Manuscript submitted to *The Econometrics Journal*, pp. 1–26.

Comparing predictive ability in presence of instability over a very short time

Fabrizio Iacone[†,‡], Luca Rossini[†,§] and Andrea Viselli[†]

[†] *Università degli Studi di Milano*
E-mail: fabrizio.iacone@unimi.it, luca.rossini@unimi.it, andrea.viselli@unimi.it
[‡] *University of York*
[§] *Fondazione Eni Enrico Mattei (FEEM)*

Summary We consider forecast comparison in the presence of instability when this affects only a short period of time. We demonstrate that global tests do not perform well in this case, as they were not designed to capture very short-lived instabilities, and their power vanishes altogether when the magnitude of the shock is very large. We then propose and discuss approaches that are more suitable to detect such situations, such as nonparametric methods like the S test from Andrews (2003) or the MAX procedure from Harvey et al. (2021). We illustrate these results in a Monte Carlo exercise and in a comparison of the nowcast of the quarterly US nominal GDP from the Survey of Professional Forecasters (SPF) against a naive benchmark of no growth, over a period that includes the GDP instability brought by the COVID-19 crisis. We recommend that the forecaster does not pool the sample, but excludes the short periods of high local instability from the evaluation exercise.

Keywords: *Forecast Evaluation, Local Diagnostics, Structural Instability Test, Change Point, SPF.*

1. INTRODUCTION

Instabilities during periods of crisis are common in time series. Forecasting becomes more challenging during the phases of crisis and sudden recovery, and yet these periods are often more important in the forecasting task, since they usually carry a greater risk of catastrophic errors. An exceptional example is the methodology proposed by Bok et al. (2018), which was implemented by the New York FED but was suspended in September 2021 due to the challenges posed by the COVID-19 pandemic and was recently reintroduced in Almuzara et al. (2023). The period related to COVID-19 and the sudden recovery (see, e.g. Ng, 2021; Lenza and Primiceri, 2022) has attracted attention among researchers and policymakers since for the first time we have had the opportunity to study an extreme, unpredictable situation.

COVID-19 caused a shock to GDP that was in many ways unprecedented in recent history and tested the capacity of current methodologies in the presence of extraordinary situations. The main challenges are in how the forecasting methods perform and in how we should evaluate the forecasts that common models provide to policymakers. The first challenge has received widespread attention in the literature: for example, Huber et al. (2023) argue that nonlinear methods may better accommodate extreme situations, while Foroni et al. (2022) and Schorfheide and Song (2024) stress the resilience of popular models, such as mixed frequency or dynamic factor models, to accommodate the severity of recessions.

On the other hand, the second task has not received the same amount of attention, despite the fact that average tests for forecast evaluation, such as the Diebold and Mariano

test of equal unconditional predictive ability (see, Diebold and Mariano, 1995; Giacomini and White, 2006) may not be informative, as they do not have much power to detect instances in which one forecast outperforms the competitor only on a fraction of the sample. To account for forecasting instability, Rossi (2021) recommends the application of local procedures, like the one-time reversal or the fluctuation tests from Giacomini and Rossi (2010). In comparison with diagnostics based on the evaluation of forecasts over an average, the fluctuation test is indeed local, as the statistic is only computed over a fraction of the out-of-sample evaluation period.

However, as the test is derived assuming that the fluctuation induced by the instability spans a relevant fraction of the sample, then it replicates on a smaller scale the difficulties incurred by the average, global Diebold and Mariano (1995) test. Indeed, the Monte Carlo study in Giacomini and Rossi (2010) shows that the attempt to run the test for a very short fluctuation period is frustrated by a relevant size distortion.

In this paper, therefore, we discuss the difficulties associated with using global diagnostics for evaluation in times of crisis, and we consider other diagnostics, based on the predictive instability tests: the S test from Andrews (2003) and the MAX diagnostic from Harvey et al. (2021). Situations of crises like the one induced by COVID-19 or by an economic recession do not typically conform well with the assumption of a large, even if local, evaluation period, as crises often span only a very small number of observations. As an example, in the COVID-19-induced recession, the instability mostly affected just two or three quarters.

One key finding of our paper is that in the presence of high, very localised instability, global tests may have no power at all, thus leading to the incorrect conclusion, and that this may also obfuscate a signal that would otherwise be clear, the analysis had not included that brief period of instability. In particular, we investigate in a Monte Carlo exercise the importance of considering the S test and the MAX procedure instead of the usual global tests when we have a short deviation. Finally, we emphasize the importance of applying a correction for the dependence in the S test statistic by using the estimated value in the pre-changed sample rather thab the one recommended in Andrews (2003) or a simple identity matrix.

In an empirical application to the US nominal GDP growth, we compare the nowcast from the Survey of Professional Forecasters (SPF) and the last available observation during the COVID-19 recession and the subsequent recovery. When COVID-19 is not considered in the analysis, Diebold and Mariano and fluctuation tests suggest that the SPF nowcast is more precise than the naive model, but this does not happen when we include the COVID-19 period. On the other hand, the S test and the MAX procedure provide more appropriate results. Given these results, we recommend that the forecaster should not pool the sample, but exclude the short periods of high local instability from the evaluation exercise.

The rest of the paper is organised as follows. In Section 2, we introduce the Diebold and Mariano (DM) and fluctuation test statistics, and we investigate their performance in case of a very large, localised shock. We also introduce some diagnostics to detect the presence of those shocks. In Section 3 we study the performance of these diagnostics in different Monte Carlo exercises; while in Section 4 we evaluate the ability of the Survey of Professional Forecasters (SPF) to outperform a naive benchmark when a period of high but localised instability due the COVID-19 shock is included in the study. In Section 5 we conclude.

2. DETECTING FORECAST BREAKDOWNS OVER VERY SMALL SAMPLES

2.1. Description of the environment

We consider the classical Giacomini and White (2006) framework, also see for example Giacomini and Rossi (2010) or Coroneo and Iacone (2020). To fix some notation, we denote the variable of interest by y_t, for which we want to compare two h-step ahead forecasts obtained from two alternative forecasting methods, based on some predictor variables x_t. We denote the observed vector by $w_t \equiv (y_t, x_t')'$, defined on a complete probability space $(\Omega, \mathcal{F}, P)$, and the information set at time t by $\mathcal{F}_t = \sigma(w_1', \ldots, w_t')$. The two h-step ahead forecasts for time t are based on the information set $\mathcal{F}_{t-h}$ and are denoted by $\widehat{y}_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right) \equiv f^{(i)}\left(w_{t-h}, w_{t-h-1}, \ldots, w_{t-h-R_i+1}; \widehat{\delta}_{t-h,R_i}^{(i)}\right)$ for $i = 1, 2$, where the forecasts are measurable functions of a sample of size R_1 for $f^{(1)}$ and R_2 for $f^{(2)}$. If a forecast is based on parametric models, the vector $\widehat{\delta}_{t-h,R_i}^{(i)}$ includes the estimates from the model. Otherwise, $\widehat{\delta}_{t-h,R_i}^{(i)}$ represents the semiparametric or nonparametric estimator used to construct the forecast. Notice that in this framework the estimates $\widehat{\delta}_{t-h,R_i}^{(i)}$ are based on a rolling window of dimension $R_i < \infty$.

For the two forecasts $\widehat{y}_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right)$, denote the forecast error by $e_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right) = y_t - \widehat{y}_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right)$ and, for a real function $L(\cdot)$, that we interpret as a loss function, the loss associated with the forecast error is $L\left(e_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right)\right)$. Finally, the loss differential at time t between the two forecasts is

$$d_t\left(\widehat{\delta}_{t-h,R_1}^{(1)}, \widehat{\delta}_{t-h,R_2}^{(2)}\right) = L\left(e_t^{(1)}\left(\widehat{\delta}_{t-h,R_1}^{(1)}\right)\right) - L\left(e_t^{(2)}\left(\widehat{\delta}_{t-h,R_2}^{(2)}\right)\right),$$

and the null hypothesis of equal predictive ability of the two forecasting methods is

$$H_0 : E\left(d_t\left(\widehat{\delta}_{t-h,R_1}^{(1)}, \widehat{\delta}_{t-h,R_2}^{(2)}\right)\right) = 0 \tag{2.1}$$

at each point t. Denoting $R \equiv \max(R_1, R_2)$, we assume that we have a sample of dimension $R + h + T - 1$ and we can therefore evaluate the hypothesis (2.1) in the $\{R + h, \ldots, R + h + T - 1\}$ evaluation period. To abbreviate the notation, denote $s = t - (R + h) + 1$ and

$$d_s \equiv d_t\left(\widehat{\delta}_{t-h,R_1}^{(1)}, \widehat{\delta}_{t-h,R_2}^{(2)}\right),$$

where notice that the evaluation period with respect to s is $\{1, \ldots, T\}$.
Let us denote μ_s by

$$\mu_s = E(d_s).$$

Then, the null hypothesis of equal predictive ability of the two forecasting methods is

$$H_0 : \mu_s = 0 \tag{2.2}$$

at each point $s \in \{1, \ldots, T\}$.
When H_0 is not met, then there is a $\mu_s \neq 0$ for some s in the evaluation period.

2.2. Diebold and Mariano equal predictive ability test

Diebold and Mariano (1995) propose to test the hypothesis in (2.2) using the sample

average

$$\overline{d} = \frac{1}{T}\sum_{s=1}^{T} d_s.$$

Denoting the long-run variance by

$$\sigma_T^2 = Var\left(\frac{\sqrt{T}}{T}\sum_{s=1}^{T} d_s\right)$$

and by $\widehat{\sigma}_T^2$ an estimate of σ_T^2, the Diebold and Mariano test (hereafter DM test) uses the statistic

$$t_{DM} = \sqrt{T}\frac{\overline{d}}{\widehat{\sigma}_T}.$$

When $\widehat{\sigma}_T^2 - \sigma_T^2 = o_p(1)$ and given other regularity conditions (see for example Assumption GW in Subsection 2.4), Giacomini and White (2006) show that, under H_0,

$$t_{DM} \to_d Z,$$

where Z is a standard normal distributed variable. When the null hypothesis is not met, the DM test has non-trivial power in presence of local alternatives $\mu_s = \delta T^{-1/2}$ for all s.

