):=\mathbb{P}^{\star}\left(\cdot|\{\Delta\mathcal{L}_{t,T},\mathbb{X}_{t,T}\}_{t=1}^{T}\right)$ are used to denote the expectation, variance and the distribution respectively conditional on the sample $\{\Delta\mathcal{L}_{t,T},\mathbb{X}_{t,T}\}_{t=1}^{T}$. We start by making use of the following notation:

$$\sqrt{T}\mathcal{S}_{T}^{\star}=\frac{\sqrt{T}}{\sqrt{\Phi}}\int_{0}^{1}\phi(u)\widehat{\tau}^{\star}(u)du=\sqrt{T}\int_{0}^{1}\left[V_{T}^{\star}(u)+\mathcal{B}_{T}^{\star}(u)\right]du,$$

where $\Phi = \int_0^1 \phi^2(u)du$ and with the notation from (26) we can write

$$V_T^\star(u) = \frac{\sqrt{Th_1}\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\widetilde{W}_{T,0}^\star(u)}{\sqrt{\Phi}\widehat{\sigma}^\star(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}, \tag{38}$$

and

$$\mathcal{B}_T^\star(u) = \frac{\phi(u)}{\sqrt{\Phi}}\tau^\star(u) + \frac{\sqrt{Th_1}\phi(u)h_1^2\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\left(\widetilde{\mathcal{R}}_T^\star(u) + o_p(h_1^2)\right)}{\sqrt{\Phi}\widehat{\sigma}^\star(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}, \tag{39}$$

where we used Theorem 2 to substitute $\widehat{\sigma}^\star(u)$ with $\sigma^\star(u)$ and where the following notation is used:

$$\Sigma_{T,m}(u) = \frac{1}{T}\sum_{t=1}^T K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^m\mathbb{X}_t(u)\mathbb{X}_t^T(u),$$

$$\widetilde{W}_{T,0}^\star(u) = \frac{1}{T}\sum_{t=1}^T K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\xi_t^\star,$$

and

$$\widetilde{\mathcal{R}}_T^\star(u) = \frac{1}{T}\sum_{t=1}^T K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\mathbb{X}_t^T\Big\{\rho(t/T) - \widetilde{\rho}(u)\Big\}.$$

In what follows we need to show that under the conditions of Theorem 8, the following holds:

$$\sqrt{T}\int_0^1 V_T^\star(u)du \xrightarrow{d} \mathcal{N}\left(0,1\right), \tag{B1}$$

conditional on the sample $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}$ with probability tending to one, and

$$\sqrt{T}\int_0^1 \mathcal{B}_T^\star(u)du = h_1^2\sqrt{T}\mathbb{B}_T + o_p(1). \tag{B2}$$

For the proofs of (B1) and (B2) we will be using the following notation from the proof of Lemma 1: denote by $\delta^\star(u) = \sigma^\star(u)\sqrt{\nu_0\mathbb{X}_t(u)\Omega^{-1}(u)\mathbb{X}_t(u)}$ and

$$\Upsilon_{t,T}^\star = \frac{\sqrt{h_1}}{\sqrt{T\Phi}}\int_0^1 \delta^{-1,\star}(u)\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)K_{h_1}\left(t/T-u\right)\mathbb{X}_t\xi_t^\star du, \tag{40}$$

where $\xi_t^\star = \widehat{\xi}_t\eta_t$ are the bootstrapped residuals. Note that since η_t are i.i.d. it then follows that $\xi_t^\star$ have the same mixing properties as the original residuals ξ_t (see Theorem 5.2 in Bradley (2005)).

It is than straightforward to establish that $Y^{\star}_{t,T}$ is also a martingale difference sequence and by using uniform convergence results in Theorem 2, in what follows I establish that conditional on the sample with probability one it holds that

$$\sqrt{T}\int_0^1 \mathcal{B}^{\star}_T(u)du = \sqrt{T}\int_0^1 \mathcal{B}_T(u)du + o_p(1), \tag{B3}$$

and

$$\mathbb{P}^{\star}\left(\sqrt{T}\int_0^1 V^{\star}_T(u)du \leq x\right) \xrightarrow{p} \Phi(x), \tag{B4}$$

where $\Phi(x)$ is the standard Gaussian distribution. Therefore,

$$\mathbb{P}^{\star}\left(\sqrt{T}\left(S^{\star}_T - \mathbb{B}_T\right) \leq x\right) \xrightarrow{p} \Phi(x),$$

