SFB
823

Robust inference under time-varying volatility: A real-time evaluation of professional forecasters

Matei Demetrescu, Christoph Hanck,
Robinson Kruse

Discussion Paper

Nr. 6/2021

SFB
823

Robust Inference under Time-Varying Volatility: A Real-Time Evaluation of Professional Forecasters*

Matei Demetrescu[a], Christoph Hanck[b] and Robinson Kruse[c]

[a]Christian-Albrechts-University of Kiel† [b]University of Duisburg-Essen‡
[c]University of Hagen and CREATES, Aarhus University§

December 16, 2020

Abstract

In many forecast evaluation applications, standard tests (e.g., Diebold and Mariano, 1995) as well as tests allowing for time-variation in relative forecast ability (e.g., Giacomini and Rossi, 2010) build on heteroskedasticity-and-autocorrelation consistent (HAC) covariance estimators. Yet, the finite-sample performance of these asymptotics is often poor. "Fixed-b" asymptotics (Kiefer and Vogelsang, 2005), used to account for long-run variance estimation, improve finite-sample performance under homoskedasticity, but lose asymptotic pivotality under time-varying volatility. Moreover, loss of pivotality due to time-varying volatility is found in the standard HAC framework in certain cases as well. We prove a wild bootstrap implementation to restore asymptotically pivotal inference for the above and new CUSUM- and Cramér-von Mises based tests in a fairly general setup, allowing for estimation uncertainty from either a rolling window or a recursive approach when fixed-b asymptotics are adopted to achieve good finite-sample performance. We then investigate the (time-varying) performance of professional forecasters relative to naive no-change and model-based predictions in real-time. We exploit the Survey of Professional Forecasters (SPF) database and analyze nowcasts and forecasts at different horizons for output and inflation. We find that not accounting for time-varying volatility seriously affects outcomes of tests for equal forecast ability: wild bootstrap inference typically yields convincing evidence for advantages of the SPF, while tests using non-robust critical values provide remarkably less. Moreover, we find significant evidence for time-variation of relative forecast ability, the advantages of the SPF weakening considerably after the "Great Moderation".

Keywords: Forecast evaluation, Hypothesis testing, HAC estimation, Structural breaks, Bootstrap

JEL classification: C12 (Hypothesis Testing), C32 (Time-Series Models), C51 (Forecasting Models)

*The authors would like to thank three anonymous referees and the editor Barbara Rossi for their very constructive comments. Seminar and conference participants in Aarhus, Amsterdam, Bath, Bonn, Cologne, Graz, Hannover, Konstanz, Liverpool, Maastricht, Montreal, Münster, Rostock and Vienna, in particular Ulrich Müller provided very useful comments. The first two authors gratefully acknowledge the support of the German Research Foundation (DFG) through the projects DE 1617/4-2 and SFB 823, TP4. Kruse gratefully acknowledges financial support from CREATES funded by the Danish National Research Foundation (DNRF78).

†Institute for Statistics and Econometrics, Christian-Albrechts-University of Kiel, Olshausenstr. 40-60, D-24118 Kiel, Germany, e-mail address: `mdeme@stat-econ.uni-kiel.de`.

‡Faculty of Economics and Business Administration, University of Duisburg-Essen, Universitätsstraße 12, D-45117 Essen, Germany, e-mail address: `christoph.hanck@vwl.uni-due.de`.

§**Corresponding author:** University of Hagen, Faculty of Economics, Universitätsstr. 41, 58097 Hagen, Germany, e-mail address: `robinson.kruse-becher@fernuni-hagen.de` and CREATES, Aarhus University, School of Economics and Management, Fuglesangs Allé 4, DK-8210 Aarhus V, Denmark.

1 Introduction

Forecasting plays a crucial role in economics, finance and many other disciplines. Policy makers, firms, investors and households have various needs for macroeconomic predictions. Many of those are available, e.g., from the IMF and OECD, governmental forecasts like 'Teal Book' forecasts from the Federal Reserve, or commercial forecasters (e.g., Blue Chip Economic Indicators, Data Resources Inc. or the Survey of Professional Forecasters [SPF]). The SPF is the most comprehensive database available to assess the performance of professional forecasters. A fundamental question is then whether SPF forecasts outperform simple (model-based) alternatives, that is, have significantly smaller forecast error loss differentials on average. E.g., Zarnowitz and Braun (1993) reveal that SPF forecasts perform well in comparison to standard time series models (see also Croushore, 1993; Stark, 2010). With data from 1969 to 2017, we re-evaluate SPF forecasts for US output growth and GDP deflator inflation using robust inference methods.

This long evaluation period contains subsamples with structural changes mainly due to the "Great Moderation", but also during and after the "Great Financial Crisis". The "Great Moderation" is a period of considerable reduction in macroeconomic volatility as well as of sharp decline in predictability (Campbell, 2007). The "Great Financial Crisis" changed volatility, although to a lesser extent than the "Great Moderation", and yet less is known about its consequences on predictability. Changing macroeconomic volatility and changing predictability have important implications for forecast evaluation tests. While the first feature typically leads to time-varying volatility (in the sense of possibly unconditional heteroskedasticity over time) in forecast error loss differentials, the second might imply an instability of their mean. Ignoring these features may lead to significant size distortions and power losses; see the rich literature on forecasting in unstable environments (e.g., Giacomini and Rossi, 2010; Rossi, 2013; Coroneo and Iacone, 2020).

Here, we discuss the Diebold and Mariano (DM, 1995), fluctuation (Giacomini and Rossi, 2010) as well as new CUSUM and Cramér-von Mises tests from the perspective of time-variation, in particular time-varying volatility. While the DM test focuses on comparisons in stable environments, the latter three statistics capture time-varying relative forecast performance explicitly. The fluctuation, CUSUM and Cramér-von Mises statistics are however generally not robust to time-varying volatility, as their limiting null distributions depend on limit processes for partial sums, which do not converge to standard Wiener processes under time-varying volatility (cf. Section 2.2).

Moreover, we conduct the discussion in the "fixed-b" paradigm as pioneered by Kiefer and Vogelsang (2005). This paradigm goes beyond the standard heteroskedasticity- and autocorrelation consistent [HAC] framework (see the seminal contributions of Newey and West, 1987; Andrews, 1991), in which e.g. Diebold and Mariano (1995) and Giacomini and Rossi (2010) also derive their limiting distributions for the cited test statistics. HAC permits to use critical values from standard distributions, like the χ^2 or standard normal. These asymptotic distributions, however, turn out to be rather poor approximations to actual finite-sample distributions. Hence, substantial size distortions arise in practice. In particular, test results turn out to be sensitive to the choice of bandwidth B and kernel k employed for long-run variance estimation. The poor performance of HAC's asymptotic approximation can be explained by the "small-b" requirement that a vanishing fraction $b := B/P \to 0$ of the number of observations P be used for estimating autocovariances, while of course $b > 0$ in finite-samples. To tackle this issue, Kiefer et al. (2000) and Kiefer and Vogelsang (2002a,b, 2005) propose "fixed-b" asymptotics, which do not assume that $b \to 0$. This leads to nonstandard distributions (reviewed in Section 2). Conveniently and unlike in the standard small-b HAC framework, the new distributions reflect the choice of B and k even in the limit. The above papers convincingly demonstrate that the new distributions provide, in the absence of time-varying variances, substantially better approximations to actual finite-sample distributions. For these reasons, Choi and Kiefer (2010) advocate the use of Diebold and Mariano (1995) tests with fixed-b critical values; see also Li and Patton (2018). However, fixed-b critical values rely too on asymptotics for partial sums, which are affected by time-varying volatility, such that the fixed-b based Diebold and Mariano (1995) test then lacks pivotality, too.

Our main theoretical contribution is then to develop time-varying volatility-robust wild bootstrap versions of DM, fluctuation (Giacomini and Rossi, 2010) as well as the new CUSUM and Cramér-von Mises statistics under the fixed-b paradigm. We allow for parameter estimation error (West, 1996) in estimated nonnested forecast models, and cover both rolling window and recursive estimation for a fairly general nonlinear GMM setup.

In more detail, Section 2 rigorously shows time-varying variances to affect fixed-b limiting distributions of all the above four statistics (discussed in more detail in Subsection 2.1) and thus to lead to a loss of asymptotic pivotality (see also Müller, 2014, p. 314). This actually emphasizes a strength of the fixed-b approach, as it implies that the variability of the variances—influencing finite-sample behavior—is reflected in the limiting distribution. It does, however, come at the cost of yet different critical values. Such time-varying variances are pervasive in applied work in general and in our empirical application in Section 3 specifically.[1]

Adopting the parameter estimation framework of West (1996) (see Subsection 2.2), we characterize the resulting additional terms affecting the fixed-b distribution of the discussed tests for a class of generic nonlinear GMM estimators. We then develop a wild bootstrap correction (Subsection 2.3) replicating these features of the asymptotic distribution and establish its asymptotic validity. An appendix provides numerical results indicating considerable size distortions, due to time-varying volatility, resulting from using the non-bootstrapped conventional asymptotic critical values even in the limit. At the same time, the proposed bootstrap is shown to work well.

Section 3 compares the predictive ability of SPF forecasts for output and inflation to no-change and model-based approaches based on rolling window and recursive estimation. We focus on nowcasts, one-quarter and one-year ahead forecasts and evaluate these by considering the first and the final release of data. Overall, we find forecast error loss differentials to exhibit substantial heteroskedasticity. This has a direct impact on test decisions when comparing outcomes of traditional and our new robust tests: while the bootstrap provides strong evidence for the superiority of SPF forecasts (especially for nowcasts), there are notably fewer and weaker rejections when using asymptotic critical values. Our findings strongly suggest that SPF forecasts perform better early in the sample, but also that this advantage shrank considerably in the 1980s, leading to equal predictive ability starting in the mid-1980s. There are some signs of recoveries of forecast superiority around 2000 for GDP deflator inflation. We discuss our findings in relation to the literature on SPF accuracy, in general as well as with emphasis on the loss in relative predictability related to the "Great Moderation".

In recent related work, Coroneo and Iacone (2020) study the use of the full-sample Diebold and Mariano (1995) statistic $\mathcal{T}^{DM}$ for unconditional predictive ability testing. They adopt the framework of Giacomini and White (2006), i.e., they work with observed loss differentials—estimated from rolling forecasts—directly, and hence do not explicitly model effects of parameter estimation in the limiting distributions as we do in our nonlinear GMM setup. Next to an application of fixed-b inference using the Bartlett kernel, Coroneo and Iacone (2020) use an alternative weighted periodogram estimate of the long-run variance with associated "fixed-m" asymptotics to improve the finite-sample performance of $\mathcal{T}^{DM}$. Additionally, they compare the effectiveness of these testing approaches to a stationary block bootstrap (Politis and Romano, 1994). Their fixed-b and fixed-m

[1] Indeed, Groen et al. (2013) find variance changes to be important for inflation forecasting. More generally, time-varying volatility is present in many macroeconomic (e.g., Stock and Watson, 2002; Sensier and van Dijk, 2004; Justiniano and Primiceri, 2008; Clark and Ravazzolo, 2015) and financial (e.g., Guidolin and Timmermann, 2006; Rapach and Strauss, 2008; Amado and Teräsvirta, 2013) series such as economic growth, inflation and excess returns.

approaches rule out time-varying volatility.[2] Under time-varying volatility, as is also present in, e.g., their empirical applications to the SPF, Coroneo and Iacone (2020) suggest to split the sample into subsamples for which an assumption of constant variance is more credible and hence would allow for the use of standard fixed-b or fixed-m asymptotics. Sometimes, economic considerations (e.g., the "Great Moderation") may provide useful guidance about suitable splits of the whole sample. However, there are several problems with ad hoc choices regarding selected sample splits. These issues touch upon the unknown existence, number and locations of break points see, e.g., Rossi and Sekhposyan (2016). Our proposed tests do not require the researcher to possess such knowledge.