2.3. Giacomini and Rossi fluctuation test

Giacomini and Rossi (2010) describe the DM test as an average or global test, as it primarily detects deviations of the null hypothesis that are constant over the whole evaluation period. The DM test is less effective in detecting deviations from H_0 when they occur only on a fraction of the sample, and it might even have no power at all when μ_s changes sign over the evaluation sample so that $\sum_s \mu_s = 0$ is possible.

Therefore, Giacomini and Rossi (2010) propose to consider a local statistic, called the fluctuation statistic

$$Fl_{s,k} = \frac{\sqrt{k}}{k}\frac{1}{\widehat{\sigma}}\sum_{l=s-k/2}^{s+k/2-1} d_l,$$

where $k = \lfloor \kappa T \rfloor$ and we assume $k/T \to \kappa \in (0, \infty)$ as $k \to \infty$ and $T \to \infty$ as in Assumption 1(c) in Giacomini and Rossi (2010). They show that, under H_0 and regularity conditions,

$$Fl_{s,k} \Rightarrow \frac{B(\rho + \kappa/2) - B(\rho - \kappa/2)}{\sqrt{\kappa}},$$

where $B(\cdot)$ is a standard univariate Brownian motion and $\rho \in [\kappa/2, 1 - \kappa/2]$ is such that $s = \lfloor \rho T \rfloor$. The fluctuation test statistic is defined as

$$FL_\kappa = max_s |Fl_{s,k}|,$$

hence Giacomini and Rossi (2010) characterise the convergence to the limit distribution of the test statistic and provide simulated critical values for the test. The fluctuation test has power against a wide range of alternatives, requiring $\mu_s \neq 0$ only over an asymptotically non-negligible portion of the sample.

In comparison with the DM test, the fluctuation test should have less power when μ_s is constant, but more when the predictive ability is different only on a subsample, and

in situations of more general instability, including the case in which the total variation $\sum_s \mu_s$ is small relative to the sample size.

It is noteworthy that the critical values for the fluctuation test depend on the length of the fraction κ: the Monte Carlo study in Giacomini and Rossi (2010) suggests a certain size-power trade-off in the choice of κ, in the sense that very small values ($\kappa = 0.1$) are associated to size distortion in finite samples, whereas larger values are associated to lower power in presence of instability.

2.4. Equal predictive ability tests in presence of brief events

Giacomini and Rossi (2010) demonstrate the value of the fluctuation test in the forecasting of exchange rate macroeconomic models. As their example makes it clear, the natural application is in situations where the economic dynamics are slowly changing over time, and we can use the test to study the evolution, as in Rossi and Sekhposyan (2010) and Galvão et al. (2021). In other words, the test detects forecast differential instability across periods, or regimes, as it happens when forecasts from a macroeconomic model are compared against a benchmark in the presence of changes to the fundamental economic relations.

The fluctuation test is also effective in detecting the existence of what Timmermann (2008), emphasising the local nature of the predictability of returns, refers to as pockets of predictability, that only appear in some periods in time corresponding to fractions in the sample, as in Hillebrand et al. (2023). This situation, however, does not cover well the differential forecasting instability that occurs over only a small period of time, as is sometimes the case for economic recessions or other short-lived events.

In this case, the expected value of the differential predictive ability is neither constant over the entire evaluation sample, as assumed by the DM test, nor does it take on distinct values across specific segments of the sample, as addressed by the fluctuation test. Instead, it is better characterized as

$$\mu_s = \delta_2 T^a I_s(\tau),$$

where $I_s(\tau)$ is an indicator function, taking value 1 if $s = \lfloor \tau T \rfloor$ and 0 otherwise; the factor $\delta_2 T^a$ characterises the dimension of the change in the prediction differential in relation to the sample size.

For our study, we then assume for the loss differential d_s the data generating process

$$d_s = \delta_1 T^{-1/2} + \delta_2 T^a I_s(\tau) + u_s, \tag{2.3}$$

where u_s is a zero-mean process. In this case, the loss differential d_s is constant, but for a point in time $s = \lfloor \tau T \rfloor$: the situation of equal predictive ability corresponds to $\delta_1 = 0$ and $\delta_2 = 0$, while the constant, non-zero mean that is usually considered to investigate the local power of the DM test occurs when $\delta_1 \neq 0$ and $\delta_2 = 0$. Finally, $\delta_2 \neq 0$ is used to study the situation in which the difference in the forecasting ability occurs only for a very short time. The inclusion of $\delta_1 \neq 0$ when $\delta_2 \neq 0$ is necessary to establish one key conclusion, that the power of the DM test may drop to 0 even in presence of systematic deviations from the null hypothesis.

We first study the limit properties of the DM statistic assuming that the long-run

variance is estimated using the Bartlett kernel,

$$\widehat{\sigma}_T^2 = c_0 + 2\sum_{l=1}^{M} \frac{M-l}{M} c_l,$$

where c_l is the l-th sample covariance of d_s (with $l=0$ the sample variance) and M is a user-chosen bandwidth such that $M/T \to 0$ as $T \to \infty$. To establish the limit properties of the DM statistic, we introduce the following assumptions.

Assumption GW

(GW.1) u_s *is mixing with* ϕ *of size* $-r/(2r-2)$, $r \geq 2$; *or* α *of size* $-r/(r-2)$, $r > 2$;
(GW.2) $E\left(|u_s|^{2r}\right) < \infty$ *for all* s;
(GW.3) $Var\left(\frac{\sqrt{T}}{T}\sum_{s=1}^{T} u_s\right) > 0$ *for all* T *sufficiently large.*

REMARK 2.1. *When* $\delta_2 = 0$, *Assumption GW.*1 *may be formulated in terms of* w_s, *and GW.*2 *and GW.*3 *in terms of* d_s, *as in Giacomini and White (2006), to which we refer for a discussion of these assumptions.*

To characterise the local power of the test, it is also convenient to assume that there is $\sigma^2 = \lim_{T\to\infty} Var\left(\frac{\sqrt{T}}{T}\sum_{s=1}^{T} u_s\right)$.

THEOREM 2.1. *Under Assumptions GW.*1 – *GW.*3,

(i) if $a < 1/2$, *then* $t_{DM} \to_d Z + \frac{\delta_1}{\sigma}$;
(ii) if $a > 1/2$, *then* $|t_{DM}| \to_p 1$;
(iii) if $a = 1/2$, *then* $t_{DM} \to_d \frac{\sigma Z + \delta_1 + \delta_2}{\sqrt{\sigma^2 + \delta_2^2}}$.

We refer to Appendix A for a complete proof.

REMARK 2.2. *The limit in part (i) is the same that occurs for the DM test in the standard situation, in presence of a Pitman drift, and it is routinely used to present the local power of the test. The factor* $T^{-1/2}$ *multiplying* δ_1 *in* (2.3) *can be heuristically interpreted saying that drifts* δ_1 *that are too small with respect to the sample size are not detected; conversely, non-negligble drifts are eventually detected as the sample gets larger. The power is increasing in the signal-to-noise ratio,* δ_1/σ. *This limit is not affected by the presence of additional instability at time* s. *This means that a local instability does not affect asymptotically the usual properties of the DM test, provided that the magnitude is not too large* ($a < 1/2$). *Again, whether the magnitude of the sudden jump is too large is relative to the sample size. When the magnitude of the local instability is very large, as in part (ii), the absolute value of the DM statistic converges to 1, and the DM test has no power for conventional levels of significance. Notice that this also occurs when* $\delta_1 \neq 0$, *so one forecaster has relevant and systematic superior predictive ability against the other forecaster. The limit in part (iii) is intermediate between the other two.*
The DM test is therefore not able to detect superior forecasting ability when this is limited

to just one point in time. In fact, in case of very large local differentials, the power of the test drops to 0.
As the fluctuation test uses a fraction of the sample size that is proportional to the whole sample, and the same estimate for the long-run variance, qualitatively similar results also hold for the fluctuation test.

A similar argument of course holds when the instability affects more than one observation, provided that the number is very small relative to the sample. The COVID-19-induced instability seems the typical example of this situation, but Theorem 2.1 suggests that including in the evaluation exercise the recession induced by the financial crisis may also generate a power loss, although this should be more subdued as the size of the shock is less.

2.5. Detecting predictive superiority in presence of brief, extreme instability

Detecting forecasting superiority is therefore more difficult in cases of events that are limited in time and large in scale, and yet it is also usually more important for its policy implications. When the location of the event is known in advance, as it may be in the case of the recession induced by COVID-19, we propose to apply the predictive instability test of Andrews (2003), and we show that its application to evaluate forecasts is justified. When the location of the potential predictive instability is not known, the approach is rather similar to detecting outliers or extreme values, as in Leadbetter et al. (1983).

2.5.1. Predictive instability test when the location is known, the S test The test defined in Andrews (2003) is based on testing the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta \neq 0$ in

$$d_s = \mu + \theta I_s(\tau) + u_s, \tag{2.4}$$

where $I_s(\tau)$ is treated as a dummy variable taking a non-zero value only when $s = \lfloor \tau T \rfloor$, and u_s is a zero-mean process.
To simplify notation, we assume $\tau = 1$, as in Andrews (2003), where the generic τ situation is also briefly discussed.

In general, comparing residuals sum of squares from an unrestricted and restricted regression is done using an F test or, if the data are not normally distributed, using a χ^2 limit distribution. However, in this case, the usual asymptotic convergence in distribution of the F statistics to a χ^2_1 limit does not hold: intuitively, this is because δ is estimated using only one observation, so it is not possible to invoke a central limit theorem to establish the limit distribution of this estimate.