which then completes the proof of Theorem 8. ■

Below we prove (B3) and (B4). However, before proving (B3) and (B4), we first consider $\widehat{\sigma}^{\star}(u)$. We can write

$$\widehat{\sigma}^{\star}(u) = \frac{\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\left(\widehat{\xi}^{\star}_t\right)^2}{\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)} = \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\left(\widehat{\xi}^{\star}_t\right)^2 =$$

$$= \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\left(\Delta\mathcal{L}^{\star}_t - \mathbb{X}^T_t\widehat{\rho}^{\star}(t/T)\right)^2 =$$

$$= \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\left(\mathbb{X}^T_t\widetilde{\rho}(t/T) + \xi^{\star}_t - \mathbb{X}^T_t\widehat{\rho}^{\star}(t/T)\right)^2 = \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\left(\xi^{\star}_t\right)^2 +$$

$$+ \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\mathbb{X}^T_t\left(\widetilde{\rho}(t/T) - \widehat{\rho}^{\star}(t/T)\right)\left(\widetilde{\rho}(t/T) - \widehat{\rho}^{\star}(t/T)\right)\mathbb{X}_t +$$

$$+ 2\widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi^{\star}_t\left(\widetilde{\rho}(t/T) - \widehat{\rho}^{\star}(t/T)\right).$$

Using results (i)-(vi) in the proof of Theorem 2, and the definition of $\widehat{\rho}^{\star}(t/T)$, it is then straightforward to establish that

$$\sup_{u\in I_{h_2}}\left|\widehat{\sigma}^{\star}(u) - \sigma^{\star}(u)\right| = O_p\left(\sqrt{\frac{\log T}{Th_2}} + h_2^2\right), \tag{41}$$

where I_{h_2} is defined in the statement of Theorem 2. We now prove (B3) and (B4). Using the result

in (41) we substitute $\widehat{\sigma}^{\star}(u)$ with $\sigma^{\star}(u)$ in the expressions of (38), (39) and (40).

Proof of B3.

It then holds that

$$\sqrt{T}\int_0^1 \mathcal{B}_T^{\star}(u)du = \frac{\sqrt{T}}{\sqrt{\Phi}}\int_0^1 \phi(u)\tau^{\star}(u)du + \sqrt{T}\int_0^1 \frac{\sqrt{Th_1}\phi(u)h_1^2\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\left(\widetilde{\mathcal{R}}_T^{\star}(u)+o_p(h_1^2)\right)}{\sqrt{\Phi}\widehat{\sigma}^{\star}(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}du =$$

$$= \frac{\sqrt{T}}{2}\int_0^1 \frac{\sqrt{Th_1}\phi(u)h_1^2\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\left(\Sigma_{T,2}(u)\ddot{\rho}(u)+o_p(h_1^2)\right)}{\sqrt{\Phi}\sigma^{\star}(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}du =$$

$$= \sqrt{T}\int_0^1 \mathbb{B}_T(u)du + o_p(1) \quad , \quad (42)$$

where in the second line of (42) we used the fact that $\int_0^1 \phi(u)\tau^{\star}(u)du = 0$ by construction in (17)-(18). This completes the proof of (B3). ■

Proof of B4.

Denoting by $\zeta_t := \xi_t\eta_t$, it holds that

$$\xi_t^{\star} := \widehat{\xi}_t\eta_t = (\Delta\mathcal{L}_t - \widehat{\mu}_t)\,\eta_t = (\mu_t + \xi_t - \widehat{\mu}_t)\,\eta_t =$$
$$= \xi_t\eta_t + \mathbb{X}_t^T\left(\widehat{\rho}(t/T) - \rho(t/T))\eta_t = \zeta_t + \mathbb{X}_t^T\left(\widehat{\rho}(t/T) - \rho(t/T)\right)\eta_t.\right.$$

First recall that $\widehat{\rho}(u) - \rho(u) = O_p\left(\sqrt{\frac{\log T}{Th_1}} + h_1^2\right)$, and the distribution of $Y_{t,T}^{\star}$ in (40) will be given by the expression involving ζ_t, i.e.