Section 4 concludes. A series of appendices collects proofs (unless indicated otherwise in the main text), other derivations, simulation results and further empirical results.

2 Fixed-b inference under time-varying volatility

2.1 Hypotheses and tests

We test the null of equal predictive ability of two competing forecasts for a target series z_t, either generated by models or obtained from surveys. We shall not assume a specific loss function but work with generic loss differentials directly (Diebold and Mariano, 1995),

$$y_t = \mathcal{L}_t\left(z_{t+h}, f_{1,t}\right) - \mathcal{L}_t\left(z_{t+h}, f_{2,t}\right). \tag{1}$$

Here, $f_{i,t}$, $i = 1, 2$, denote the competing h-step ahead forecasts for time $t + h$ and $\mathcal{L}_t(u_1, u_2)$ the loss function relevant at time t for horizon h. Typically, one focuses on one horizon h at a time, and we therefore avoid any explicit dependence of $f_{i,t}$ and $\mathcal{L}_t$ on h in the following.

The forecasts $f_{i,t}$ depend on various predictors (including, e.g., z_t and lags of z_t) in the model-based case, gathered in the vector $\boldsymbol{x}_{i,t}$, and on parameters of a model, say $\boldsymbol{\theta}_i \in \mathbb{R}^{M_i}$. Sometimes, $\boldsymbol{\theta}_i$ is known, and we write $f_{i,t} = f_i\left(\boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i\right)$ as "ideal forecasts".[3] In practice, however, parameters of forecast models are typically unknown, and one uses $\hat{f}^{\mathrm{r}}_{i,t} = f_i\big(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t}\big)$. The notation $\hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t}$ emphasizes that one can update the estimators over time, either in a rolling (r $=$ *rol*) or a recursive (r $=$ *rec*) fashion.

[2] The sampling properties of the periodogram also depend on time-varying volatility (see, e.g., Demetrescu and Sibbertsen, 2016).

[3] This includes cases such as driftless random-walk forecasts that do not require parameter estimation.

Time-variation in the loss differentials (1) may arise for a variety of reasons. The most obvious are time-varying features in the series z_{t+h} and the forecasts $f_{i,t}$, but changes in the loss function (such as different weights attached to forecast errors at different times) may also play a role. Less apparent but potentially no less important is the effect of parameter estimation, $\hat{f}^{\mathrm{r}}_{i,t} - f_{i,t}$; see Subsection 2.2.

We focus on tests of unconditional (cf. Remark 6 below for alternative cases) equal predictive accuracy for all t. Hence, the null of interest is that of a zero loss differential at each point in time (Giacomini and Rossi, 2010)

$$H_0 : \mathrm{E}\left(y_t\right) \equiv \mu_t = 0 \ \forall t,$$

extending the pair of hypotheses of “average” equal vs. unequal predictive ability as pioneered by Diebold and Mariano (1995). One may also consider one-sided alternatives, cf., e.g., Remark 3 below. Imposing constancy of μ_t has important consequences: as pointed out by Giacomini and Rossi (2010), one can expect some loss of power and reduced interpretability of rejections by tests based on falsely assuming a (time-)homogenous alternative. We follow the seminal work of Giacomini and Rossi (2010) and allow for time-variation in μ_t under the alternative (e.g., as a consequence of forecast breakdowns or other forms of structural instabilities in the relative predictive performance).

To accommodate parameter estimation, we follow closely the setup pioneered by West (1996). There are R preliminary observations used to obtain estimates $\hat{\boldsymbol{\theta}}_{1,R}$ and $\hat{\boldsymbol{\theta}}_{2,R}$. These are used to set up the forecasts $\hat{f}_{1,R}$ and $\hat{f}_{2,R}$, which are compared with z_{R+h}. Then, for the rolling window approach, one estimates the parameters using observations $t = 2, \ldots, R+1$ (resulting in $\hat{\boldsymbol{\theta}}^{rol}_{i,R+1}$), while the estimation sample is expanded by one observation for the recursive approach (resulting in $\hat{\boldsymbol{\theta}}^{rec}_{i,R+1}$). The forecast comparison is then conducted for $t = R+1$, until $t = R+P-1$. In total, P observations are available for forecast comparison, $z_{R+h}, \ldots, z_{R+P-1+h}$ together with $\hat{f}_{i,R}, \ldots, \hat{f}_{i,R+P-1}$. According to West (1996), R and P should go to infinity jointly, with $P/R \to \pi > 0$ to ensure that the estimation effect is reflected in the asymptotics.[4] To fix ideas, we focus on the class of (possibly overidentified) GMM estimators with at least as many moment conditions N_i as parameters M_i. Like in West (1996), pseudo-true values $\boldsymbol{\theta}_i$ are taken to exist, such that, as the sample size grows, one may write $\hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t} \xrightarrow{p} \boldsymbol{\theta}_i \ \forall \, t \geq R$, for $\mathrm{r} \in \{rol, rec\}$. Subsection 2.2 states precise assumptions on the estimators. The observed forecast losses are then given by $\mathcal{L}_t\big(z_{t+h}, \hat{f}^{\mathrm{r}}_{i,t}\big) \equiv$

[4] By considering the contribution of estimation uncertainty, our framework therefore focuses on (adopting the taxonomy of Giacomini and Rossi, 2010) comparing forecasting *models* (and, in so doing, on non-nested models) rather than comparing forecasting *methods*, as in, e.g., Giacomini and White (2006), where the losses depend on parameters estimated in sample using so-called limited-memory estimators.

$\mathcal{L}_t\big(z_{t+h}, f_i\big(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t}\big)\big)$, so one uses

$$\hat{y}^{\mathrm{r}}_t = \mathcal{L}_t\big(z_{t+h}, \hat{f}^{\mathrm{r}}_{1,t}\big) - \mathcal{L}_t\big(z_{t+h}, \hat{f}^{\mathrm{r}}_{2,t}\big), \qquad t = R, \dots, R+P-1, \tag{2}$$

for testing rather than the infeasible y_t.

Testing the null restriction $\mathrm{E}(y_t) = 0$ under the assumption of (time-)homogeneity may be done via a Wald-type statistic building on $\hat{y}^{\mathrm{r}}_t$ (Diebold and Mariano, 1995; West, 1996). Concretely, let

$$\mathcal{T}^{DM} = \frac{1}{P} \frac{\left(\sum_{t=R}^{R+P-1} \hat{y}^{\mathrm{r}}_t\right)^2}{\hat{\Omega}}, \tag{3}$$

where $\hat{\Omega}$ is a suitable estimator of the relevant long-run variance. Estimation of $\hat{\Omega}$ is discussed in more detail below. Considering heterogeneity, the first method used here to test $\mu_t = 0$ against $\mu_t \neq 0$ without imposing constant expectations is the fluctuations test of Giacomini and Rossi (2010). With $\hat{\Omega}$ based on all P pseudo out-of-sample observations available,[5] consider

$$\mathcal{T}^{F} = \max_{t \in \{\lfloor S/2 \rfloor + R, \dots, P+R-\lfloor S/2 \rfloor\}} \left| \frac{1}{\sqrt{S\hat{\Omega}}} \sum_{j=t-\lfloor S/2 \rfloor}^{t+\lfloor S/2 \rfloor - 1} \hat{y}^{\mathrm{r}}_j \right|, \quad S = \lfloor \nu P \rfloor \quad \text{with} \quad \nu \in (0,1)\,. \tag{4}$$

We consider two additional statistics to deal with time-varying relative predictive ability, namely a CUSUM-type and a Cramér-von Mises functional.[6] The CUSUM-type statistic is directly based on the partial sums of $\hat{y}^{\mathrm{r}}_t$,[7]

$$\mathcal{T}^{Q} = \max_{R \le t \le R+P-1} \sqrt{\frac{S_t^2}{\hat{\Omega} P}} \qquad \text{with} \qquad S_t = \sum_{j=R}^{t} \hat{y}^{\mathrm{r}}_j. \tag{5}$$

The Cramér-von Mises statistic is given by

$$\mathcal{T}^{C} = \frac{1}{P^2} \sum_{t=R}^{R+P-1} \frac{S_t^2}{\hat{\Omega}}. \tag{6}$$

Standard regularity conditions assumed, the small-b limiting distribution of $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$

[5] We hence follow Giacomini and Rossi (2010) and focus on a full-sample estimate of the long-run variance. In a time-varying framework like the present one, it is, following a suggestion of a referee, natural to also study time-varying estimates $\hat{\Omega}_t$ of the long-run variance. We investigate this option in our Monte-Carlo study, but find full-sample estimates to typically perform better, at least in the experiments considered there.

[6] These appear to be more popular in the statistical literature, with prominent econometric exceptions such as the KPSS test for stationarity.

[7] The (perhaps more familiar) CUSUM statistic for a break in mean involves $S_t/t - S_P/P$. This effectively demeans the series, and such a test is rather for a break in relative predictive power. We however test for departures from the null $\mu_t = 0$ rather than μ_t being a constant unknown mean, so centering S_t at zero is the natural choice here.

are known under unconditional homoskedasticity, and can be obtained as particular cases of Proposition 1 below, which deals with the encompassing case of time-varying volatility.

Let us now take a closer look at the long-run variance estimator. Given suitable choices for the kernel k and the bandwidth $B = \lfloor bP \rfloor$(see Newey and West, 1987; Andrews, 1991),

$$\hat{\Omega} = \hat{\gamma}_0 + 2 \sum_{j=1}^{P-1} k\left(j/B\right) \hat{\gamma}_j \tag{7}$$

is a long-run variance estimator with $\hat{\gamma}_j = P^{-1} \sum_{t=|j|+R}^{R+P-1} \left(y_t - \bar{y}\right)\left(y_{t-|j|} - \bar{y}\right)$. Regularity conditions assumed, $\hat{\Omega}$ is consistent for the long-run variance of y_t. Whenever y_t is unobserved, one computes $\hat{\Omega}$ based on $\hat{y}_t^{\mathrm{r}}$. However, West (1996) shows that, when parameters need to be estimated, the resulting long-run variance estimator does not standardize the partial sums of $\hat{y}_t^{\mathrm{r}}$ correctly in general. See Theorem 4.1 of West (1996), which also indicates how to explicitly correct the long-run variance estimator. Yet, we shall not require West's *explicit* correction here, since the wild bootstrap we use to deal with time-varying volatility in the fixed-b framework (see Subsection 2.3, and in particular Step 4 of Algorithm 1) *implicitly* correctly replicates the behavior of the test statistics in the limit by constructing bootstrap samples in such a way that they do capture the effect of estimation error.

Although (cf. Remark 1 below) the small-b asymptotic distributions of the above statistics do not depend on k and b,[8] Kiefer and Vogelsang (2005) argue for $\mathcal{T}^{DM}$ (and this extends to $\mathcal{T}^{x}$, $x \in \{F, Q, C\}$) that finite-sample dependence on tuning parameters translates into poor finite-sample behavior. To alleviate this, Choi and Kiefer (2010) resort to fixed-b asymptotics for $\mathcal{T}^{DM}$.

However, fixed-b based limiting distributions are affected by time-varying variances, such that one solution immediately prompts the next problem. Proposition 1 below contains a formal treatment; see also Demetrescu et al. (2019) and the references therein. To illustrate the main issues with such time-varying variances, consider the case of known parameters and tests based on $\mathcal{T}^{DM}$. To make the dependence of the distribution of $\mathcal{T}^{DM}$ on k and b explicit, Kiefer and Vogelsang (2005) let $b \in (0, 1]$ in the limit. Under homoskedasticity, the resulting limiting distribution is free of nuisance parameters (any scale matrix cancelling out), but is nonstandard. Concretely, Choi and

[8] Since $B = \lfloor bP \rfloor$, we may switch freely between the use of the bandwidth B and the fraction b; however, since b appears in the limit distributions, we use it from now on.