Instead, denote $\widetilde{\mu} = \frac{1}{T}\sum_{s=1}^{T} d_s$, the estimate of μ in the restricted model, and $\widehat{\mu} = \frac{1}{T-1}\sum_{s=1}^{T-1} d_s$ the estimate in the unrestricted model, so that the restricted and unrestricted residuals are $\widetilde{u}_s = d_s - \widetilde{\mu}$ and $\widehat{u}_s = d_s - \widehat{\mu}$, respectively. The idea is then to estimate the distribution of $\widetilde{u}^2_T$ with the sample distribution of $\widehat{u}^2_s$ for $s = 1, \ldots, T-1$.

To improve the empirical size performance in finite sample, Andrews proposes a slight modification of this procedure, where the critical distribution is estimated from $\widehat{u}_{2(s)} = d_s - \widehat{\mu}_{2(s)}$, where $\widehat{\mu}_{2(s)} = \frac{1}{T-2}\sum_{j=1, j\neq s}^{T-1} y_j$ (we refer to Andrews (2003) for the computation of $\widehat{\mu}_{2(s)}$ when the instability spans more than one period). The null hypothesis is rejected at α asymptotic significance level if $\widetilde{u}^2_T$ exceeds the $(1-\alpha)$ sample quantile of $\widehat{u}^2_{2(s)}$ for $s = 1, \ldots, T-1$.

To state the properties of the test, we introduce some additional notation, adapting the one presented in Andrews (2003). Denote the test statistic by S, so $S = \widetilde{u}_T^2$, and, for any $s \neq T$, let $S_s(\mu) = d_s - \mu = u_s$: under strict stationarity, all these variables have the same distribution, denoted as $F_S(x)$. Also, let $S_s = \widehat{u}_{2(s)}^2$, with empirical distribution $\widehat{F}_{S,T}(x) = \frac{1}{T-1}\sum_{s=1}^{T-1} 1(S_s \leq x)$, and let $q_{S,1-\alpha}$ denote the $(1-\alpha)$ quantile of $F_S(x)$, and $\widehat{q}_{S,1-\alpha}$ denote the $(1-\alpha)$ sample quantile of $S_{s,s\neq T}$. Finally, let S_∞ be a random variable with the same distribution as $d_T - \mu$.

Then we introduce the following assumptions:

Assumption A

A.1 w_t *is strictly stationary for all* t *if* H_0 *holds, and for* $t \neq T$ *otherwise;*
A.2 w_t *is ergodic for all* t *if* H_0 *holds, and for* $t \neq T$ *otherwise;*
A.3 $E(u_1)^2 < \infty$*;*
A.4 u_1 *has continuous and increasing distribution at the quantile* $1-\alpha$*.*

In Assumptions A.3 and A.4, we refer to u_1 for a generic u_s as the strict stationarity of w_t and the nature of the forecasting functions ensure the strict stationarity of u_s.

THEOREM 2.2. *Under assumptions A.1-A.4, as* $T \to \infty$*,*

(i) $S \to_d S_\infty$ *as* $T \to \infty$ *under* H_0 *and* H_1*;*
(ii) $\widehat{F}_{S,T}(x) \to_p F_S(x)$ *in a neighbourhood of* $q_{S,1-\alpha}$ *under* H_0 *and* H_1*;*
(iii) $\widehat{q}_{S,1-\alpha} \to_p q_{S,1-\alpha}$ *under* H_0 *and* H_1*;*
(iv) $P\left(S > \widehat{q}_{S,1-\alpha}\right) \to \alpha$ *under* H_0*.*

We refer to Appendix A for a complete proof.

REMARK 2.3. *Assumptions A.*1 *and A.*2 *are in terms of the observables* $w_t = (y_t, x_t')'$*. As in Andrews (2003), they are only referred to the period of stability.*
The restriction to stationarity rules out heterogeneity in the distribution, including heteroskedasticity, and in this sense it is stronger than requirements in Giacomini and White (2006) or in Theorem 2.1, where mixing is allowed. That was possible since the limit distribution of the test statistics was derived using central limit theorem arguments. In this case, we use the assumption of identical distribution to estimate the distribution of the residuals.
*Assumptions A.*3 *and A.*4 *are in terms of* u_s*: this is equivalent to (b) and (d) in Assumption LS in Andrews (2003): as similar assumptions on the unobservable term in Giacomini and White (2006) and Giacomini and Rossi (2010), these can be investigated on a case by case basis.*

REMARK 2.4. *The presentation and the statement of Theorem 2.2 allow for instability only at the end of the sample; this follows the outline in Andrews (2003). Instabilities at different points in time can also be considered, as it is discussed in Andrews (2003).*

In the interest of simplicity, we present Theorem 2.2 assuming that the instability only affects one point. Of course, it is possible to consider a longer span of observations,

as long as this remains finite and, in practice, also small concerning the sample size. When instability affects k observations $(T-k,\dots,T)$, Andrews (2003) shows that the procedure can be easily applied to quadratic forms of $\widetilde{U}_s = (\widetilde{u}_s,\dots,\widetilde{u}_{s+k-1})'$, $\widehat{U}_{2(s)} = (\widehat{u}_{2(s)},\dots,\widehat{u}_{2(s+k-1)})'$, adjusting the definition of $\widehat{\mu}_{2(s)}$ to account for the fact that more residuals are considered jointly.

One could compute quadratic forms directly from $\iota_k'\widetilde{U}_s$ and $\iota_k'\widehat{U}_{2(s)}$, where ι_k is a $k\times 1$ vector of ones. Andrews (2003) also proposes to account for the autocorrelation in u_s: denoting Σ the variance-covariance matrix of $U_s = (u_s,\dots,u_{s+k-1})'$, this quadratic form is computed from $\iota_k'\Sigma^{-1}\widetilde{U}_s$ as $\left(\iota_k'\Sigma^{-1}\widetilde{U}_s\right)'\left(\iota_k'\Sigma^{-1}\iota_k\right)^{-1}\left(\iota_k'\Sigma^{-1}\widetilde{U}_s\right)$ and similarly if $\iota_k'\Sigma^{-1}\widehat{U}_{2(s)}$ is used.

REMARK 2.5. *As Σ^{-1} is unobservable, Andrews (2003) proposes to estimate Σ using the restricted residuals $\widetilde{u}_s$ over the whole sample. We denote this estimate as $\widetilde{\Sigma}$.*
Andrews (2003) assumes that the explanatory variables do not depend on the sample size (T), ruling out in the regression (2.4) *a factor proportional to T^α, as in* (2.3)*. Thus, the estimates $\widetilde{\mu}$ and $\widetilde{\Sigma}$ are consistent, see Lemma 1 in Andrews (2003). When, however, model* (2.3) *is correct, the estimate $\widetilde{\Sigma}$ may fail to be consistent. This is clearly a relevant issue in our situation, and we explore it further in the Monte Carlo experiment. In this case, a consistent estimate of Σ may still be obtained using the residuals from the regression in the stability part of the sample only, $\widehat{u}_s$, and we denote it as $\widehat{\Sigma}$.*

2.5.2. Predictive instability when the exact location is not known, the MAX diagnostic
The test in Andrews (2003) is designed for situations in which the exact location of the suspected instability is known. In many cases, however, the exact location of the occurrence of a brief differential in predictive ability is not known in advance. This situation is rather more similar to the problem of detecting one of the "pockets of predictability" discussed in Timmermann (2008), as the forecast of a model is measured against a benchmark of no predictive ability.

It follows from Theorem 2.1 that global diagnostics like the DM test would not be able to detect such situations, regardless of the dimension of the occurrence. A recent procedure to detect such pockets was proposed by Harvey et al. (2021), where it is characterised as the occurrence of an outlier at an unknown point in time in a series. As they are interested in detecting the location of a pocket, as well as in testing its statistical significance, Harvey et al. (2021) divide their sample into many short intervals and compute a t statistic for each one. Our situation of interest is somewhat simpler, as we can just look at the statistic d_s^2.

The procedure consists of splitting the sample in a training period, $s=1,\dots,T^*$, and a monitoring period $s=T^*+1,\dots,E$ where $E\le T$, and T^* and E are fractions of the sample period $T^*=\lfloor\lambda_1 T\rfloor$, $E=\lfloor\lambda_2 T\rfloor$, for $0<\lambda_1<\lambda_2\le 1$. We assume that no instability occurs during the training period, but instability may take place during the monitoring period. We then compare the maximum of the statistic d_s^2 during the training period, $\max_{s=1,\dots,T^*} d_s^2$, and during the monitoring period, $\max_{s=T^*+1,\dots,E} d_s^2$. In practice, we use the training period to estimate the probability that $\max_{s=T^*+1,\dots,E} d_s^2 > \max_{s=1,\dots,T^*} d_s^2$ if no instability has occurred.

Notice that, heuristically, if the statistic of interest d_s^2 was independently and identically distributed, and the training and monitoring period constituted fractions λ_1 and $(1-\lambda_1)$

of the sample, respectively, then in large sample that probability of incorrectly detecting instability should be $(1-\lambda_1)$. Harvey et al. (2021) establish this result formally, and under conditions that allow for dependence in d_s.

Recall $w_t = (y_t, x_t')'$ and denote the element in position ι as $\{w_{\iota,t}\}$.

Assumption B

Let $\{w_{\iota,t}\}_{t\geq 1}$ be a strictly stationary sequence of random variables and $\{v_t(\xi)\}_{t\geq 1}$, $\{v_t(\zeta)\}_{t\geq 1}$ with $\xi,\ \zeta \in \mathbb{R}$, sequences of real numbers. For each $1 \leq i \leq j$, set $\mathcal{F}_i^j(v_t(\cdot))$ as the $\sigma-$algebra generated by the events $\{w_{\iota,r} \leq v_t(\cdot)\}$, $i \leq r \leq j$, and, for $1 \leq l \leq t-1$, denote

$$\alpha_{t,l}(\xi,\zeta) = \max_{1\leq k\leq t-l} \left\{ \left| P\left(A \cap B\right) - P(A)P(B) \right| \ :\ A \in \mathcal{F}_1^k\left(v_t(\xi)\right),\ B \in \mathcal{F}_{k+1}^t\left(v_t(\zeta)\right) \right\},$$

then there exists a sequence $l_t(\xi;\zeta) = o(t)$, as $t \to \infty$, such that $\lim_{t\to\infty} \alpha_{t,l_t}(\xi,\zeta) = 0$.