$$\widetilde{Y}_{t,T}^{\star} = \frac{\sqrt{h_1}\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\int_0^1 K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\zeta_t du}{\sqrt{\Phi}\sigma^{\star}(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}} =$$

$$= \frac{\sqrt{h_1}\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\int_0^1 K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\xi_t^{\star} du}{\sqrt{\Phi}\sigma^{\star}(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}},$$

where $\zeta_t := \xi_t \eta_t$. Note also, that provided the definition of η_t, the mixing properties of the original residual sequence ξ_t (Theorem 5.2 of Bradley (2005)) are preserved by the new residual ζ_t and therefore to establish (B4) we need to verify conditions (C1) and (C2) in the proof of Theorems 3-6. Since the proof follows exactly the same steps as the one in the proof of Theorems 3-6, it is omitted. ■

11 Appendix C.

This Appendix presents more results from the application section. In particular, we report results for two more forecast horizons: $k = 6$ months and $k = 12$ months.

Table 16: Results for one-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 6$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.49					−1.93					−2.75					−2.25				
	p	0.930					0.980					0.996					0.988				
DI	$\mathcal{S}_T$	1.09	1.54				−0.43	1.82				−3.05	2.78				−3.91	2.39			
	p	0.154	0.068				0.692	**0.040**				0.998	**0.006**				1.00	**0.014**			
AR	$\mathcal{S}_T$	1.42	1.60	1.20			0.82	1.94	1.59			−2.66	2.79	2.71			−4.016	1.82	1.48		
	p	0.084	**0.046**	0.220			0.192	**0.028**	0.056			0.994	**0.008**	**0.008**			1.00	**0.046**	0.062		
RW	$\mathcal{S}_T$	1.46	1.53	0.02	−0.78	–	0.15	1.81	0.82	−0.50	–	3.84	2.55	3.32	6.83	–	2.19	2.31	4.77	7.11	–
	p	0.074	0.062	0.968	0.786	–	0.444	**0.034**	0.200	0.688	–	**0.000**	**0.010**	**0.000**	**0.000**	–	**0.012**	**0.012**	**0.000**	**0.000**	–

Note: Table reports the value of the one-sided test statistic $\mathcal{S}_T$ that corresponds to the null of superior predictive ability, see eq. (2), for horizon $k = 6$ months. The p-values are obtained via the wild 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.49$ (indicating that Lasso is better) with the p-value of 0.930. The p-values in bold indicate rejection of the null (2) at the 5% level of significance.

Table 17: Results for two-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 6$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.50					−1.87					−2.77					−2.16				
	p	0.140					0.072					**0.016**					**0.040**				
DI	$\mathcal{S}_T$	1.12	1.56				−0.43	1.82				−3.34	2.88				−3.87	2.21			
	p	0.252	0.120				0.640	0.080				**0.004**	**0.004**				**0.000**	**0.030**			
AR	$\mathcal{S}_T$	1.44	1.48	1.21			0.81	1.81	1.60			−2.95	2.92	2.68			−3.86	2.11	1.47		
	p	0.136	0.148	0.240			0.445	0.080	0.108			**0.000**	**0.008**	**0.008**			**0.000**	**0.040**	0.148		
RW	$\mathcal{S}_T$	1.38	1.53	0.02	−0.80	–	0.15	1.82	0.84	−0.51	–	3.83	2.89	3.29	7.11	–	2.21	2.30	4.79	7.12	–
	p	0.188	0.156	0.948	0.456	–	0.780	0.064	0.396	0.640	–	**0.000**	**0.008**	**0.004**	**0.000**	–	**0.036**	**0.032**	**0.000**	**0.000**	–

Note: Table reports the value of the two-sided test statistic $\mathcal{S}_T$ that corresponds the null of equal predictive ability, see eq. (1), for horizon $k = 6$ months. The p-values are obtained via the wild 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.50$ (indicating that Lasso is better) with the p-value of 0.140. The p-values in bold indicate rejection of the null (1) at the 5% level of significance.