Kiefer (2010) show that

$$
\begin{aligned}
&\mathcal{T}^{DM} \xrightarrow{d} \mathcal{B}_{k,b} \qquad \text{with} \qquad \mathcal{B}_{k,b} = W^2(1)/\Lambda_{k,b}(W) \quad \text{and} \\
&\Lambda_{k,b}(W) \;\equiv\; \begin{cases} -\int_0^1 \int_0^1 \frac{1}{b^2} k''\left(\frac{r-s}{b}\right) \bar{W}(r)\bar{W}(s)\, \mathrm{d}r\mathrm{d}s & \text{for } k \text{ differentiable twice} \\ \frac{2}{b}\left(\int_0^1 \bar{W}(r)^2 \mathrm{d}r - \int_0^{1-b} \bar{W}(r+b)\bar{W}(r)\mathrm{d}r\right) & \text{for the Bartlett kernel,} \end{cases}
\end{aligned} \tag{8}
$$

where $\bar{W}(s) \equiv W(s) - sW(1)$ with $W(s)$ a standard Wiener process. The distinct feature of fixed-b asymptotics is that $\mathcal{B}_{k,b}$ depends on the *entire* path of the Wiener process $W(s)$ obtained as the limit process of the partial sums of y_t—and not only on $W(1)$, like for small-b. Since time-varying volatility implies a *different* limit for partial-sums processes (see, e.g., Cavaliere, 2004), this has important consequences for fixed-b when the volatility of y_t varies over time. Such dependence of the limiting distributions on the variance pattern extends to the case of estimated parameters and forecast instabilities; see Proposition 1 below.

Remark 1. For $b \to 0$, $\Lambda_{k,b}(W) \xrightarrow{d} 1$ and $\mathcal{B}_{k,b} \xrightarrow{d} \chi^2_1$ (Kiefer and Vogelsang, 2005). In this sense, small-b asymptotics are a particular case of fixed-b asymptotics. Interestingly, $\mathcal{T}^{DM}$ is asymptotically robust under the null to time-varying volatility under small-b asymptotics.[9] Yet, as mentioned above, the finite-sample quality of the HAC-based χ^2-approximation is poor, so the two extant options presented above effectively force practitioners to choose for $\mathcal{T}^{DM}$ between two problems under possible time-varying volatility: either non-pivotal fixed-b distributions, or asymptotically robust small-b distributions with poor finite-sample quality. □

2.2 Assumptions and limiting behavior

This subsection states our maintained assumptions on the DGP and GMM estimation with $N_i \geq M_i$ moment conditions, and provides relevant asymptotic theory.

Assumption 1. *Let* $\bar{\mathbf{C}}^b_{i,a} \equiv \sum_{j=a}^b \mathbf{C}_{i,j,\boldsymbol{\theta}_i}$. *For* $t = R, \ldots, R+P-1$ *and* $\mathrm{r} \in \{rol, rec\}$, *let the following decompositions hold:*

$$
\hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t} = \boldsymbol{\theta}_i + \left(\bar{\mathbf{C}}^{t,\prime}_{i,\mathcal{R}} \mathbf{W}_{i,\boldsymbol{\theta}_i} \bar{\mathbf{C}}^{t}_{i,\mathcal{R}}\right)^{-1} \bar{\mathbf{C}}^{t,\prime}_{i,\mathcal{R}} \mathbf{W}_{i,\boldsymbol{\theta}_i} \sum_{j=\mathcal{R}}^{t} \boldsymbol{a}_{i,j,\boldsymbol{\theta}_i} + \boldsymbol{r}^{\mathrm{r}}_{i,t}
$$

where $\mathcal{R} = t - R + 1$ *for* $\mathrm{r} = rol$ *and* $\mathcal{R} = 1$ *for* $\mathrm{r} = rec$. *Furthermore,*

[9] The explanation is that the full-sample sum in the numerator of the $\mathcal{T}^{DM}$ converges upon normalization to a normal distribution even under time-varying volatility, while the long-run variance estimator converges under small-b to the average long-run variance of the loss differentials as required for robustness (see Cavaliere, 2004).

(i) $\sup_{R<t\leq R+P}\left\|\boldsymbol{r}^{\mathrm{r}}_{i,t}\right\| = o_p\left(R^{-1/2}\right)$ *as* $R, P \to \infty$ *with* $P/R \to \pi$*,*

(ii) $\mathbf{W}_{i,\boldsymbol{\theta}_i} > 0$ *are deterministic, symmetric full-rank matrices,*

(iii) $E(\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}) = \mathbf{0}$ *and*

(iv) $\bar{\mathbf{C}}^t_{i,\mathcal{R}}$ *are full-rank with probability approaching unity as specified in Assumption 4 below.*

This assumption gives the usual linearized representation of a standard nonlinear GMM estimator which minimizes the suitably weighted quadratic form of sample moment conditions. The condition that $\mathrm{E}\left(\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}\right) = \mathbf{0}$ at the true $\boldsymbol{\theta}_i$ follows from specifying moment conditions for estimating $\boldsymbol{\theta}_i$. The $\mathbf{C}_{i,j,\boldsymbol{\theta}_i}$ are the Jacobians of the moment conditions and the $\mathbf{W}_{i,\boldsymbol{\theta}_i}$ are the limiting weighting matrices (note that the formulation allows for estimated optimal weights). The dependence on $\boldsymbol{\theta}_i$ arises from having possibly nonlinear moment conditions which are linearized for the asymptotics.

In the linear GMM case, the $\mathbf{C}_{i,j,\boldsymbol{\theta}_i}$ are simply the cross-products of instruments and regressors, while the $\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}$ are the products of instruments and regression errors, say, $\epsilon_{i,t}$. Moreover, $\boldsymbol{r}^{\mathrm{r}}_{i,t} = \mathbf{0}$ in the linear setup. For OLS, of course, regressors serve as instruments and weight matrices cancel out. We thus simply have that $\hat{\boldsymbol{\theta}}^{rol}_{i,t} = \boldsymbol{\theta}_i + \left(\sum_{j=t-R+1}^{t} \boldsymbol{x}_{j,t}\boldsymbol{x}'_{j,t}\right)^{-1} \sum_{j=t-R+1}^{t} \boldsymbol{x}_{i,t}\epsilon_{i,t}$ (and analogously for $\hat{\boldsymbol{\theta}}^{rec}_{i,t}$). Appendix A provides further details for the important special case of a linear regression.

In line with the literature (again, see West, 1996), we assume the loss and forecast functions to be smooth enough to allow for an evaluation of the impact of the estimation noise. The assumption covers leading loss functions such as squared error loss as well as generic forecast functions, cf. again Appendix A for a specific example. The gradient characterizing the impact of changes in the parameters on the loss is

$$\boldsymbol{d}_i(f, \boldsymbol{t}) = \left.\frac{\partial \mathcal{L}_t}{\partial u_2}\right|_{\substack{u_1 = z_{t+h} \\ u_2 = f}} \left.\frac{\partial f_i}{\partial \boldsymbol{\theta}}\right|_{\substack{\boldsymbol{x}_{i,t} \\ \boldsymbol{\theta} = \boldsymbol{t}}}, \tag{9}$$

and we assume it to be uniformly continuous in the following sense.

Assumption 2. *There exists* $0 < \epsilon < 1/2$ *such that, for the neighbourhood* $\Phi_P = \times_{i=1,2}\left\{\tilde{\boldsymbol{\theta}}_i : \left\|\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right\| < CP^{-1/2+\epsilon}, C > 0\right\}$ *of* $\left(\boldsymbol{\theta}'_1; \boldsymbol{\theta}'_2\right)'$*, it holds as* $R, P \to \infty$ *with* $P/R \to \pi$ *that*

$$\sup_{(\tilde{\boldsymbol{\theta}}'_1, \tilde{\boldsymbol{\theta}}'_2)' \in \Phi_P; t = R, \ldots, P+R-1} \left\|\boldsymbol{d}_i(\tilde{f}_{i,t}, \tilde{\boldsymbol{\theta}}_i) - \boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i)\right\| \xrightarrow{p} 0$$

where $\tilde{f}_{i,t} = f_i\left(\boldsymbol{x}_{i,t}, \tilde{\boldsymbol{\theta}}_i\right)$*,* $i = 1, 2$*.*

As a consequence, we may write

$$\hat{y}_t^{\mathrm{r}} = y_t + \sum_{i=1}^{2}(-1)^{i+1}\boldsymbol{d}_i'(f_{i,t},\boldsymbol{\theta}_i)\cdot\left(\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} - \boldsymbol{\theta}_i\right) + o_p(1), \qquad t = R,\ldots,R+P-1, \tag{10}$$

where the $o_p(1)$ term is negligible uniformly in t (see the proof of Lemma 2 below) and (the transpose of) $\boldsymbol{d}_i'(f_{i,t},\boldsymbol{\theta}_i)$ is defined in (9). Assumption 2 serves the same purpose as the corresponding Assumption 1(b) of West (1996) requiring a certain boundedness of second derivative of the $f_{i,t}$. The conditions are useful in this form for dealing with the bootstrap later on; see in particular the proof of consistency of our proposed bootstrap approach (Proposition 2) below. It is fulfilled, e.g., when the Jacobians of $\boldsymbol{d}_i$ are bounded on Φ_P. To describe the effect of the "estimation noise" terms $\boldsymbol{d}_i'(f_{i,t},\boldsymbol{\theta}_i)\cdot\left(\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} - \boldsymbol{\theta}_i\right)$, we make the following mild high-level assumption serving to guarantee a law of large numbers for the average of the derivatives to hold.[10]

Assumption 3. *As $P, R \to \infty$ with $P/R \to \pi$, the weak convergence $P^{-1}\sum_{t=R}^{R+[sP]-1}\boldsymbol{d}_i(f_{i,t},\boldsymbol{\theta}_i) \Rightarrow \boldsymbol{h}_i(s)$, $i = 1,2$ holds on $s \in [0,1]$, where $\boldsymbol{h}_i$ are Lipschitz-continuous deterministic vector functions.*

To quantify the departures from the standard small-b limits, we specify the behavior of the moment conditions *jointly* with that of y_t (and also characterize the limit behavior of the Jacobians of the moment conditions $\mathbf{C}_{i,j,\boldsymbol{\theta}_i}$):

Assumption 4. *Let $\boldsymbol{\xi}_t = \left(\boldsymbol{a}_{1,t,\boldsymbol{\theta}_1}', \boldsymbol{a}_{2,t,\boldsymbol{\theta}_2}', y_t - \mu_t\right)' \in \mathbb{R}^{N_1+N_2+1}$ s.t. $\boldsymbol{\xi}_t = \mathbf{G}(t/R)\tilde{\boldsymbol{v}}_t$. Assume that*

(i) $\mathbf{G}(u)$ is a matrix of piecewise Lipschitz functions, full-rank at all $u \in [0,1+\pi]$,

(ii) $\tilde{\boldsymbol{v}}_t$ has zero mean and unit long-run covariance, and is $L_{2+\delta}$-bounded for some $\delta > 0$, strictly stationary and strong mixing with mixing coefficients $\alpha(j)$ satisfying the summability condition $\sum_{j\geq 0}\alpha(j)^{1/p-1/(2+\delta)} < \infty$ for some $2 < p < 2+\delta$, and

(iii) there exist matrices $\mathbf{C}_i(u)$ of deterministic Lipschitz functions, full-rank for all $u > 0$, such that the weak convergence $R^{-1}\sum_{t=1}^{[uR]}\mathbf{C}_{i,t,\boldsymbol{\theta}_i} \Rightarrow \mathbf{C}_i(u)$ holds on $[0,1+\pi]$.