Assumption B is discussed in Ferreira and Scotto (2002) and is a very mild mixing condition, which relaxes regularity condition $D(u_n)$ in Leadbetter et al. (1983), page 53, which in turn is already a relaxation of strong mixing. Notice that the stationarity and mixing condition in Ferreira and Scotto (2002) is referred to the series d_s: differently, we formulate this requirement in terms of the observables w_t: the validity of this stationarity and mixing conditions for d_s follows from Theorems 3.35 and 2.49 of White (2000), recalling that R is finite so the application of this result is justified.

Assumption B is sufficient to establish Proposition 1 of Harvey et al. (2021). We then obtain

THEOREM 2.3. *Under assumption B, as $T \to \infty$,*

$$\lim P\left(\max_{s=T^*+1,\ldots,E} d_s^2 > \max_{s=1,\ldots,T^*} d_s^2 \right) = \frac{\lambda_2 - \lambda_1}{\lambda_2}.$$

Following Harvey et al. (2021) we therefore suggest considering $\max_{s=T^*+1,\ldots,E} d_s^2$ and $\max_{s=1,\ldots,T^*} d_s^2$ and conclude that there is instability if $\max_{s=T^*+1,\ldots,E} d_s^2 > \max_{s=1,\ldots,T^*} d_s^2$. Harvey et al. (2021) refer to this as the MAX procedure, and this is a test with size $\frac{\lambda_2-\lambda_1}{\lambda_2}$.

Notice that although the monitoring interval $[\lfloor \lambda_1 T \rfloor + 1, \lfloor \lambda_2 T \rfloor]$ is assumed to be proportional to the sample size, we still consider the MAX a diagnostic to detect instability even at a single point in time, or for a very short period, as that could be sufficient to cause $\max_{s=T^*+1,\ldots,E} d_s^2$ to exceed the threshold from $\max_{s=1,\ldots,T^*} d_s^2$.

For example, consider a sequence $\{\upsilon_s\}$ of independent standard normal distributions, observed at $s = 1,\ldots,T$, with a possible point of instability υ_{s^*}, with $E(\upsilon_{s^*}) = \mu_s^*$ and $s^* > T^*$. The MAX procedure compares $\max_{s=T^*+1,\ldots,E} \upsilon_s$ against $\max_{s=1,\ldots,T^*} \upsilon_s$, and it is well-known that the latter has stochastic order $O_p((\ln(T^*))^{1/2})$, see Theorem 1.5.3 in Leadbetter et al. (1983). Thus, the MAX procedure would have non-trivial power against cases with $\mu_s^* = c(\ln(T^*)^{1/2})$ when $c > 0$. Interestingly, this is slightly less than the local power for the test based on Andrews (2003) (which has non-trivial power if $\mu_s^* > 0$), but much more than the local power of the Fluctuation test or of the Diebold and Mariano test (as we have seen these have no power in this situation). Thus, we propose that the Andrews (2003) test and the MAX procedure may complement the information from the Fluctuation and DM tests.

3. MONTE CARLO STUDIES

In this section we investigate the size and power properties of the tests described in Section 2 by means of a Monte Carlo experiment.

We simulate

$$d_s = \mu + \theta I_s(\tau; m) + u_s,$$

where u_s is a zero-mean stationary process and $I_s(\tau; m)$ is an indicator function taking value 1 when $s \in \{\lfloor \tau T \rfloor, ..., \lfloor \tau T \rfloor + m - 1\}$ for a range of realistic parameters. The term m denotes the number of observations subject to instability: in Theorem 2.1 and in the discussion in Section 2 for ease of exposition we only considered $m = 1$ but in practice m may span more than one point in time: for example with quarterly data the COVID-19 period is better described as having $m = 3$.

To make sure that our results from Theorem 2.1 have practical relevance in reality, in the experiment we consider a data generating process based on our empirical example, in which we analyse the loss differential for two alternative nowcasts of the GDP growth, a naive forecast of the GDP consisting of the value of the previous quarter, and the nowcast from the Survey of Professional Forecasters.

For u_s we therefore assume an autoregressive model of order 1 (AR(1))

$$u_s = \phi u_{s-1} + \varepsilon_s,$$

where ε_s is independently and normally distributed, with $E(\varepsilon_s) = 0$, and $E(\varepsilon_s) = \sigma_\varepsilon^2$.

To match the empirical situation in Section 4, in our baseline design we assume: $\phi = 0.25$, $\sigma_\varepsilon^2 = 1.8$, $\mu = -1.3$, $\theta = -52$, and instability of length $m = 3$ (and spanning the last three observations), with a sample $T = 80$. However, we also consider different specifications of these parameters to investigate size and power. For each experiment, we run 10.000 replications.

For the size study, we set $\mu = 0$ and $\theta = 0$. As we are primarily interested in discussing the results in Section 2, we only consider $T = 80$. We, however, consider three values of ϕ ($\phi \in \{0, 0.25, 0.5\}$), as we want to verify that any significance that is observed in an empirical exercise is not spuriously generated by size distortion.

We summarise the results in Table 1. For the DM test, we consider the Bartlett (triangular) kernel, and two situations, setting the bandwidth M equal to 3 and 8. The first choice may be more common, reflecting the popular rules $M = \lfloor T^{2/9} \rfloor$ or $M = \lfloor T^{1/3} \rfloor$. It is well known, however, that the DM and similar tests may have very poor size properties in forecast evaluations, as the samples are usually relatively small (see Clark (1999) and Clark and McCracken (2013)). In such situations, treating the bandwidth-to-sample ratio as vanishing, as it is commonly done in standard inference, may be inappropriate: Kiefer and Vogelsang (2005) propose to treat it as a constant (fixed) b instead. The limit distribution for the test statistic is not a standard normal when this different assumption is imposed, but the alternative distribution may be tabulated and critical values for the appropriate b and kernel computed. This alternative approach to deriving the limit distribution is often referred to as *fixed-b* or *fixed smoothing* asymptotics. Moreover, treating the bandwidth-to-sample ratio as constant also allows the possibility of setting a relatively large bandwidth (even as large as the sample size, corresponding to $b = 1$). Kiefer and Vogelsang (2005) show that there is a trade-off between size and power, as larger values of b are associated with better size even in the presence of autocorrelation, but with less power. Simulations in Coroneo and Iacone (2020) suggest that $M = \lfloor T^{1/2} \rfloor$ is

Table 1. Global and local equal predictive ability tests, size study.

ϕ	DM_3	DM_8	$Fl_{0.1}$	$Fl_{0.3}$	S_1	S_I	$S_{\widetilde{\Sigma}}$	$S_{\widehat{\Sigma}}$	Max
0	0.047	0.045	0.017	0.036	0.043	0.055	0.051	0.056	0.052
0.25	0.071	0.055	0.028	0.067	0.045	0.057	0.054	0.058	0.056
0.5	0.118	0.070	0.072	0.145	0.050	0.062	0.061	0.066	0.061

Note: empirical size of the equal predictive ability tests DM, Fluctuation, and S test and the MAX procedure when $T = 80$. The theoretical size is set at 5% for all the tests.

a reasonable compromise, so our second choice is $M = 8$. The frequencies of rejections of the correct null hypothesis for these two tests are in columns DM_3 and DM_8, respectively.

For the fluctuation test, we consider κ equal to 0.1 and $\kappa = 0.3$. In this test too we need to estimate the long run variance and, as the critical values are only given for vanishing M/T ratio, we use $M = 3$. The empirical size is provided in columns $Fl_{0.1}$ and $Fl_{0.3}$.

As for the S test, the size when $m = 1$ is assumed in Column S_1. When $m > 1$, the test depends on Σ, and we consider three cases: in the first case we do not apply a correction for the dependence in the test statistic, corresponding to using the identity matrix; in the second case we use $\widetilde{\Sigma}$ as recommended in Andrews (2003), and in the third case we employ $\widehat{\Sigma}$, where Σ is estimated only using the pre-change sample. These three cases are denoted as S_I, $S_{\widetilde{\Sigma}}$ and $S_{\widehat{\Sigma}}$, respectively. Finally, for the MAX procedure, we consider a training period of 76 observations, i.e. $T^* = 76$ and $E = 80$, so that the asymptotic size is also 5% (see Column Max).

On balance, the results of the size study confirm the findings of other similar exercises: the DM test with standard asymptotics tends to be oversized in presence of positive autocorrelation (DM_3), but fixed smoothing asymptotics and a larger bandwidth address this problem (DM_8); the fluctuation tests are also oversized. The problem seems to be more severe for the $\kappa = 0.3$ case, and this is a surprise as the order is reversed compared to the finding in Giacomini and Rossi (2010). On the other hand, serial correlation does not seem to be a concern for the S test (regardless of the choice for Σ) and for the MAX procedure.

In the second part of the simulation exercise we study the power. Here we set $\phi = 0$ to avoid interference from the size distortion that could be generated by the autocorrelation, and we consider a range of values for μ and θ, for $m = 1$ and $m = 3$. We focus on the $T = 80$ sample but we also present results for $T = 40$ for comparison, in this case we focus only on the $m = 1$ case (we also adjust the bandwidths accordingly, and set them as 2 and 6, respectively). The results for $T = 80$, $m = 1$, for $T = 80$, $m = 3$, and for $T = 40$, $m = 1$ are in Table 2, 3 and 4, respectively. Notice that our reference values for μ and θ are both negative: we display the opposite value in the tables to facilitate readability.

We first consider situations in which $\mu = 0$ but $\theta \neq 0$: consistently with the prediction from Theorem 2.1, part *ii*, the DM and Fl tests cannot detect a local change in the mean, and the power goes in fact to 0 (which is even below the theoretical 5% that is set for the size). The only exception is for the $m = 3$, $\kappa = 1$ fluctuation test case, where power is associated to some values of θ: even when $T = 80$, $m = 3$ may be close enough to the interval that is considered in the test to still deliver power, at least for breaks of the dimension that we consider.