Table 18: Results for one-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 12$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.78					−2.06					−1.78					−2.07				
	p	0.960					0.972					0.962					0.980				
DI	$\mathcal{S}_T$	0.90	1.79				−0.66	2.14				−3.42	1.77				−2.80	1.75			
	p	0.192	**0.036**				0.766	**0.028**				0.996	**0.046**				0.998	**0.046**			
AR	$\mathcal{S}_T$	1.82	1.79	0.97			0.79	2.08	1.72			−3.21	1.68	3.52			−4.55	1.93	3.58		
	p	**0.040**	**0.030**	0.178			0.206	**0.020**	**0.034**			1.00	0.060	**0.000**			1.00	**0.034**	**0.000**		
RW	$\mathcal{S}_T$	1.42	1.70	0.34	−0.87	–	0.36	2.14	1.34	−0.34	–	1.51	1.74	6.30	7.21	–	1.36	1.95	3.92	6.15	–
	p	0.076	**0.036**	0.398	0.814	–	0.370	**0.018**	0.092	0.624	–	0.074	**0.048**	**0.000**	**0.000**	–	0.092	**0.032**	**0.000**	**0.000**	–

Note: Table reports the value of the one-sided test statistic $\mathcal{S}_T$ that corresponds to the null of superior predictive ability, see eq. (2), for horizon $k = 12$ months. The p-values are obtained via the wild 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.78$ (indicating that Lasso is better) with the p-value of 0.960. The p-values in bold indicate rejection of the null (2) at the 5% level of significance.

Table 19: Results for two-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 12$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.86					−2.06					−1.75					−1.95				
	p	0.052					**0.036**					0.072					0.060				
DI	$\mathcal{S}_T$	0.89	1.81				−0.69	2.06				−3.53	1.47				−2.80	1.68			
	p	0.380	0.072				0.476	0.052				**0.000**	0.144				**0.008**	0.100			
AR	$\mathcal{S}_T$	1.73	1.76	0.99			0.82	1.89	1.61			−3.29	1.66	3.47			−4.72	1.96	3.59		
	p	0.088	0.092	0.312			0.396	0.084	0.100			**0.000**	0.112	**0.000**			**0.000**	**0.044**	**0.000**		
RW	$\mathcal{S}_T$	1.30	1.80	0.32	−0.87	−	0.36	2.19	1.32	−0.34	−	1.47	1.55	6.42	7.36	−	1.42	1.95	3.74	5.97	−
	p	0.216	0.068	0.772	0.352	−	0.736	**0.032**	0.172	0.768	−	0.144	0.132	**0.000**	**0.000**	−	0.156	**0.050**	**0.000**	**0.000**	−

Note: Table reports the value of the two-sided test statistic $\mathcal{S}_T$ that corresponds the null of superior predictive ability, see eq. (1), for horizon $k = 12$ months. The p-values are obtained via the wild 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.86$ (indicating that Lasso is better) with the p-value of 0.052. The p-values in bold indicate rejection of the null (1) at the 5% level of significance.

Table 20: Sign forecasting for $\Delta\mathcal{L}_{T+1}$ for horizon $k = 6$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\widehat{Pr}_{T+1}$	0.986					0.986					0.999					0.991				
	$\widehat{FI}_l$	0.983					0.979					0.998					0.990				
	$\widehat{FI}_u$	0.988					0.989					1.000					0.992				
	$\widehat{C}$	0.087					0.049					0.028					0.061				
DI	$\widehat{Pr}_{T+1}$	0.542	0.005				0.896	0.031				0.869	0.000				0.852	0.004			
	$\widehat{FI}_l$	0.500	0.003				0.864	0.027				0.831	0.000				0.819	0.003			
	$\widehat{FI}_u$	0.581	0.006				0.906	0.042				0.901	0.001				0.885	0.005			
	$\widehat{C}$	-0.024	-0.082				0.022	-0.072				0.006	-0.028				0.011	-0.050			
AR	$\widehat{Pr}_{T+1}$	0.444	0.005	0.382			0.801	0.017	0.136			0.746	0.001	0.246			0.577	0.000	0.278		
	$\widehat{FI}_l$	0.408	0.003	0.352			0.751	0.013	0.126			0.697	0.000	0.201			0.539	0.000	0.245		
	$\widehat{FI}_u$	0.481	0.007	0.415			0.836	0.025	0.176			0.769	0.002	0.272			0.601	0.001	0.307		
	$\widehat{C}$	-0.040	-0.092	-0.015			-0.012	-0.075	-0.025			0.049	-0.034	0.025			0.039	-0.060	0.016		
RW	$\widehat{Pr}_{T+1}$	0.456	0.008	0.202	0.571		0.935	0.030	0.639	0.883		0.151	0.000	0.088	0.248		0.158	0.000	0.102	0.338	
	$\widehat{FI}_l$	0.424	0.006	0.178	0.537	–	0.916	0.024	0.614	0.859	–	0.127	0.000	0.059	0.231	–	0.147	0.000	0.074	0.329	–
	$\widehat{FI}_u$	0.491	0.011	0.254	0.600	–	0.945	0.036	0.684	0.894	–	0.202	0.001	0.125	0.279	–	0.191	0.001	0.143	0.376	–
	$\widehat{C}$	-0.034	-0.104	-0.026	0.022		-0.031	-0.068	-0.035	-0.001		-0.028	-0.027	-0.012	-0.057		-0.013	-0.052	-0.001	-0.054	