The structure of $\mathbf{G}$ is not restricted, since its role is to generate time-varying, symmetric, positive definite (local) long-run covariance matrices $\mathbf{G}(t/R)\mathbf{G}'(t/R)$ for $\boldsymbol{\xi}_t$. Assumption 4 allows for a wide range of patterns of time-varying volatility, including (possibly multiple) abrupt or smooth changes, as well as periodic patterns of heteroskedasticity. The assumption of a non-stochastic

[10] One could alternatively state slightly more low-level assumptions on average of $\boldsymbol{d}_i(f_{i,t},\boldsymbol{\theta}_i)$ for the full sample $t = 1,\ldots,R+P-1$. However, as can be seen in (10), one only needs observations at times $R,\ldots,R+P-1$, so that we state our assumption on $P^{-1}\sum_{t=R}^{R+[sP]-1}\boldsymbol{d}_i(f_{i,t},\boldsymbol{\theta}_i)$ directly.

variance function $\mathbf{G}(u)$ can moreover be relaxed, e.g., under independence conditions between $\mathbf{G}(u)$ and $\tilde{\boldsymbol{v}}_t$. The strong mixing condition is fairly mild, too; it is a typical requirement for CLTs and invariance principles for dependent sequences and allows, under suitable restrictions, for various forms of e.g. Markov switching or GARCH models (the surveys of Bradley, 2005, and Lindner, 2009, provide more technical discussions).

Partitioning $\mathbf{G}$ conformably with the components of $\boldsymbol{\xi}_t$, we note that the off-diagonal blocks induce (long-run) correlation of the moment conditions and the loss differentials, which may therefore be time-varying. Correspondingly, block diagonality of $\mathbf{G}$ implies asymptotic independence of the average moment conditions and the loss differentials, case in which the time-variation is rather in their marginal covariance matrices. Clearly, the mixing requirement on $\boldsymbol{\xi}_t$ and the deterministic limit of the sample averages of the Jacobians of the moment conditions imply short memory, so we do not allow for unit root behavior of regressors or instruments in the GMM estimation procedure. We obtain from, e.g., Smeekes and Urbain (2014, Lemma 1) the following partial sum behavior:

Lemma 1. *Under Assumption 4 with $\boldsymbol{W}$ a $N_1 + N_2 + 1$ vector of independent Wiener processes, $R^{-1/2} \sum_{t=1}^{[uR]} \boldsymbol{\xi}_t \Rightarrow \int_0^u \mathbf{G}(s) \mathrm{d}\boldsymbol{W}(s) \equiv (\boldsymbol{A}_1'(u), \boldsymbol{A}_2'(u), A_y(u))'$ on $[0, 1+\pi]$.*

The process $\int_0^u \mathbf{G}(s)\mathrm{d}\boldsymbol{W}(s)$ is Gaussian with independent, zero-mean increments, but not a Brownian motion as its quadratic variation $\int_0^s \mathbf{G}(r)\mathbf{G}'(r)\mathrm{d}r$ is nonlinear whenever $\mathbf{G}(\cdot) \neq const.$ In particular, this can occur due to breaks or smooth transitions in variances or covariances of $\boldsymbol{\xi}_t$. Its components $\boldsymbol{A}_i$ and A_y are simply the limit processes for the partial sums of the GMM moment conditions and the loss differentials, respectively. We then have the following behavior of the partial sums of $\hat{y}_t^{\mathrm{r}}$, $\mathrm{r} \in \{rol, rec\}$, in the evaluation period $t = R+1, \ldots, R+P$.

Lemma 2. *Let $\mathcal{A}(s) \equiv (A_y(1+s\pi) - A_y(1))/\sqrt{\pi}$, and, for $\mathrm{r} \in \{rol, rec\}$, $\tilde{\mathbf{C}}_i^{rol}(s) \equiv \mathbf{C}_i(1+\pi s) - \mathbf{C}_i(\pi s)$, $\tilde{\mathbf{C}}_i^{rec}(s) \equiv \mathbf{C}_i(1+\pi s)$, $\tilde{\boldsymbol{A}}_i^{rol}(s) \equiv \boldsymbol{A}_i(1+\pi s) - \boldsymbol{A}_i(\pi s)$ and $\tilde{\boldsymbol{A}}_i^{rec}(s) \equiv \boldsymbol{A}_i(1+\pi s)$. Under Assumptions 1–4 and the null $\mu_t = 0\ \forall t$, we have, for $s \in [0,1]$,*

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[sP]-1} \hat{y}_t^{\mathrm{r}} \Rightarrow \mathcal{A}(s) + \sqrt{\pi} \sum_{i=1}^{2} (-1)^{i+1} \int_0^s \boldsymbol{N}_i^{\mathrm{r}\prime}(r)(\mathbf{M}_i^{\mathrm{r}})^{-1}(r)\mathrm{d}\boldsymbol{h}_i(r) \equiv B_{\mathbf{G},\pi}^{\mathrm{r}}(s),$$

where $\mathbf{M}_i^{\mathrm{r}}(s) \equiv \tilde{\mathbf{C}}_i^{\mathrm{r}\prime}(s)\, \mathbf{W}_{i,\boldsymbol{\theta}_i}\, \tilde{\mathbf{C}}_i^{\mathrm{r}}(s)$ and $\boldsymbol{N}_i^{\mathrm{r}}(s) \equiv \tilde{\mathbf{C}}_i^{\mathrm{r}\prime}(s)\, \mathbf{W}_{i,\boldsymbol{\theta}_i}\, \tilde{\boldsymbol{A}}_i^{\mathrm{r}}(s)$.

Proof: *See Online Appendix.*

Remark 2. As already discussed by West (1996, Sec. 4), there are situations in which the effect of estimation error is negligible. Lemma 2 shows that it is sufficient that $\boldsymbol{h}_i(s) = \mathbf{0}$ for all s, as the weak limit of $P^{-1/2}\sum_{t=R+1}^{R+[sP]}\hat{y}_t^{\mathrm{r}}$ then only depends on the limit process for the loss differential, A_y. Verifying whether the condition $\boldsymbol{h}_i(s) = \mathbf{0}$ holds or not in a particular application requires information beyond the observed forecast errors. A sufficient condition for this to hold is that $\frac{\partial \mathcal{L}_t}{\partial u_2}$ has zero expectation and is uncorrelated with $\frac{\partial f_i}{\partial \boldsymbol{\theta}}$ for both $i = 1, 2$. The first condition (unbiasedness) is quite mild. The second, however, implies both $f_{1,t}$ and $f_{2,t}$ to be rational forecasts. The statistics under study test for equal predictive accuracy only, so rationality may be quite restrictive. It will, however, at least approximately be met in an interesting situation: under stationarity and estimation under the relevant loss, their product $\boldsymbol{d}_i\left(f_{i,t}, \boldsymbol{\theta}_i\right)$ may be close to zero because it represents a f.o.c. for the estimators (following from minimizing the observed loss, $\sum \mathcal{L}\left(z_{t+h}; f_{i,t}(\boldsymbol{\theta})\right)$ w.r.t. $\boldsymbol{\theta}$). See, e.g., Appendix A for a leading example. The bottom line is that, for all tests considered here, the estimation effect depends in general on the examined forecasting procedures via $\frac{\partial f_i}{\partial \boldsymbol{\theta}}$. In order to compare forecasts, one therefore requires information regarding their construction, i.e., information in addition to the point forecasts and the actual realizations, see West (1996) again.

Lemma 2 confirms that one also recovers the case without estimation error for $\pi \to 0$ (i.e., when "many" preliminary observations R are available relative to the forecasting periods P), where, again $P^{-1/2}\sum_{t=R}^{R+[sP]-1}\hat{y}_t \Rightarrow \mathcal{A}(s)$. At the same time, for $\pi \to \infty$, the estimation effect dominates.

When the researcher knows that she is in a situation like one of those discussed in this remark, she may simply set $\boldsymbol{d}_i = \mathbf{0}$ in Step 4 of the bootstrap algorithm 1 introduced in the following subsection. □

Since the processes $B_{\mathbf{G},\pi}^{\mathrm{r}}$ are not Brownian motions is general, Lemma 2 implies non-pivotal null distributions for the statistics of interest. With $\Lambda_{k,b}$ from (8), we have the following

Proposition 1. *Under the assumptions of Lemma 2 and the null $\mu_t = 0$ $\forall t$, we have for $B_{\mathbf{G},\pi}^{\mathrm{r}}$, $\mathrm{r} \in \{rol, rec\}$,*

$$\mathcal{T}^{DM} \xrightarrow{d} \left(B_{\mathbf{G},\pi}^{\mathrm{r}}(1)\right)^2 / \Lambda_{k,b}\big(B_{\mathbf{G},\pi}^{\mathrm{r}}\big), \qquad \mathcal{T}^{F} \Rightarrow \sup_{s\in[\nu/2,1-\nu/2]} \frac{1}{\nu} \frac{\left|B_{\mathbf{G},\pi}^{\mathrm{r}}\left(s+\frac{\nu}{2}\right) - B_{\mathbf{G},\pi}^{\mathrm{r}}\left(s-\frac{\nu}{2}\right)\right|}{\sqrt{\Lambda_{k,b}\big(B_{\mathbf{G},\pi}^{\mathrm{r}}\big)}}$$

$$\mathcal{T}^{Q} \Rightarrow \sup_{s\in[0,1]} \frac{\left|B_{\mathbf{G},\pi}^{\mathrm{r}}(s)\right|}{\sqrt{\Lambda_{k,b}\big(B_{\mathbf{G},\pi}^{\mathrm{r}}\big)}}, \qquad \mathcal{T}^{C} \Rightarrow \frac{1}{\Lambda_{k,b}\big(B_{\mathbf{G},\pi}^{\mathrm{r}}\big)} \int_0^1 \left(B_{\mathbf{G},\pi}^{\mathrm{r}}(s)\right)^2 \mathrm{d}s\,.$$

Proof: *See Online Appendix.*

Remark 3. Evidently, the limiting random variables presented in Proposition 1 may, together with suitable critical values (see Section 2.3), also be adopted for one-sided testing whenever the researcher has specific alternatives in mind. E.g., a signed version of (5), $\max_{R\leq t\leq R+P-1}\ S_t\Big/\sqrt{\hat{\Omega}P}$, together with large quantiles of $\sup_{s\in[0,1]} B^{\mathrm{r}}_{\mathbf{G},\pi}(s)\Big/\sqrt{\Lambda_{k,b}\big(B^{\mathrm{r}}_{\mathbf{G},\pi}\big)}$ may be used for right-tailed CUSUM-type tests. See Section 3 for an illustration of one-sided testing. □

Remark 4. Notwithstanding Remark 1, the limiting distributions of $\mathcal{T}^F$, $\mathcal{T}^Q$ and $\mathcal{T}^C$ depend on the entire path of the processes $B^{\mathrm{r}}_{\mathbf{G},\pi}$ via their numerator even when $b\to 0$. Therefore, small-b robustness to time-varying volatility is only given for $\mathcal{T}^{DM}$ in general. □

Given the dependence on time-varying variances in this particular form, a wild bootstrap is a natural candidate to restore asymptotically valid inference. See, e.g., Hansen (2000, p. 106) for an early application of the wild bootstrap to replicate sampling distributions affected by unconditional heteroskedasticity. We provide implementation details in the next subsection.