Next, we consider situations of $\mu \neq 0$ and $\theta = 0$. In this case the DM and fluctuation

Table 2. Global and local equal predictive ability tests, power study when $T = 80$, $m = 1$.

$-\mu$	$-\theta$	DM_3	DM_8	$Fl_{0.1}$	$Fl_{0.3}$	S_I	$S_{\widehat{\Sigma}}$	$S_{\widetilde{\Sigma}}$	Max
0	2.08	0.045	0.044	0.015	0.031	0.315	0.315	0.315	0.185
0	5.2	0.042	0.04	0.011	0.018	0.962	0.962	0.962	0.88
0	13	0.029	0.023	0.045	0.005	1	1	1	1
0	26	0.004	0.003	0.087	0	1	1	1	1
0	52	0	0	0.067	0	1	1	1	1
0	104	0	0	0.009	0	1	1	1	1
0.13	0	0.124	0.116	0.025	0.075	0.043	0.043	0.043	0.052
0.325	0	0.555	0.519	0.115	0.325	0.043	0.043	0.043	0.052
0.65	0	0.988	0.981	0.553	0.888	0.043	0.043	0.043	0.052
1.3	0	1	1	0.999	1	0.043	0.043	0.043	0.052
0.13	13	0.145	0.115	0.084	0.024	1	1	1	1
0.325	13	0.589	0.511	0.184	0.118	1	1	1	1
0.65	13	0.99	0.981	0.46	0.584	1	1	1	1
1.3	13	1	1	0.944	1	1	1	1	1
0.13	26	0.032	0.018	0.15	0.001	1	1	1	1
0.325	26	0.286	0.184	0.296	0.019	1	1	1	1
0.65	26	0.942	0.867	0.597	0.215	1	1	1	1
1.3	26	1	1	0.959	0.947	1	1	1	1
0.13	52	0	0	0.12	0	1	1	1	1
0.325	52	0.006	0.001	0.249	0	1	1	1	1
0.65	52	0.319	0.117	0.543	0.001	1	1	1	1
1.3	52	1	0.997	0.951	0.264	1	1	1	1

Note: empirical power of the equal predictive ability tests DM, fluctuation, and S test and the MAX procedure when $T = 80$ and $m = 1$. The theoretical size is set at 5% for all the tests.

tests have power against non-negligible deviations from the null hypothesis. These results are well established in the literature: we introduced this design to compare the situation to when $\theta \neq 0$ is also introduced, to investigate the prediction from Theorem 2.1, part *iii*. We find that, for given μ, introducing $\theta \neq 0$ causes a drop in power. Again, the case $\kappa = 0.1$ for the fluctuation test seems to be less sensitive to this phenomenon, at least for the parameters that we considered.

The S test and MAX diagnostics have no power when the deviation from the null hypothesis occurs at every point. On the other hand, they have the best power in the case of a very short deviation: interestingly, the S test has marginally more power compared to the MAX, perhaps reflecting the fact the S test assumes exact knowledge of the location of the shift, whereas the MAX procedure is more agnostic about this information. As for the three S statistics when $m > 1$, it seems that using $\widetilde{\Sigma}$ may have a very small advantage in size, and slightly less power. Overall, using $\Sigma = I$ seems to be a robust and economical choice from our experiment.

For the last exercise, we simulate the DGP with all the realistic choices for the parameters of interest, so $T = 80$, $\phi = 0.5$, $\mu = -1.3$, $\delta = -52$, $\sigma^2 = 1.8$, and $m = 3$. The results are provided in Table 5. In view of the poor size properties, we do not include the DM test when $M = 3$; despite the systematic deviation from the null hypothesis and the additional discontinuity at the end of the sample, the exercise suggests that we should

Table 3. Global and local equal predictive ability tests, power study when $T = 80$, $m = 3$.

$-\mu$	$-\theta$	DM_3	DM_8	$Fl_{0.1}$	$Fl_{0.3}$	S_I	$S_{\widehat{\Sigma}}$	$S_{\widehat{\Sigma}}$	Max
0	2.08	0.05	0.042	0.018	0.022	0.723	0.715	0.725	0.393
0	5.2	0.047	0.025	0.166	0.017	1	1	1	0.992
0	13	0.006	0	0.791	0.002	1	1	1	1
0	26	0	0	0.997	0	1	1	1	1
0	52	0	0	1	0	1	1	1	1
0	104	0	0	1	0	1	1	1	1
0.13	0	0.124	0.116	0.025	0.075	0.055	0.051	0.056	0.052
0.325	0	0.555	0.519	0.115	0.325	0.055	0.051	0.056	0.052
0.65	0	0.988	0.981	0.553	0.888	0.055	0.051	0.056	0.052
1.3	0	1	1	0.999	1	0.055	0.051	0.056	0.052
0.13	13	0.043	0.005	0.874	0.011	1	1	1	1
0.325	13	0.316	0.062	0.953	0.066	1	1	1	1
0.65	13	0.948	0.667	0.994	0.403	1	1	1	1
1.3	13	1	1	1	0.98	1	1	1	1
0.13	26	0	0	0.999	0	1	1	1	1
0.325	26	0.007	0	1	0	1	1	1	1
0.65	26	0.335	0.006	1	0.013	1	1	1	1
1.3	26	1	0.901	1	0.599	1	1	1	1
0.13	52	0	0	1	0	1	1	1	1
0.325	52	0	0	1	0	1	1	1	1
0.65	52	0	0	1	0	1	1	1	1
1.3	52	0.281	0	1	0	1	1	1	1

Note: empirical power of the equal predictive ability tests DM, fluctuation, and S test and the MAX procedure when $T = 80$ and $m = 3$. The theoretical size is set at 5% for all the tests.

not expect to see a rejection of the null hypothesis when the DM test is used; the $Fl_{0.1}$ on the other hand may still be informative.

Overall, these power simulations support the theoretical results from Section 2. For practical purposes, it is also important to understand what may realistically generate the situations that we considered. In particular, our choice of θ suggests a very extreme form of instability, like the one induced by the COVID-19 recession; however, our choice of μ is also somewhat exaggerated, as it is generated assuming a particularly poor benchmark. We, therefore, think that this design may suggest a realistic concern when evaluating situations as extreme as the recession induced by the COVID-19, but also a caution for situations less extreme, for example where the two predictions that are considered are not so different, so that μ is not very large relative to the noise. Finally, notice that the factor T^a in Theorem 2.1 would require an extremely large shock when T is large, as it is often the case with financial markets data. The risk is more relevant with macro data, as T tends to be smaller, especially when survey or quarterly data are used.

In view of these results, we recommend not to pool the sample when a large (relative to the sample size), localised instability takes place. Considering together the results of the DM and of the fluctuation tests, these findings also suggest that in practical situations the DM and fluctuation tests should be treated as complementary, rather than competing, diagnostics. The S test and MAX diagnostics may be of help to detect these situations.

Table 4. Global and local equal predictive ability tests, power study when $T = 40$, $m = 1$.

$-\mu$	$-\theta$	DM_2	DM_6	$Fl_{0.1}$	$Fl_{0.3}$	S_I	$S_{\widetilde{\Sigma}}$	$S_{\widehat{\Sigma}}$	Max
0	2.08	0.048	0.044	0.009	0.027	0.3	0.3	0.3	0.251
0	5.2	0.046	0.035	0.011	0.014	0.951	0.951	0.951	0.916
0	13	0.016	0.009	0.061	0.002	1	1	1	1
0	26	0	0	0.085	0	1	1	1	1
0	52	0	0	0.034	0	1	1	1	1
0	104	0	0	0.001	0	1	1	1	1
0.13	0	0.091	0.082	0.016	0.052	0.041	0.041	0.041	0.054
0.26	0	0.218	0.194	0.033	0.116	0.041	0.041	0.041	0.054
0.65	0	0.845	0.79	0.219	0.578	0.041	0.041	0.041	0.054
1.3	0	1	1	0.864	0.994	0.041	0.041	0.041	0.054
0.13	13	0.06	0.031	0.099	0.009	1	1	1	1
0.26	13	0.163	0.096	0.149	0.021	1	1	1	1
0.65	13	0.778	0.625	0.388	0.18	1	1	1	1
1.3	13	1	0.998	0.805	0.807	1	1	1	1
0.13	26	0.002	0	0.132	0	1	1	1	1
0.26	26	0.012	0.002	0.194	0	1	1	1	1
0.65	26	0.296	0.101	0.455	0.01	1	1	1	1
1.3	26	0.992	0.918	0.861	0.274	1	1	1	1
0.13	52	0	0	0.06	0	1	1	1	1
0.26	52	0	0	0.097	0	1	1	1	1
0.65	52	0	0	0.293	0	1	1	1	1
1.3	52	0.266	0.017	0.757	0	1	1	1	1

Note: empirical power of the equal predictive ability tests DM, Fluctuation, and S test and the MAX procedure when $T = 40$ and $m = 1$. The theoretical size is set at 5% for all the tests.

Table 5. Global and local equal predictive ability tests, power study.

DM_8	$Fl_{0.1}$	$Fl_{0.3}$	S_I	$S_{\widetilde{\Sigma}}$	$S_{\widehat{\Sigma}}$	Max
0.005	1	0.013	1	1	1	1

Note: the table exhibits the performance of the equal predictive ability tests DM, Fluctuation, and S test and the MAX procedure when $T = 80$ for a relistic design. The theoretical size is set at 5% for all the tests.

4. EVALUATING THE NOWCAST OF US NOMINAL GDP GROWTH

In this section, we illustrate the results obtained from the Monte Carlo exercise with an empirical example dedicated to evaluating the nowcast of US nominal GDP growth by using the GDP nowcast from the Survey of Professional Forecasters (SPF) over the period 2000:Q1 to 2020:Q3. In particular, we consider for the SPF the nowcast of nominal GDP, denoted $\widehat{y}_t$, that is made when the information on nominal GDP is only available up the previous quarter, y_{t-1}. As the survey covers many individuals, we use as $\widehat{y}_t$ the median of the responses for each point in time[1].