Note: Table reports the results of the sign forecasting for $\Delta\mathcal{L}_{T+1}$ for the forecast horizon $k = 6$ months. $\widehat{Pr}_{T+1}$ is an abbreviation of $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, i.e. the forecasted probability at the very end of the sample. $\widehat{FI}_u$ and $\widehat{FI}_l$ denotes the upper and lower bounds of the forecast interval, such that $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) \in [\widehat{FI}_l,\ \widehat{FI}_u]$. Finally, $\widehat{C}$ denotes the value of the criterion in eq.(15). 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$, for which $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = 0.986$ with the corresponding forecast interval $[0.983,\ 0.988]$.

Table 21: Sign forecasting for $\Delta\mathcal{L}_{T+1}$ for horizon $k = 12$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\widehat{Pr}_{T+1}$	0.415					1.000					0.706					0.821				
	$\widehat{FI}_l$	0.387					0.999					0.683					0.795				
	$\widehat{FI}_u$	0.457					1.000					0.724					0.839				
	$\widehat{C}$	0.072					0.049					0.050					0.102				
DI	$\widehat{Pr}_{T+1}$	0.485	0.540				0.686	0.000				0.823	0.284				0.518	0.197			
	$\widehat{FI}_l$	0.453	0.504				0.659	0.000				0.789	0.262				0.487	0.179			
	$\widehat{FI}_u$	0.526	0.573				0.702	0.001				0.862	0.312				0.561	0.228			
	$\widehat{C}$	-0.017	-0.064				-0.035	-0.054				0.016	-0.073				0.023	-0.101			
AR	$\widehat{Pr}_{T+1}$	0.524	0.560	0.602			0.644	0.000	0.244			0.828	0.326	0.700			0.799	0.197	0.796		
	$\widehat{FI}_l$	0.491	0.523	0.579			0.599	0.000	0.216			0.779	0.305	0.654			0.760	0.176	0.751		
	$\widehat{FI}_u$	0.552	0.588	0.620			0.671	0.001	0.273			0.854	0.352	0.732			0.826	0.230	0.828		
	$\widehat{C}$	-0.018	-0.068	-0.029			-0.011	-0.069	-0.030			0.049	-0.004	0.002			0.031	-0.102	-0.003		
RW	$\widehat{Pr}_{T+1}$	0.491	0.516	0.603	0.372		0.752	0.000	0.734	0.880		0.222	0.276	0.191	0.137		0.234	0.174	0.568	0.110	
	$\widehat{FI}_l$	0.460	0.476	0.573	0.348	–	0.724	0.000	0.702	0.833	–	0.202	0.267	0.154	0.111	–	0.214	0.158	0.532	0.088	–
	$\widehat{FI}_u$	0.520	0.548	0.619	0.394	–	0.770	0.001	0.749	0.897	–	0.268	0.300	0.226	0.193	–	0.276	0.203	0.610	0.157	–
	$\widehat{C}$	-0.026	-0.065	-0.018	0.033		-0.029	-0.072	-0.041	-0.020		-0.017	-0.046	-0.038	-0.057		0.002	-0.097	-0.018	-0.044	

Note: Table reports the results of the sign forecasting for $\Delta\mathcal{L}_{T+1}$ for the forecast horizon $k = 12$ months. $\widehat{Pr}_{T+1}$ is an abbreviation of $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, i.e. the forecasted probability at the very end of the sample. $\widehat{FI}_u$ and $\widehat{FI}_l$ denotes the upper and lower bounds of the forecast interval, such that $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) \in [\widehat{FI}_l,\ \widehat{FI}_u]$. Finally, $\widehat{C}$ denotes the value of the criterion in eq.(15). 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, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$, for which $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = 0.415$ with the corresponding forecast interval $[0.387,\ 0.457]$.