Remark 5. There are alternative ways to deal with time-varying (co)variances, some of which we explore in related work (Demetrescu et al., 2019). These build i) on estimating $\mathbf{G}$ and using the estimate to time-transform the series so as to restore homoskedasticity and hence apply standard fixed-b inference, or ii) on using a pretesting approach where, depending on the outcome of a test of no unconditional heteroskedasticity, either standard or heteroskedasticity robust fixed-b methods are used. We provide evidence that the wild bootstrap's performance is superior in terms of both size and power. We therefore focus in a wild bootstrap implementation here. □

Remark 6. Tests of equal *conditional* predictive ability are obtained by leveraging the loss differentials with a vector $\boldsymbol{w}_t$ of K suitable test functions (Giacomini and White, 2006). To cover this case, one may set $\boldsymbol{y}_t = \boldsymbol{w}_t\left(\mathcal{L}_t\left(z_{t+h}, f_{1,t}\right) - \mathcal{L}_t\left(z_{t+h}, f_{2,t}\right)\right)$ and correspondingly test the null $H_0 : \mathrm{E}\left(\boldsymbol{y}_t\right) = \mathbf{0}$. Appendix C contains the details of a multivariate implementation of tests of equal predictive accuracy. Of course, $w_t = 1$ recovers the unconditional approach on which we focus here. In any case, conditional tests are of course equally affected by time-varying volatility. □

2.3 A wild bootstrap correction

To correct for inherent non-pivotality via the wild bootstrap, the bootstrap scheme must replicate the properties of $B^{\mathrm{r}}_{\mathbf{G},\pi}$, $\mathrm{r}\in\{rol, rec\}$, in the limit. In particular, the wild bootstrap algorithm we propose focuses at replicating the *volatility-related* time-varying properties of all involved series. These properties depend, among others, on $\boldsymbol{h}_i(\cdot)$, $\mathbf{C}_i(\cdot)$, and the joint behavior of $A_y(\cdot)$ and $\boldsymbol{A}_i(\cdot)$. Since $\mathbf{C}_i(\cdot)$, $\mathbf{W}_i$ and $\boldsymbol{h}_i(\cdot)$ are deterministic, this can be achieved by jointly bootstrapping y_t and

$\boldsymbol{a}_{i,t}$. To do so, one must however resort to estimated quantities, since y_t and especially $\boldsymbol{a}_{i,t}$ are not observed directly (unless there is no estimation error, such that y_t is observed and the other quantities do not enter the test statistics at all). While $\hat{y}_t^{\mathrm{r}}$, $\mathrm{r} = \{rec, rol\}$, is a natural estimator for y_t, estimates of $\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}$, $\mathbf{W}_{i,\boldsymbol{\theta}_i}$ and $\mathbf{C}_{i,t,\boldsymbol{\theta}_{i,t}}$ require plugging in estimates of $\boldsymbol{\theta}_i$, leading to $\hat{\mathbf{C}}^{\mathrm{r}}_{i,t}$, $\hat{\mathbf{W}}^{\mathrm{r}}_{i,t}$ and $\hat{\boldsymbol{a}}^{\mathrm{r}}_{i,t}$:

Algorithm 1

1. Compute $\hat{y}_t^{\mathrm{r}}$ from (2) and $\hat{\mathbf{C}}^{\mathrm{r}}_{i,t}$, $\hat{\mathbf{W}}^{\mathrm{r}}_{i,t}$ and $\hat{\boldsymbol{a}}^{\mathrm{r}}_{i,t}$, $\mathrm{r} = \{rec, rol\}$ as follows:

   - For rolling window estimation:

$$\begin{aligned} \hat{\mathbf{C}}^{rol}_{i,t} &= \mathbf{C}_{i,t,\hat{\boldsymbol{\theta}}^{rol}_{i,R}}, \quad \hat{\mathbf{W}}^{rol}_{i,t} = \mathbf{W}_{i,\hat{\boldsymbol{\theta}}^{rol}_{i,R}}, \; \hat{\boldsymbol{a}}^{rol}_{i,t} = \boldsymbol{a}_{i,t,\hat{\boldsymbol{\theta}}^{rol}_{i,R}}, \quad \text{for} \quad t = 1, \ldots, R \\ \hat{\mathbf{C}}^{rol}_{i,t} &= \mathbf{C}_{i,t,\hat{\boldsymbol{\theta}}^{rol}_{i,t}}, \quad \hat{\mathbf{W}}^{rol}_{i,t} = \mathbf{W}_{i,\hat{\boldsymbol{\theta}}^{rol}_{i,t}}, \; \hat{\boldsymbol{a}}^{rol}_{i,t} = \boldsymbol{a}_{i,t,\hat{\boldsymbol{\theta}}^{rol}_{i,t}}, \quad \text{for} \quad t = R+1, \ldots, R+P-1. \end{aligned}$$

   - For recursive estimation: set $\hat{\boldsymbol{\theta}}^{rec}_{i,t} = \mathbf{0}$ for $t < N_i$ and compute

$$\hat{\mathbf{C}}^{rec}_{i,t} = \mathbf{C}_{i,t,\hat{\boldsymbol{\theta}}^{rec}_{i,t}}, \; \hat{\mathbf{W}}^{rec}_{i,t} = \mathbf{W}_{i,\hat{\boldsymbol{\theta}}^{rec}_{i,t}}, \; \hat{\boldsymbol{a}}^{rec}_{i,t} = \boldsymbol{a}_{i,t,\hat{\boldsymbol{\theta}}^{rec}_{i,t}}, \quad t = 1, \ldots, R+P-1.$$

   To save computing time, one may evaluate $\hat{\mathbf{C}}^{\mathrm{r}}_{i,t}$, $\hat{\mathbf{W}}^{\mathrm{r}}_{i,t}$ and $\boldsymbol{a}_{i,t,\cdot}$ at $\hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,R+P-1}$.

2. For $t = 1, \ldots, R+P-1$, construct wild bootstrap variates $\left(\boldsymbol{a}^{*,\prime}_{1,t}, \boldsymbol{a}^{*,\prime}_{2,t}, y^*_t\right)'$ as $\left(\hat{\boldsymbol{a}}^{\mathrm{r},\prime}_{1,t}, \hat{\boldsymbol{a}}^{\mathrm{r},\prime}_{2,t}, \hat{y}^{\mathrm{r}}_t\right)' r^*_t$, where the multipliers r^*_t are an i.i.d.(0,1) sequence, independent of the data, with $\mathrm{E}\left(|r^*_t|^w\right) < \infty \; \forall w \in \mathbb{N}$. Note that, for $t < R$, one may use any values for y_t and $\hat{y}^{\mathrm{r}}_t$ since these do not enter the test statistics $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$.

3. Construct the bootstrap analogues

$$\hat{\boldsymbol{\theta}}^{*,\mathrm{r}}_{i,t} = \left(\sum_{j=\mathcal{R}}^{t} \hat{\mathbf{C}}^{\mathrm{r},\prime}_{i,j} \hat{\mathbf{W}}^{\mathrm{r}}_{i,t} \sum_{j=\mathcal{R}}^{t} \hat{\mathbf{C}}^{\mathrm{r}}_{i,j} \right)^{-1} \sum_{j=\mathcal{R}}^{t} \hat{\mathbf{C}}^{\mathrm{r},\prime}_{i,j} \hat{\mathbf{W}}^{\mathrm{r}}_{i,t} \sum_{j=\mathcal{R}}^{t} \boldsymbol{a}^*_{i,j} + \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,R+P}$$

   for $t = R, \ldots, R+P-1$, where $\mathcal{R} = t - R + 1$ for $\mathrm{r} = rol$ and $\mathcal{R} = 1$ for $\mathrm{r} = rec$.

4. Letting $\hat{f}^{\mathrm{r},*}_{i,t} = f_i\big(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{i,t}\big)$, $\mathrm{r} \in \{rol, rec\}$, construct the bootstrap sample

$$\hat{y}^{\mathrm{r},*}_t = y^*_t + \boldsymbol{d}'_1(\hat{f}^{\mathrm{r},*}_{1,t}, \hat{\boldsymbol{\theta}}^{*,\mathrm{r}}_{1,t}) \cdot \left(\hat{\boldsymbol{\theta}}^{*,\mathrm{r}}_{1,t} - \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{1,R+P} \right) - \boldsymbol{d}'_2(\hat{f}^{\mathrm{r},*}_{2,t}, \hat{\boldsymbol{\theta}}^{*,\mathrm{r}}_{2,t}) \cdot \left(\hat{\boldsymbol{\theta}}^{*,\mathrm{r}}_{2,t} - \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{2,R+P} \right)$$

   for $t = R, \ldots, R+P-1$.

5. Using the bootstrap sample $\hat{y}_t^{\mathrm{r},*}$, $t = R, \ldots, R+P-1$, compute the bootstrap analogues $\mathcal{T}^{x,*}$, $x \in \{DM, F, Q, C\}$, of the test statistics (3)-(6).
6. Obtain the quantile(s) $q_{1-\alpha}^{x,*}$, $x \in \{DM, F, Q, C\}$, of the respective bootstrap distributions.

In practice, the distribution functions of the bootstrap statistics $\mathcal{T}^{x,*}$ are not known, but can be simulated in the usual way by repeating Steps 2 – 5 M times for a reasonably large M to obtain consistent empirical analogues via Monte Carlo simulation. Typical choices for the distribution of r_t^* are the Gaussian, Rademacher, or Mammen (1993) distributions.

Some additional conditions are required for establishing the validity of this bootstrap.

Assumption 5.

(i) $\mathbf{W}_{i,\boldsymbol{\theta}_i}$ *is continuous in* $\boldsymbol{\theta}_i$*,*

(ii) for $\max\{N_1, N_2\} \leq t \leq R + P - 1$*,* $\sup_t \left\| \hat{\mathbf{C}}_{i,t}^{\mathrm{r}} - \mathbf{C}_{i,t,\boldsymbol{\theta}_i} \right\| \xrightarrow{p} 0$*,*

(iii) $\exists \gamma > 0$ *such that* $\sup_t \|\boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i)\| = O_p\left(P^{1/2-\gamma}\right)$ *and* $\sup_t \left\| \hat{\boldsymbol{a}}_{i,t}^{\mathrm{r}} - \boldsymbol{a}_{i,t,\boldsymbol{\theta}_i} \right\| = O_p\left(P^{-\gamma}\right)$*,*

(iv) $\mathrm{E}\left(\tilde{\boldsymbol{v}}_t \tilde{\boldsymbol{v}}_t'\right) = c \cdot \mathbf{I}_{N_1+N_2+1}$ *with* $c > 0$*.*

Proposition 2. *Under Assumptions 1–5, it holds under the null* $\mu_t = 0 \; \forall t$ *that*

$$P\left(\mathcal{T}^x \geq q_{1-\alpha}^{x,*}\right) \to \alpha, \quad x \in \{DM, F, Q, C\}, \qquad \textit{as} \quad R, P \to \infty \textit{ with } P/R \to \pi.$$

Proof: *See Online Appendix.*

Remark 7. The additional Assumption 5(i)-(iii) refers essentially to required smoothness of $\hat{\mathbf{C}}_{i,t}^{\mathrm{r}}$ and $\hat{\boldsymbol{a}}_{i,t}^{\mathrm{r}}$ as functions of the estimators, and is fulfilled in, e.g., the linear GMM case; see for example Appendix A. In a nutshell, it transfers the smoothness requirements from Assumption 2 to the bootstrap world. Assumption 5(iv) implies the proposed bootstrap scheme to asymptotically work under the additional condition that $\mathrm{E}\left(\tilde{\boldsymbol{v}}_t \tilde{\boldsymbol{v}}_t'\right) = c \cdot \mathbf{I}_{N_1+N_2+1}$, namely that the covariance and long-run covariance matrices of $\tilde{\boldsymbol{v}}_t$ are proportional. This is trivially fulfilled in the case without estimation error, and may for example also be side-stepped when there is one factor driving the volatility changes having the same impact on all components, i.e., when $\mathbf{G}(s) = g(s) \cdot \mathbf{G}_0$ for some constant full-rank matrix $\mathbf{G}_0$ and $g(s)$ a piecewise Lipschitz scalar function. A further slightly more restrictive example of this condition being fulfilled is given in case of common dynamics. That we require this condition is a consequence of using a plain-vanilla wild bootstrap in step 2 of the above algorithm, which imposes no serial correlation in the bootstrap error replicates, therefore producing equal covariance and long-run covariance matrices (conditional on the data). The condition would be violated when, e.g., the researcher overdifferences the involved series to obtain a reduced-rank long-run covariance matrix. In such cases, one could for example resort to a sieve wild bootstrap (see, e.g., Cavaliere et al., 2010, for an implementation in co-integrated models with time-varying volatility) or, in a less parametric vein, to a block wild bootstrap (see, e.g., Smeekes and Urbain, 2014, who explicitly permit singular long-run covariance matrices), both of which allow to capture the relevant long-run covariance matrix. □