The nowcast of the quarterly growth rate is then $\frac{\widehat{y}_t - y_{t-1}}{y_{t-1}}$ and the error associated with the SPF nowcast is denoted as $e_t^{(1)} = \frac{y_t - \widehat{y}_t}{y_{t-1}}$. As a benchmark, we consider nowcasting

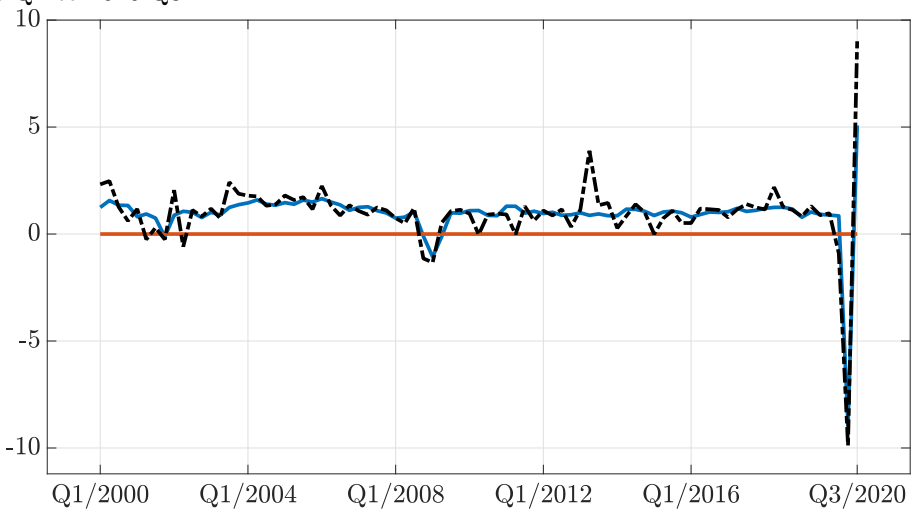
[1]We use the Q1:2024 vintage and consider column (4) of the SPF data for the Nominal GDP, labelled NGDP2, accordingly with the SPF documentation available at `https://www.philadelphiafed.org/-/media/frbp/assets/surveys-and-data/survey-of-professional-forecasters/spf-documentation.pdf?la=en&hash=F2D73A2CE0C3EA90E71A363719588D25`.

nominal GDP growth as 0, which corresponds to nowcasting the GDP as the last available observation, $\widetilde{y}_t = y_{t-1}$, so the error associated with this naive benchmark nowcast, $e_t^{(2)}$, is therefore $e_t^{(2)} = \frac{y_t - \widetilde{y}_t}{y_{t-1}}$. Using the quadratic loss function, the loss differential is then $d_t = e_t^{(1)2} - e_t^{(2)2}$.

Clearly, the benchmark in this exercise is not a very effective nowcast; for example, it even neglects the long-run growth in nominal GDP due both to economic growth and inflation. Moreover, it deviates from the common practice of using a constant growth or an autoregressive model, or a random walk, as it is the case when the benchmark is given by the previous value of GDP growth. However, it is a convenient one, in the sense that we expect that $E(d_t) < 0$, and therefore the null hypothesis should be rejected by the DM and fluctuation tests. If this prediction is not met, the exercise would highlight the inability of the DM and fluctuation tests to detect differences in predictive ability, even when these are large, in the presence of a short and large instability. This is then a fitting benchmark to check the predictions from Theorem 2.1.

Figure 1 provides the plot of the GDP growth, along with the SPF nowcast and the naive benchmark, while Figure 2 shows the errors.

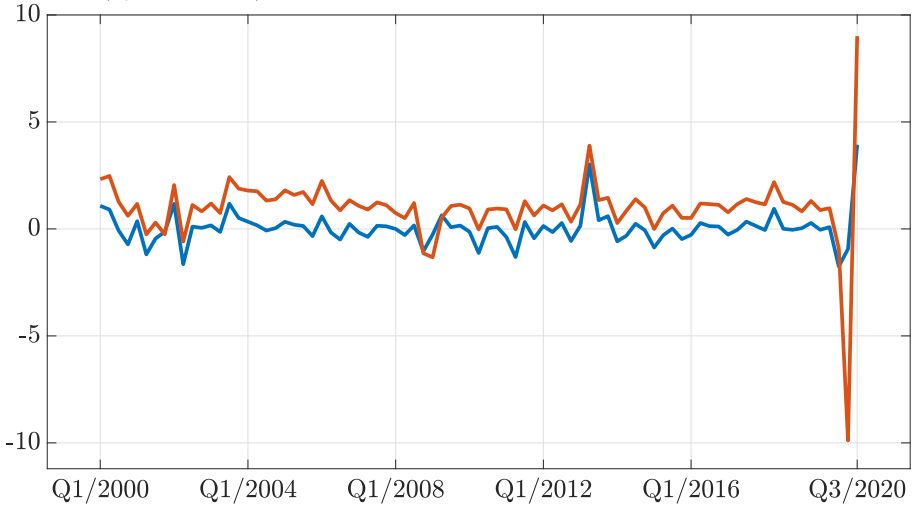
Figure 1. US nominal GDP growth (black dotted), GDP growth nowcast from the SPF (blue line, denoted as $\widehat{y}_t$), and GDP growth naive nowcast (red line, $\widetilde{y}_t$) over the period 2000:Q1 to 2020:Q3.



The SPF nowcast is always close to the target, except for an unanticipated and temporary surge of the growth rate in 2013; as anticipated, the naive benchmark forecast $\widetilde{y}_t$ does not take into account long-run real economic growth and inflation and the errors associated with it are then always positive, except for the 2008-2009 recession. Crucially, the SPF tracked quite well even the 2020 shock, whereas the naive errors are much higher for that period (in absolute value).

As a result (see Table 6), the performance in root mean square error terms (RMSE) of the naive benchmark worsens compared to the SPF nowcast, the ratio of the two RMSEs passing from 0.47 over the period up to 2019, to 0.39 when the three quarters in 2020 are

Figure 2. Errors associated to the SPF nowcast ($e_t^{(1)}$, blue line) and to the naive benchmark model ($e_t^{(2)}$, red line) over the period 2000:Q1 to 2020:Q3.



added, even though the RMSE increased for both sources. We then turn to the DM and the Fl tests, again evaluated with the RMSE loss function: in both cases, we estimated the long-run variance using the Bartlett kernel, but using bandwidth $M = 8$ for the DM test, and $M = 3$ for the fluctuation tests.

A summary of the results is available in Table 6. A negative entry for the DM statistic means that the average RMSE for the SPF nowcast is lower than the average RMSE of the benchmark; $Fl_{l:\kappa}$ and $Fl_{u:\kappa}$ are the minimum and maximum of the fluctuation test statistic for the two values of κ, respectively, where again a negative entry refers to the situation in which the RMSE of the benchmark exceeds the RMSE of the SPF nowcast, so the latter is more precise. The critical values are 2.261 for DM test (when $T = 80$; it is 2.250 when $T = 83$), and 3.393 and 3.012 for the $Fl_{0.1}$ and $Fl_{0.3}$ tests, respectively. In this example, both the DM and the Fluctuation tests suggest that the SPF nowcast is more precise than the benchmark, and significantly so if the sample is limited to up to 2019. As we anticipated the SPF does not seem to outperform the naive benchmark using the DM and the $Fl_{0.3}$ tests when the observations for 2020 are included in the sample, although the RMSE ratio is even more favorable.

Table 6. Nowcast evaluation over the period 2000:Q1 to 2019:Q4 and 2000:Q1 to 2020:Q3.

Period	RMSE_{SPF}	$\text{RMSE}_{\text{Naive}}$	Ratio	DM	$\{Fl_l, Fl_u\}_{0.1}$	$\{Fl_l, Fl_u\}_{0.3}$
2000:Q1-2019:Q4	0.608	1.289	0.472	-5.761	-4.800, -0.284	-5.414, -1.901
2000:Q1-2020:Q3	0.769	1.941	0.396	-1.751	-3.716, -0.031	-2.380, -0.206

Note: columns RMSE_{SPF} and $\text{RMSE}_{\text{Naive}}$ are the average RMSE for the SPF and Naive benchmark, respectively. Column Ratio refers to the ratio $\text{RMSE}_{\text{SPF}}/\text{RMSE}_{\text{Naive}}$. Columns DM, Fl_l, and Fl_u are the DM and lower and upper Fl test statistics when $\kappa = 0.1$ or 0.3. The 5% critical values for two-sided tests are 2.261 for the DM test (when $T = 80$; 2.250 when $T = 83$) and 3.393 and 3.012 for the two Fl_κ tests.

In conclusion, we analyse the performance of the two nowcasting models using the S

test from Andrews (2003) and the MAX diagnostic from Harvey et al. (2021) over the COVID-19 period (2020:Q1 - 2020:Q3). As the location of the COVID-19 recession and recovery can be treated as a shock at a known date, the application of the S test seems appropriate, and in our case we focus on just three observations. We thus set $m = 3$ for the S test, and $T^* = 80$, $E = 83$ for the MAX diagnostic (which corresponds to having size of approximately 3.6%). The application of the MAX test is particularly interesting in this situation: the test allows for continuous monitoring, indeed this is one feature that distinguishes it from the S test. We assume that no instability occurred until 2019, and start monitoring in the first quarter of 2020. As we continue the monitoring for three periods, the asympotic size is appropriately 3.6%.

Results for the two diagnostics are in Table 7: these both suggest an increase of the forecasting differential in the second period, but for the S statistic weighting the errors with the restricted residuals we fail to reject the null, consistently with the finding in the Monte Carlo experiments that this has less power, especially in presence of large shocks. For the MAX procedure, the last column q_{MAX} is the critical value that is obtained from the pre-COVID-19 period (2000:Q1 - 2019:Q4).

Table 7. Andrews (2003) S test evaluated for three different values of Σ and MAX procedure over the period 2020:Q1 to 2020:Q3 (3 observations).

S_I	q_{S_I}	$S_{\widetilde{\Sigma}}$	$q_{S_{\widetilde{\Sigma}}}$	$S_{\widehat{\Sigma}}$	$q_{S_{\widehat{\Sigma}}}$	MAX	q_{MAX}
7576	10.9	0.21	1.92	3060	3.6	96.84^2	6.03^2

Note:: Columns S_I, $S_{\widetilde{\Sigma}}$, and $S_{\widehat{\Sigma}}$, denote the S test statistics when the Identity matrix, the restricted residuals, and the unrestricted residuals are used to weight the errors, respectively. Columns q_{S_I}, $q_{S_{\widetilde{\Sigma}}}$, and $q_{S_{\widehat{\Sigma}}}$ are the respective critical values. The theoretical size is 5%. Column MAX denotes the maximum procedure over the period 2020:Q1 to 2020:Q3, while column q_{MAX} denotes the MAX over the period 2000:Q1 to 2019:Q4. The false positive rate of the procedure is 3.6%.

5. CONCLUSIONS

COVID-19 was an exceptionally challenging event for forecasting and evaluation since it was a moment of extreme instability spanning over a very short period. Tests like the Diebold and Mariano (1995) test for equal forecasting ability or the fluctuation test from Giacomini and Rossi (2010) test have less power, if the time span of the instability is very short.

In this paper, we show that in these situations, using non-parametric diagnostics (such as the S test or the MAX procedure) for local breaks or extreme values leads to more appropriate conclusions. We illustrate these results in a Monte Carlo exercise, and we provide evidence of the importance of selecting the correct time span for the test in a nowcasting exercise for the nominal US GDP, where we compare the SPF and a naive benchmark. Given these results, we recommend that the forecaster does not pool the sample, but excludes the short periods of high local instability from the evaluation exercise.

ACKNOWLEDGEMENTS

The authors thank the Editor, Raffaella Giacomini, and an anonymous referee for their very constructive and insightful comments, which helped improve the paper substantially.

The authors gratefully acknowledge the participants at the 31st SNDE Symposium in Padova, Fondazione Eni Enrico Mattei, and the 2024 European Association of Young Economists (EAYE) annual meeting in Paris for their helpful feedback. The authors also gratefully acknowledge Laura Coroneo and Annika Camehl for their useful comments and feedback. This research used the Computational resources provided by the Core Facility INDACO, which is a project of High-Performance Computing at the University of Milan. Fabrizio Iacone and Luca Rossini acknowledge financial support from the Italian Ministry of University and Research (MUR) under the Department of Excellence 2023-2027 grant agreement "Centre of Excellence in Economics and Data Science" (CEEDS).

REFERENCES

Almuzara, M., K. Baker, H. O'Keeffe, and A. Sbordone (2023). The New York Fed Staff Nowcast 2.0. New York FED Staff Technical Paper.

Andrews, D. W. (2003). End-of-sample instability tests. Econometrica 71, 1661–1694.

Bok, B., D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti (2018). Macroeconomic nowcasting and forecasting with big data. Annual Review of Economics 10, 615–643.

Clark, T. and M. McCracken (2013). Chapter 20 - Advances in Forecast Evaluation. In G. Elliott and A. Timmermann (Eds.), Handbook of Economic Forecasting, Volume 2, pp. 1107–1201. Elsevier.

Clark, T. E. (1999). Finite-sample properties of tests for equal forecast accuracy. Journal of Forecasting 18, 489–504.

Coroneo, L. and F. Iacone (2020). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. Journal of Applied Econometrics 35, 391–409.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics 20, 134–144.

Ferreira, H. and M. Scotto (2002). On the asymptotic location of high values of a stationary sequence. Statistics and Probability Letters 60, 475–482.

Foroni, C., M. Marcellino, and D. Stevanovic (2022). Forecasting the Covid-19 recession and recovery: lessons from the financial crisis. International Journal of Forecasting 38, 596–612.

Galvão, A. B., A. Garratt, and J. Mitchell (2021). Does judgment improve macroeconomic density forecasts? International Journal of Forecasting 37, 1247–1260.

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

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

Harvey, D. I., S. J. Leybourne, R. Sollis, and A. R. Taylor (2021). Real-time detection of regimes of predictability in the US equity premium. Journal of Applied Econometrics 36, 45–70.

Hillebrand, E., J. G. Mikkelsen, L. Spreng, and G. Urga (2023). Exchange rates and macroeconomic fundamentals: evidence of instabilities from time-varying factor loadings. Journal of Applied Econometrics 38, 857–877.

Huber, F., G. Koop, L. Onorante, M. Pfarrhofer, and J. Schreiner (2023). Nowcasting in a pandemic using non-parametric mixed frequency VARs. Journal of Econometrics 232, 52–69.

Kiefer, N. M. and T. J. Vogelsang (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. Econometric Theory 21, 1130–1164.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). Extremes and Related Properties of Random Sequences and Processes. New York, NY: Springer New York.

Lenza, M. and G. E. Primiceri (2022). How to estimate a vector autoregression after March 2020. Journal of Applied Econometrics 37, 688–699.

Ng, S. (2021). Modeling macroeconomic variations after Covid-19. Working Paper 29060, NBER.

Rossi, B. (2021). Forecasting in the presence of instabilities: how we know whether models predict well and how to improve them. Journal of Economic Literature 59, 1135–1190.

Rossi, B. and T. Sekhposyan (2010). Have economic models' forecasting performance for US output growth and inflation changed over time, and when? International Journal of Forecasting 26, 808–835.

Schorfheide, F. and D. Song (2024). Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic. International Journal of Central Banking 20, 275–320.

Timmermann, A. (2008). Elusive return predictability. International Journal of Forecasting 24, 1–18.

White, H. (2000). Asymptotic Theory for Econometricians. Emerald Group Publishing Limited.

A. APPENDIX

Proof of Theorem 2.1
Let us denote $\nu_s = \delta_2 T^a I_s(\tau)$, then we rewrite c_l as

$$c_l = c_l(uu) + c_l(u\nu) + c_l(\nu u) + c_l(\nu\nu),$$

where

$$c_l(uu) = \frac{1}{T}\sum_{s=l+1}^{T}(u_s - \overline{u})(u_{s-l} - \overline{u}),$$
$$c_l(u\nu) = \frac{1}{T}\sum_{s=l+1}^{T}(u_s - \overline{u})(\nu_{s-l} - \overline{\nu}),$$
$$c_l(\nu u) = \frac{1}{T}\sum_{s=l+1}^{T}(\nu_s - \overline{\nu})(u_{s-l} - \overline{u}),$$
$$c_l(\nu\nu) = \frac{1}{T}\sum_{s=l+1}^{T}(\nu_s - \overline{\nu})(\nu_{s-l} - \overline{\nu}).$$

So

$$\widehat{\sigma}_T^2 = \widehat{\sigma}_T^2(uu) + \widehat{\sigma}_T^2(\nu u) + \widehat{\sigma}_T^2(u\nu) + \widehat{\sigma}_T^2(\nu\nu),$$

where

$$\widehat{\sigma}_T^2(uu) = c_0(uu) + 2\sum_{l=1}^{M}\frac{M-l}{M}c_l(uu),$$

and $\widehat{\sigma}_T^2(\nu u)$, $\widehat{\sigma}_T^2(u\nu)$, $\widehat{\sigma}_T^2(\nu\nu)$ are defined in the same manner.
Under Assumptions GW.1 - GW.3, $\widehat{\sigma}_T(uu) - \sigma_T \to_p 0$ as in Theorem 4 of Giacomini and White (2006).
For the contribution to $\widehat{\sigma}_T^2(\nu\nu)$, first notice that

$$\overline{\nu} = \frac{1}{T}\sum_{s=1}^{T}\nu_s = \frac{1}{T}\delta_2 T^a.$$

Thus,

$$c_0(\nu\nu) = \frac{1}{T}\sum_{s=1}^{T}(\nu_s - \overline{\nu})^2 = \frac{1}{T}\sum_{s=1}^{T}\nu_s^2 - \overline{\nu}^2$$
$$= \frac{1}{T}(\delta_2 T^a)^2 - \left(\frac{1}{T}\delta_2 T^a\right)^2 = \delta_2^2 T^{2a-1} - \delta_2^2 T^{2a-2} = \delta_2^2 T^{2a-1} + o\left(T^{2a-1}\right).$$

Looking at $c_l(\nu\nu)$, for $\tau \in (0,1)$, and for T large enough, then $l < \lfloor \tau T \rfloor < T - l$, and

$$\begin{aligned}
&\frac{1}{T}\sum_{s=l+1}^{T} \nu_s \nu_{s-l} = 0, \\
&-\frac{1}{T}\sum_{s=l+1}^{T} \nu_s \overline{\nu} = -\frac{1}{T}\left(\delta_2 T^a\right)\left(\frac{1}{T}\delta_2 T^a\right) = -\delta_2^2 T^{2a-2}, \\
&-\frac{1}{T}\sum_{s=l+1}^{T} \nu_{s-l} \overline{\nu} = -\frac{1}{T}\left(\delta_2 T^a\right)\left(\frac{1}{T}\delta_2 T^a\right) = -\delta_2^2 T^{2a-2}, \\
&\frac{1}{T}\sum_{s=l+1}^{T} \overline{\nu}^2 = \frac{1}{T}(T-l)\left(\frac{1}{T}\delta_2 T^a\right)^2 = \frac{T-l}{T}\delta_2^2 T^{2a-2},
\end{aligned}$$

we obtain

$$c_l(\nu\nu) = -\delta_2^2 T^{2a-2} \ - \frac{l}{T}\delta_2^2 T^{2a-2} = -\delta_2^2 T^{2a-2} + o\left(T^{2a-2}\right)$$

so

$$\widehat{\sigma}_T^2(\nu\nu) = \delta_2^2 T^{2a-1} - M\delta_2^2 T^{2a-2} + o\left(T^{2a-1} + MT^{2a-2}\right) = \delta_2^2 T^{2a-1} + o\left(T^{2a-1}\right).$$

The approximation $\widehat{\sigma}_T^2(\nu\nu) = \delta_2^2 T^{2a-1} + o\left(T^{2a-1}\right)$ when $\tau = 0$ or $\tau = 1$ can be established in the same way, using $\frac{1}{T}\sum_{s=l+1}^{T} \nu_{s-l}\overline{\nu} = 0$ when $\tau = 1$, and $\frac{1}{T}\sum_{s=l+1}^{T} \nu_s \overline{\nu} = 0$ when $\tau = 0$.