Remark 8. As argued in the proof of Proposition 2, $q_{1-\alpha}^{x,*}$ remains unaffected under local alternatives $\mu_t = R^{-1/2}\mu(t/R)$ with μ a non-zero deterministic Lipschitz function $\mu(\cdot)$; see the discussion following eq. (16) in Appendix B. At the same time, the limiting behavior of $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$ can easily be seen to change, so that the bootstrap tests have nontrivial local power. □

Remark 9. The algorithm is easily modified to account for the case where only one of the forecasts involves estimated parameters, or when the two forecasts resort to different estimation schemes, one rolling and the other recursive. □

Remark 10. While the bootstrap from Algorithm 1 is feasible when a researcher possesses all the necessary information regarding the construction of the forecast, some external sources (cf. Section 3) only publish point forecasts and actual realizations. Such information is not sufficient to assess the relative strengths of privately constructed forecast *models*. Among others, the covariance of $\boldsymbol{A}_i$ and A_y is often not known to "outsiders", making it impossible to apply a suitable bootstrap. □

Remark 11. Appendix D presents the results of extensive Monte Carlo simulations confirming good finite-sample performance of the bootstrap versions of all statistics considered in this section. □

Remark 12. Multiple forecast comparisons, e.g., of the kind used for model confidence sets (Hansen et al., 2011), may also be implemented using the proposed bootstrap procedure. □

3 Empirical results

3.1 The Survey of Professional Forecasters data - summary statistics

The survey started in 1968 (conducted by the American Statistical Association and the National Bureau for Economic Research) and is administered by the Federal Reserve Bank of Philadelphia since 1990. Participants are asked to predict main US macroeconomic variables in the middle of each quarter for the current and the following four quarters. We consider two key variables: output growth (RGDP, “Real Gross National Product/Gross Domestic Product”) and inflation (PGDP, “Price Index for Gross National Product/Gross Domestic Product”).[11]

Our sample includes the 1970s with its severe oil price shocks, leading to increases in macroeconomic volatility and conversely, the “Great Moderation”, lasting until the mid-1980s, which exhibited a sharp decline in volatility and predictability (see Campbell, 2007). It is well-documented that the “Great Moderation” led to enhanced macroeconomic stability which eased forecasting in general, but also made it more difficult to beat simple time series models (see, e.g., Stock and Watson, 2007). Similarly, Groen et al. (2013) find that regime changes in the variance play an important role for real-time (inflation) forecasting. The sample also covers the “Great Financial Crisis” in 2007/2008. Such a long sample is interesting as it may be possible to identify different episodes in relative forecast performance.

We consider three horizons, viz. nowcasting ($h = 0$), one-quarter ahead ($h = 1$) and one-year ahead ($h = 4$) forecasts, and two vintages (the first and final releases). Macroeconomic data is often revised significantly, see Croushore and Stark (2001). Faust and Wright (2013) and Stark (2010) discuss and demonstrate the importance of the vintage structure when evaluating SPF (inflation) forecasts. We compare the SPF to model-based forecasts generated in real-time to enable a fair comparison with regard to the available information; see also Stark (2010), D’Agostino et al. (2006) and Coroneo and Iacone (2020).

The dynamic forecast models are economically motivated and include a predictor x_t and an autoregressive term: $z_t = \theta_0 + \theta_1 x_{t-1} + \theta_2 z_{t-1} + e_t$. For output, we use the term spread (in short: TMS), i.e., the difference between long-term bond rates and short-term yields, as a predictor. Important references include Estrella and Hardouvelis (1991) for the term spread being an important predictor of real output and Giacomini and Rossi (2006) for the instability of its forecasting performance after the “Great Moderation”. For inflation, we use a Phillips curve-based model (in short: PC), see, e.g., Stock and Watson (1999). Here, x_t is the unemployment rate. By using the unemployment rate and an intercept rather than the unemployment gap, this specification is in line with the assumption of a constant NAIRU. The forecasting performance of the model and its empirical instability are investigated in, e.g., Giacomini and Rossi (2009) and recently in Perron and Yamamoto (2019).

[11] The data files are located at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/data-files/error-statistics. Appendix J presents some results indicating robustness of our findings when investigating unemployment and housing starts, which are also available from the SPF.

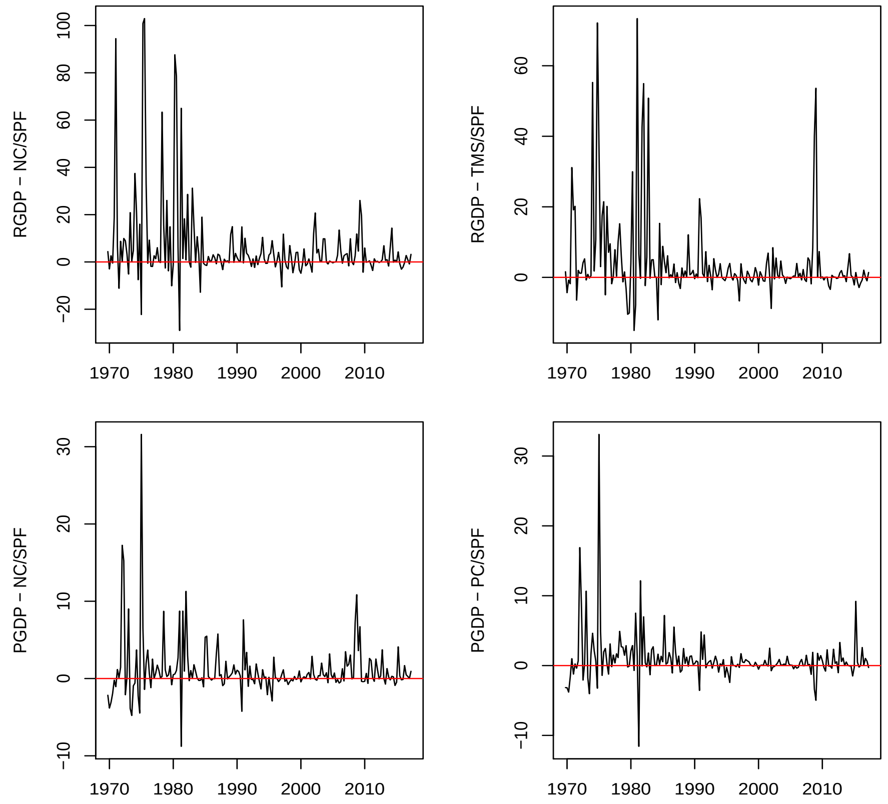


Figure 1: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve). Nowcasts are evaluated against the first release for mean squared error loss.

Real-time data from the Federal Reserve Bank of Philadelphia[12] is used to construct rolling window and recursive forecasts with $R = 60$. Interest rate data is taken from the updated data set of Welch and Goyal (2008).[13] In the following, we present evaluation results for the first release and rolling window estimation and discuss differences and similarities for the final release and recursive estimation towards the end of this section.

Figure 1 displays representative mean squared error loss differentials for $h = 0$ for the full sample, which covers 191 quarterly observations from 1969Q4 to 2017Q2.[14] The series reveal that (i) loss differentials are mostly, but not always, positive, indicating advantages of SPF forecasts, (ii) there is potentially some time-variation in the mean, (iii) there are striking volatility changes and (iv) there is some mild to intermediate autocorrelation. Appendices H and I contain further Figures

[12] The data files are located at https://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/data-files.

[13] See Amit Goyal's website http://www.hec.unil.ch/agoyal/. In the notation of Welch and Goyal (2008, p. 1459), the ten-year long-term government bond yield and the three-month Treasury bill secondary market rate are labeled as "lty" and "tbl", respectively.

[14] Some series contain a few missing values. Details on imputation are provided in Appendix G. As there are relatively many missing values in the first year of the survey, we decided to start in 1969Q4.

Table 1: Summary statistics for output growth (RGDP) and GDP deflator inflation (PGDP) using the first data release. RelLoss denotes the relative root mean squared error loss of the competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve); SD(·) labels the standard deviation of the loss differentials in the subsample I (1969-1984), II (1985-2006) or III (2007-2017). AC(1) denotes the empirical first-order autocorrelation coefficient of the loss differential series.

Statistic		RelLoss	SD(I)	SD(II)	SD(III)	AC(1)
Sample		1969-2017	1969-1984	1985-2006	2007-2017	1969-2017
RGDP - NC/SPF	$h = 0$	1.69	28.67	4.95	6.02	0.24
	$h = 1$	1.51	60.61	5.49	14.02	0.14
	$h = 4$	1.40	55.26	8.17	15.76	0.44
RGDP - TMS/SPF	$h = 0$	1.52	19.33	4.10	10.39	0.21
	$h = 1$	1.16	16.49	5.42	8.21	0.22
	$h = 4$	1.06	20.27	3.99	1.99	0.04
PGDP - NC/SPF	$h = 0$	1.38	5.88	1.68	2.41	0.08
	$h = 1$	1.23	9.82	2.01	1.91	0.26
	$h = 4$	1.12	16.62	2.33	2.57	0.29
PGDP - PC/SPF	$h = 0$	1.32	5.75	1.43	1.99	-0.02
	$h = 1$	1.26	11.00	1.65	1.83	0.25
	$h = 4$	1.29	22.55	2.41	2.58	0.41

33-37 (49-51) for other horizons and releases with similar patterns.

Table 1 provides summary statistics. We report root mean squared error ratios of competing forecasts relative to the SPF, such that values > 1 indicate a better performance of the SPF. In all cases, the SPF appears to outperform its competitors. However, there is some notable heterogeneity. The SPF is particularly successful at nowcasting (most strongly so for output). The advantages typically shrink with an increasing forecast horizon. However, the term spread model (TMS) is a strong competitor at $h = 4$, while Phillips curve-based (PC) forecasts are less competitive.

Unconditional standard deviations for the subsamples I (1969Q4-1984Q4, 61 observations), II (1985Q1-2006Q4, 88 observations) and III (2007Q1-2017Q2, 42 observations) indicate strong overall changes in volatility. This underlines the need for suitable inferential procedures. Structural changes associated with the “Great Moderation” are strongest for real GDP growth (with many break factors being even smaller than 1/5). For output, volatility of loss differentials increased a bit during the “Great Financial Crisis” (relative to the “Great Moderation”), while it stays fairly constant for inflation. Finally, the empirical first-order autocorrelation coefficient indicates a mild to intermediate degree of serial correlation in the loss differentials.

Table 2: Test decisions for the full-sample $\mathcal{T}^{DM}$-statistic for equal predictive ability of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve) - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the first data release. Evaluation sample runs from 1969Q4 to 2017Q2.