Case $a < 1/2$**:**
We obtain $\widehat{\sigma}_T(\nu\nu)^2 = o(1)$, and, by the Cauchy-Schwarz inequality, $\widehat{\sigma}_T^2(u\nu) = o_p(1)$, and $\widehat{\sigma}_T^2(\nu u) = o_p(1)$, therefore $\widehat{\sigma}_T^2 - \sigma_T^2 \to_p 0$. As for the contribution of ν_s to the numerator of the DM statistic,

$$\frac{\sqrt{T}}{T}\sum_{s=1}^{T} \nu_s = T^{-1/2}\delta_2 T^a = o(1)$$

so

$$t_{DM} \to_d Z + \frac{\delta_1}{\sigma}.$$

Case $a > 1/2$**:**
We obtain $T^{1-2a}\widehat{\sigma}_T(\nu\nu)^2 \to \delta_2^2$, and $T^{1-2a}\widehat{\sigma}_T(uu)^2 \to_p 0$, so, by the Cauchy-Schwarz inequality, $T^{1-2a}\widehat{\sigma}_T(u\nu)^2 = o_p(1)$, and $T^{1-2a}\widehat{\sigma}_T(\nu u)^2 = o_p(1)$, therefore $T^{1-2a}\widehat{\sigma}_T^2 \to_p \delta_2^2$. As for the contribution of ν_s to the numerator of the DM statistic,

$$\begin{aligned}
&\frac{T^{1/2-a}\sqrt{T}}{T}\sum_{s=1}^{T} \nu_s \to \delta_2, \\
&\frac{T^{1/2-a}\sqrt{T}}{T}\sum_{s=1}^{T} \left(u_s + \delta_1 T^{-1/2}\right) \to_p 0,
\end{aligned}$$

so

$$t_{DM} \to_p \frac{\delta_2}{|\delta_2|}.$$

Case $a = 1/2$**:**
The proof follows combining arguments from the two previous examples, establishing for

the numerator

$$\frac{\sqrt{T}}{T}\sum_{s=1}^{T}\left(u_s+\delta_1 T^{-1/2}+\nu_s\right)\to_d \sigma Z+\delta_1+\delta_2$$

and, for the denominator,

$$\widehat{\sigma}_T^2(uu)\to_p \sigma^2,$$
$$\widehat{\sigma}_T^2(\nu\nu)\to_p \delta_2^2.$$

However, in this situation, we cannot rely on the Cauchy-Schwarz inequality to establish $\widehat{\sigma}_T^2(u\nu)=o_p(1)$, and $\widehat{\sigma}_T^2(\nu u)=o_p(1)$. Instead, we establish that these two terms converge to 0 in mean square.
We first observe that

$$E\left(c_l(\nu u)\right)=\frac{1}{T}\sum_{s=l+1}^{T}(\nu_s-\overline{\nu})E\left(u_{s-l}-\overline{u}\right)=0,$$

which also means that in $E\left(\widehat{\sigma}_T^2(u\nu)\right)=0$. To complete the argument we need to study $E\left(\widehat{\sigma}_T^2(u\nu)\right)^2$. To facilitate the discussion, we assume that u_s is stationary, denoting $\gamma_l=E\left((u_s-\overline{u})(u_{s-l}-\overline{u})\right)$ (notice that $\gamma_l=Cov\left(u_s,u_{s-l}\right)+O(1/T)$). The proof for the general case uses the same approach but with heavier notation:

$$\begin{aligned}E\left(c_l(\nu u)c_k(\nu u)\right)&=\frac{1}{T^2}\sum_{s=l+1,t=k+1}^{T}(\nu_s-\overline{\nu})(\nu_t-\overline{\nu})E\left\{\left(u_{s-l}-\overline{u}\right)\left(u_{t-k}-\overline{u}\right)\right\}\\&=I+II+III+IV\end{aligned}$$

where

$$\begin{aligned}I&=\frac{1}{T^2}\sum_{s=l+1,t=k+1}^{T}\nu_s\nu_t E\left\{\left(u_{s-l}-\overline{u}\right)\left(u_{t-k}-\overline{u}\right)\right\}=\\&=\frac{1}{T^2}\sum_{s=\max\{l,k\}+1}^{T}\nu_s^2 E\left\{\left(u_{s-l}-\overline{u}\right)\left(u_{s-k}-\overline{u}\right)\right\}\\&=\frac{1}{T^2}\sum_{s=\max\{l,k\}+1}^{T}\nu_s^2\gamma_{l-k}\end{aligned}$$

using the fact that $\nu_s\nu_t=0$ unless $s=t$. The latter expression is 0 if $\lfloor\tau T\rfloor<\max\{l,k\}$, and it is

$$I=\frac{1}{T^2}\delta_2^2 T\gamma_{l-k}=\frac{1}{T}\delta_2^2\gamma_{l-k}$$

otherwise.

Moving to IV, we obtain

$$\begin{aligned}IV&=\frac{1}{T^2}\sum_{s=l+1,t=k+1}^{T}\overline{\nu}^2E\left\{\left(u_{s-l}-\overline{u}\right)\left(u_{t-k}-\overline{u}\right)\right\}=\frac{1}{T^2}\delta_2^2\frac{1}{T}\sum_{s=l+1}^{T}\left(\sum_{t=k+1}^{T}\gamma_{s-l-t+k}\right)\\&=O\left(\frac{1}{T^2}\right),\end{aligned}$$

Regarding II, instead, we have

$$II = -\frac{1}{T^2} \sum_{s=l+1,t=k+1}^{T} \nu_s \overline{\nu} E\left\{\left(u_{s-l} - \overline{u}\right)\left(u_{t-k} - \overline{u}\right)\right\},$$

where this expression is 0 if $l < \lfloor \tau T \rfloor$; otherwise, letting $s^* = \lfloor \tau T \rfloor$ it is

$$II = -\frac{1}{T^2}\frac{1}{T^{1/2}}\delta_2 T^{1/2}\delta_2 \sum_{t=k+1}^{T} \gamma_{s^*-l-t+k} = O\left(\frac{1}{T^2}\right)$$

and proceeding in the same way we can establish

$$III = -\frac{1}{T^2} \sum_{s=l+1,t=k+1}^{T} \nu_t \overline{\nu} E\left\{\left(u_{s-l} - \overline{u}\right)\left(u_{t-k} - \overline{u}\right)\right\} = O\left(\frac{1}{T^2}\right).$$

Therefore,

$$E\left(c_l(\nu u)c_k(\nu u)\right) = \frac{1}{T}\delta_2^2\gamma_{l-k} + O\left(\frac{1}{T^2}\right)$$

if $\lfloor \tau T \rfloor < \max\{l,k\}$, and $O(\frac{1}{T^2})$ otherwise.
Finally, we now consider

$$E\left(\widehat{\sigma}_T^2(u\nu)\right)^2 = \sum_{l,k=-M}^{M} \frac{M-l}{M}\frac{M-k}{M} E\left(c_l(\nu u)c_k(\nu u)\right) = V + VI,$$

where the leading element V is

$$V = \delta_2^2 \frac{1}{T} \sum_{l,k=-M}^{M} \frac{M-l}{M}\frac{M-k}{M}\gamma_{l-k},$$

where notice that, using summation by parts,

$$\sum_{k=1}^{M}\left|\frac{M-k}{M}\gamma_{l-k}\right| \leq 0 + \sum_{j=1}^{M-1}\left|\frac{M-(j+1)}{M} - \frac{M-j}{M}\right|\left|\sum_{k=1}^{j}\gamma_{l-k}\right| = O\left(\sum_{j=1}^{M}\frac{1}{M}\right) = O(1)$$

so

$$V = O\left(\frac{1}{T}\sum_{l=-M}^{M}\frac{M-l}{M}\right) = O\left(\frac{1}{T}\sum_{l=0}^{M}\frac{l}{M}\right) = O\left(\frac{M^2}{TM}\right) = o(1)$$

and the last element VI is

$$VI = O\left(\sum_{l,k=-M}^{M}\frac{M-l}{M}\frac{M-k}{M}\frac{1}{T^2}\right) = O\left(\frac{M^2}{T^2}\right) = o(1).$$

Thus the proof follows.
□

Proof of Theorem 2.2
To prove this theorem, we need to show that the assumptions A.1-A.4 correspond to similar assumptions in Andrews (2003).

As $\widehat{y}_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right)$ is a function of $\{w_{t-h}, w_{t-h-1}, \ldots, w_{t-h-R_i+1}\}$ in view of Theorem 3.35 of White (2000), then $\widehat{y}_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right)$ are also stationary and ergodic. Similarly, $e_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right)$, $L\left(e_t^{(i)}\left(\widehat{\delta}_{t-h,R_i}^{(i)}\right)\right)$, and $d_t\left(\left(\widehat{\delta}_{t-h,R_1}^{(1)}\right), \left(\widehat{\delta}_{t-h,R_2}^{(2)}\right)\right)$ are also stationary and ergodic.

Assumptions A.1 and A.2 are sufficient to establish Assumption 1 in Andrews (2003), and Assumptions A.3 and A.4 correspond to similar assumptions in sufficient condition LS Andrews (2003), where, however, our situation is simpler because in our restricted model we only have a regression on a constant. □