	RGDP - NC/SPF						RGDP - TMS/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	***	***	***	***	***	**	**	*	*
0.1	***	***	***	***	***	**	***	***	**	*	*	*
0.2	***	**	***	**	***	**	***	**	**	**	**	*
0.3	***	*	***	*	**	*	**	**	**	**	**	*
0.4	***	*	***	*	**	*	**	**	**	**	**	
0.5	**	*	***	*	**	*	**	*	**	**	**	
0.6	**	*	***	*	**	*	**	*	**	**	**	
0.7	**	*	***	*	*		**	*	**	**	*	
0.8	**	*	***	*	*		**	*	**	**	*	
0.9	**	*	***	*	*		**	*	**	**	*	
1	**	*	***	*	*		**	*	**	**	*	

	PGDP - NC/SPF						PGDP - PC/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	**			***	***	***	***	***	**
0.1	***	***	***	***	**	*	***	***	***	***	***	*
0.2	***	***	***	***	**	*	***	**	***	**	***	*
0.3	***	**	***	***	**	*	***	**	***	**	***	*
0.4	***	**	***	**	**	*	**	*	***	**	***	*
0.5	***	**	***	**	**		**	*	***	**	***	
0.6	***	**	***	**	**		**	*	***	**	***	
0.7	***	**	***	**	**		**	*	***	*	***	
0.8	***	**	***	**	**		**	*	***	**	***	
0.9	***	**	***	**	**		**	*	***	**	***	
1	***	**	***	**	**		**	*	***	**	***	

3.2 Tests for equal predictive ability and time-variation

For all statistics $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$, we consider $b \in \{0, 0.1, \ldots, 1\}$ for the fixed-b bandwidth parameter. We thus include a classic Newey-West type statistic ($b = 0$, see also App. D, fn. 24) and also the fixed-b versions proposed by Choi and Kiefer (2010). We focus on the Bartlett kernel (i.e., $k(x) = 1 - |x|$ for $|x| < 1$ and $k(x) = 0$ otherwise) due to its higher power relative to the Quadratic Spectral kernel, where both have similar size (cf. Appendix D). Test decisions and their strengths based on asymptotic, non-robust ("asy") and wild bootstrap ("bs") critical values are compared.

No-change forecasts do not involve parameter estimation while model-based forecasts generally do. For the SPF, the estimation error is not available and therefore, no correction of estimation error is applied, see the discussion in Giacomini and Rossi (2010) and Rossi and Sekhposyan (2016). Therefore, we employ the bootstrap algorithm given in Algorithm 1 with the additional restrictions from Remarks 2 and 9 using $M = 5,000$ replications, see also Appendix A for further details.

First, we test for equal predictive ability using the full-sample statistic $\mathcal{T}^{DM}$. Table 2 reports rejections at significance levels of one, five and ten percent. These are labeled as '***', '**' and '*' to ease the presentation of the many results and to conserve space by not reporting six different critical values for each statistic. We consider one-sided tests against the alternative that the SPF outperforms the benchmark.

Starting with output growth (RGDP) and no-change (NC) forecasts, the bootstrap version (subscript 'bs') rejects equal predictive ability across the full sample in all cases—at least at the nominal ten percent level, but mostly at the five percent level or lower. This finding holds for all horizons h and all values of the bandwidth-parameter b. It thus clearly suggests that the SPF significantly outperforms its competitors over the full sample. On the contrary, asymptotic critical values produce far weaker and fewer rejections. Results for the term spread model (TMS) are quite similar.

For GDP deflator inflation (PGDP), bootstrap inference leads to rejections at the one percent level in all cases for the shortest horizons $h = 0$ and $h = 1$. Relying on asymptotic critical values mainly produces rejections at the five percent level. We find a clear difference in test decisions for one-year ahead forecasts ($h = 4$): while the bootstrap detects significant differences, asymptotic inference hardly indicates any significant deviation from equal predictive ability. The differences between the outcomes for testing the superiority of the SPF over no-change or Phillips-curve based model forecasts are quite small.

In sum, the volatility-robust full sample results convincingly indicate the usefulness of the SPF for both variables, especially at short horizons. We next consider tests suitable for detecting time-variation in the relative forecast performance. To this end, we proceed in two steps. First, we apply the $\mathcal{T}^{F}$ (with $\nu = 0.3$ as suggested in Giacomini and Rossi, 2010), $\mathcal{T}^{Q}$ and $\mathcal{T}^{C}$ statistics presented in Section 2 as two-sided versions to test for time-variation in both directions and to ensure that we do not overlook potential periods in which the SPF is outperformed by the benchmarks. It may occur that the SPF is outperformed in some periods and that this feature is reversed in another part of the sample. Second, we investigate the time-varying nature of relative predictive ability of the SPF further by studying the time-varying components of the fluctuation and the CUSUM statistic and consider signed versions of the aforementioned test statistics with one-sided (in favor of the SPF) critical values, see Remark 3. The time-varying components are in particular the (i) rolling standardized mean squared error difference and (ii) scaled partial sum of the loss differential

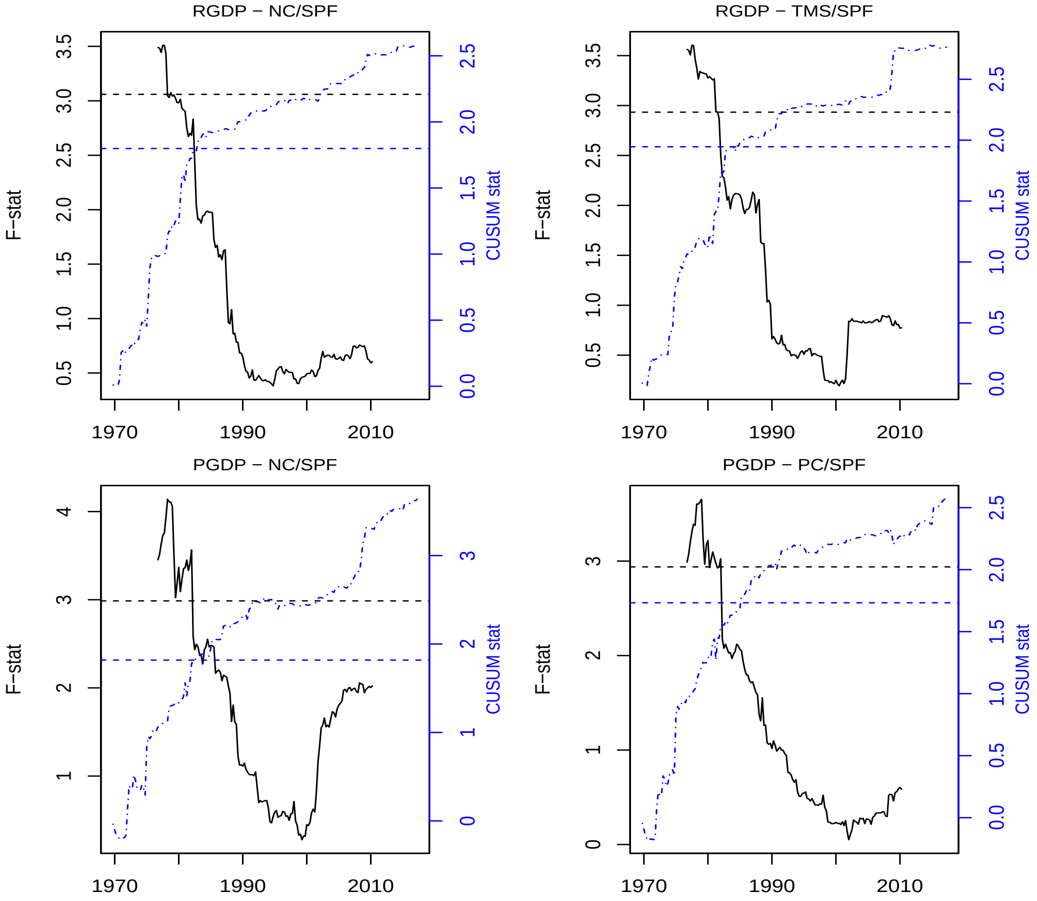


Figure 2: The plots show the signed time-varying components of the fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5 and Remark 3. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. Nowcasts are evaluated against the first release; $b = 0.2$, $\nu = 0.3$.

Table 3: Test decisions for the time-variation $\mathcal{T}^{\{Q,C,F\}}$-statistics for time-variation in the predictive ability of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve) - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the first data release. Evaluation sample runs from 1969Q4 to 2017Q2.

RGDP - NC/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	***	***	***	***	**	***	***	**	***	***	**	***
0.1	***	**	***	***	***	***	***	**	***	***	***	***	**	*	***	**	**	**
0.2	***	*	***	**	**	**	***	*	***	**	*	*	**		**	**		*
0.3	**		**	**	*		**		**	*			*		**	*		
0.4	**		**	*			**		**	*					**	*		
0.5	*		**	*			*		**	*					*			
0.6	*		**				*		**						*			
0.7	*		**				*		**						*			
0.8			**				*		**						*			
0.9			**				*		**						*			
1			**				*		*									

RGDP - TMS/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	**	**	*	**						*		*
0.1	***	**	***	***	***	***				*						*		**
0.2	**	*	***	**	**	**	*		*	*			**		***	**	**	**
0.3	*		**	**	*		**	*	**	**			*		**	*	*	*
0.4	*		**	*			**	*	**	**					**			
0.5			*	*			*	*	**	**					**			
0.6			*	*			*		*	*					*			
0.7			*	*			**		**	**					*			
0.8			*				*		**	*					*			
0.9			*				*		**	*					*			
1			*				*		**	*					*			

Table 4: continued from Table 3.

PGDP - NC/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	**	*	*	**		*						
0.1	***	***	***	***	***	***	***	***	***	***	***	***						
0.2	***	**	***	***	***	***	***	**	***	***	***	***						
0.3	***	**	***	**	**	**	***	**	***	***	***	***						
0.4	***	**	***	**	**	*	***	**	***	**	**	**			*			
0.5	***	**	***	**	**	*	***	**	***	**	**	**			*			
0.6	***	**	***	**	**	*	***	**	***	**	**	**			*			
0.7	***	**	***	**	**	*	***	*	***	**	**	**			**			
0.8	**	**	***	**	*	*	***	*	***	**	**	**			*			
0.9	***	**	***	**	**	*	***	*	***	**	**	**			*			
1	***	**	***	**	**	*	***	*	***	**	**	**			*			

PGDP - PC/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	***	**	**	***	**	**	**		**	**		
0.1	***	**	***	***	***	***	***	**	***	***	**	***	**		**	*		
0.2	**	*	***	**	**	**	***	**	***	**	**	**	*		**	*		
0.3	**		**	**	*		***	*	***	**	*	**			**	*		
0.4	*		**	*			**		***	**					**			
0.5	*		**	*			**		***	*					**			
0.6	*		**				**		***	*					**			
0.7	*		**				**		***	*					**			
0.8			**				**		***	*					**			
0.9	*		**				**		***	*					**			
1	*		**				**		***	*					**			

to identify different episodes of relative predictability, if present.

Tables 3 and 4 report results. Once more, bootstrapped versions of the test statistics provide stronger rejections than their asymptotic counterparts. Since both $\mathcal{T}^{DM}$ (aiming at testing against a constant alternative) and $\mathcal{T}^{x}$, $x \in \{F, Q, C\}$ (tests allowing for time-varying alternatives) yield quite similar rejections overall, the current results are not clear-cut as to the nature of the alternative.

Figure 2 reveals a sizable and significant deterioration in nowcast predictability in the early 1980s associated with the "Great Moderation". This breakdown is significant and permanent, while the mild recoveries observed for inflation (versus no-change forecasts only) in the early 2000s are too weak for a rejection. For output growth, the results suggest that there is no comeback in relative predictive ability of the SPF. Interestingly, relative forecast performance did not change a lot during the "Great Financial Crisis" even though volatility changed somewhat, but to a much lesser extent when compared to the "Great Moderation". Appendix H show that these results also hold for other horizons, see Figures 38-39. Figures 43-45 show the unscaled rolling window mean squared error difference between the SPF and its competitors. They generally support the previous interpretation and reveal that, at least, no-change forecasts never significantly outperform the SPF. The CUSUM statistic indicates a breakdown in relative forecast performance as it also turns significant in the early 1980s, implying that the accumulated changes are large enough for a rejection.[15]

Our interpretation is that the full sample results are mainly driven by the first part of the sample (until the mid-1980s) in which the SPF clearly performed better. As the statistics for time-variation further indicate clearly and robustly, the advantages in relative predictability largely disappear in the mid-1980s. Most of the evidence for time-variation, however, would not have been detected by a traditional analysis using asymptotic critical values.

Our results are fairly robust with respect to the vintage (first and final release) and the employed estimation scheme (rolling and recursive). Starting with the descriptive statistics reported in Tables 8 and 12 (see Appendices H and I), we observe very similar patterns to the baseline case in Table 1. The loss differentials in Figures 33-37 and 49-51 generally reveal strong heteroskedasticity. Regarding the full sample results (Tables 9 and 13), our main conclusions continue to hold. A notable difference is the case of real output growth when forecasts are evaluated against the final rather than the first release. Here, we find no more evidence for the superiority of the SPF over the term

[15] Its behavior at the beginning and end of the sample provides additional information which $\mathcal{T}^{F}$ cannot provide due to trimming. Before 1976, there are signs for time-variation in all series. GDP deflator inflation and output growth apparently exhibit some further time-variation after 2010.

spread model, except when looking at nowcasts. For inflation, on the contrary, results are quite robust throughout various settings. These findings are not affected by the estimation scheme. When looking are tests for time-variation, we obtain very similar conclusions, see Tables 10-11 and 14-15. Figures 46-48 show rolling averages of loss differentials[16] (analogous to Figure 1); see Figures 40-42 for the components of the statistics designed to detect time-variation (analogous to Figure 2).[17] In nearly all cases, we find the same pattern of advantages for the SPF in the early part of the sample (prior to the "Great Moderation") with a significant decline in the mid-eighties and limited recovery in the 2000s (if at all). An exception (for one- and four-quarter ahead forecasts) is the recursively estimated term spread model for which the relative SPF performance improves since the 2000s.

3.3 Discussion of our results in light of the related literature

We now provide a comparison of our findings with those of previous studies on the performance of the SPF. Most of these use the Diebold and Mariano (1995) test for differences in mean squared error. One strand of the literature deals with the accuracy of the SPF in general, while a second smaller one focusses on the decline of predictability in connection to the "Great Moderation". A comparison is generally complicated by the fact that studies obviously use different variables (and definitions), benchmarks, vintages, horizons, samples etc. However, two articles, viz. D'Agostino et al. (2006) and Coroneo and Iacone (2020), are particularly close to the scope of our work.

There is some consensus that the SPF provides accurate forecasts, especially nowcasts, for real output growth and inflation. Zarnowitz and Braun (1993) and Croushore (1993) (see also references therein) provide early evidence on the good performance of SPF forecasts for real GDP and inflation. Ang et al. (2007) find that surveys (including the SPF) forecast inflation better than macro variables, time series models (including no-change forecasts as advocated by Atkeson and Ohanian, 2001) and asset markets. They also find that when allowing for time-variation, the SPF dominates throughout the whole sample. Croushore (2010) finds confirmatory evidence using real-time data.

The advantages of SPF nowcasts has been documented in several influential studies, e.g., Giannone et al. (2008). Liebermann (2014) considers real-time nowcasting for output growth and compares professional forecasters and a dynamic factor model to simple autoregressive and no-change forecasts. The author finds that gains in forecasting accuracy are pronounced for $h = 0$ and decrease

[16] See Figures 55-57 for the case of recursively estimated models.

[17] See Figures 52-54 for the case of recursively estimated models.

in h. For a sample from 1985Q1 to 2007Q4, Stark (2010) similarly finds that the accuracy of the SPF declines significantly for $h > 1$, and that the SPF outperforms no-change forecasts.

We now turn to the discussion of D'Agostino et al. (2006) and Coroneo and Iacone (2020). Both use a naive benchmark (without estimation) under mean squared error loss and deal with time-variation by running tests on subsamples. In contrast to our tests, theirs are not robust to time-varying volatility and do not exploit the full sample to formally and endogenously test for time-variation.

Coroneo and Iacone (2020) propose a Diebold and Mariano (1995) statistic with fixed-m asymptotics (cf. the introduction). Their full-sample test has good size under homoskedasticity even in samples of only 40 observations, while tests using standard small-b-type asymptotics are oversized. Another advantage is the ensured positivity of the estimated long-run variance which is particularly important in small samples and with relatively long forecast horizons, see e.g. Harvey et al. (2017). In addition, Coroneo and Iacone (2020) consider a stationary block-bootstrap version of the test and find it to yield better size than standard asymptotics, again under homoskedasticity, and to be equally powerful as the fixed-m approach. In a sample ranging from 1987Q1 to 2016Q4, the SPF significantly outperforms a naive random walk in some cases for real output growth and inflation (as well as unemployment and interest rates). For output growth and inflation, there is evidence against the null at all horizons except three-quarters ahead. Generally, the evidence is stronger for shorter horizons.

In a subsample analysis with three blocks of ten years of data, the authors investigate time-variation and find: (i) for output growth, the SPF provides constantly superior nowcasts in all three subsamples, while the results for other horizons and subsamples are mixed—overall, the evidence is declining over the subsamples and for horizons beyond one-quarter; (ii) for inflation, relative advantages of the SPF are mainly observed for their last subsample period from 2007 to 2016 at all horizons (except three-quarters ahead). Thus, our findings only partly corroborate those of Coroneo and Iacone (2020, Tables 1 and 2) for these two variables. Unlike Stark (2010) and Coroneo and Iacone (2020), we do not find that the SPF easily outperforms naive output and inflation forecasts after the "Great Moderation". In order to further investigate whether the use of different testing environments may serve as an explanation for these differences, we provide an additional analysis reported in Appendix K. First, we run the $\mathcal{T}^{DM}$-statistic on each of the three subsamples (for SPF vs. no-change nowcasts and one-quarter and one-year ahead forecasts). The different tests mostly agree and give the same answers. Such an outcome is in line with the theory in Section 2 since the volatility varies much more across the individual subsamples rather than within. Second, we run our $\mathcal{T}^x, x \in \{F, Q\}$-tests on their subsample to identify periods of instability in relative forecasting performance and thereby, we are able to further compare the test results in light of the applied testing environments. Actually, we find differences as the results do not match very closely. This leads us to conclude that the observed differences in our main analysis may indeed be attributed to the different tests in use. As a by-product, we further provide evidence for instability within the subsamples studied in Coroneo and Iacone (2020) and thus recommend the usage of fluctuation and related tests in general.

D'Agostino et al. (2006) find a significant decline in relative predictive accuracy of the SPF for inflation and output growth for $h = 1$ to $h = 4$. Their full-sample (1975Q1 to 1999Q4) results indicate that the advantages of the SPF appear to be driven by the period prior to 1985 in which the SPF outperformed the naive benchmark, with no significant advantage thereafter. This points strongly to instabilities in the relative forecast performance. Our findings corroborate their results and sharpen them in showing that this phenomenon also holds for nowcasting. In addition, Campbell (2007), D'Agostino and Whelan (2008) and Gamber and Smith (2009) find, through analyses of various subsamples and consistent with our results, declining predictability of the SPF after the "Great Moderation" for output growth and inflation. Explanations regarding the causes of the forecast breakdown differ across these studies and remain an open issue.

By applying robust tests to a fairly long sample of more than 40 years, we obtain results which support several previous findings. Among these are (i) the advantages of the SPF for shortest horizons, but smaller advantages for one-year ahead forecasts; (ii) a significant decline in relative predictability during the 1980s; (iii) the robustness of the relative performance of the SPF to data revisions. Our results yield the following new insights: (i) advantages of the SPF forecasts are minimal in the 1990s, with weak signs of recoveries for GDP deflator inflation later on; (ii) relative forecast performance did not change during the "Great Financial Crisis", even though volatility increased (although relatively less than during the "Great Moderation") and (iii) the time-variation in the relative performance of the SPF is robust to the evaluation against simple no-change forecasts and dynamic models based on the term spread or the Phillips curve.

The observed recoveries possibly turn into a significant comeback of SPF forecasts in the future. In this case, the exact timing would very likely be unknown (Inoue and Rossi, 2005), rendering a subsample analysis inappropriate. In general, the ad-hoc choice of break points may easily lead to biases. Moreover, it is not always possible to invoke economic reasons like the well-studied "Great Moderation". In contrast, the methods proposed here are suitable for data containing possibly multiple unknown breakpoints in forecast performance alongside changes in volatility.

4 Concluding remarks

This paper proposes wild bootstrap tests for equal predictive ability that can be applied when volatility and relative forecast performance may be time-varying, and proves their validity. Both features are present in many macroeconomic and financial forecast comparisons. The tests account for, when needed, rolling and recursive estimation of parameters of forecast models. The considered tests are either full sample tests (Diebold and Mariano, 1995) or CUSUM, Cramér-von Mises and fluctuation statistics when testing for time-variation. All employ fixed-b asymptotics which deliver more accurately sized tests in finite-samples.

Our empirical application investigates the (time-varying) forecast performance of professional forecasters obtained from the SPF relative to simple no-change and model-based forecasts in real-time. The analysis suggests that ignoring time-varying variance seriously affects conclusions regarding the null of equal predictive ability. Traditional tests provide considerably weaker evidence against the null than the wild bootstrap versions. Tests allowing for time-variation indicate that the SPF had significant advantages until the mid-1980s, but not thereafter. Further research might address to what extent the time-varying relative forecast performance can be explained (e.g., Campbell, 2007). Another interesting avenue is to investigate the Fed's popular 'Teal Book' forecasts (e.g., Romer and Romer, 2000; D'Agostino and Whelan, 2008; Rossi and Sekhposyan, 2016).

References

Amado, C. and T. Teräsvirta (2013). Modelling volatility by variance decomposition. *Journal of Econometrics 175*(2), 142–153.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3), 817–858.

Ang, A., G. Bekaert, and M. Wei (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics 54*(4), 1163–1212.

Atkeson, A. and L. E. Ohanian (2001). Are phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review 25*, 2–11.

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

Campbell, S. D. (2007). Macroeconomic volatility, predictability, and uncertainty in the great moderation: evidence from the survey of professional forecasters. *Journal of Business & Economic Statistics 25*(2), 191–200.

Cavaliere, G. (2004). Unit root tests under time-varying variances. *Econometric Reviews 23*(3), 259–292.

Cavaliere, G., A. Rahbek, and A. M. R. Taylor (2010). Testing for co-integration in vector autoregressions with non-stationary volatility. *Journal of Econometrics 158*(1), 7–24.

Choi, H. S. and N. M. Kiefer (2010). Improving robust model selection tests for dynamic models. *The Econometrics Journal 13*(2), 177–204.

Clark, T. E. and F. Ravazzolo (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics 30*(4), 551–575.

Coroneo, L. and F. Iacone (2020). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. *Journal of Applied Econometrics 35*(4), 391–409.

Croushore, D. (1993). Introducing: the survey of professional forecasters. *Business Review-Federal Reserve Bank of Philadelphia 6*.

Croushore, D. (2010). An evaluation of inflation forecasts from surveys using real-time data. *The BE Journal of Macroeconomics 10*(1).

Croushore, D. and T. Stark (2001). A real-time data set for macroeconomists. *Journal of Econometrics 105*(1), 111–130.

D'Agostino, A., D. Giannone, and P. Surico (2006). (Un)Predictability and macroeconomic stability. *Working Paper Series 605, European Central Bank*.

D'Agostino, A. and K. Whelan (2008). Federal reserve information during the great moderation. *Journal of the European Economic Association 6*(2-3), 609–620.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.

Demetrescu, M., C. Hanck, and R. Kruse (2019). Robust fixed-b inference in the presence of time-varying volatility. *Mimeo*.

Demetrescu, M. and P. Sibbertsen (2016). Inference on the long-memory properties of time series with non-stationary volatility. *Economics Letters 144*, 80–84.

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

Estrella, A. and G. A. Hardouvelis (1991). The term structure as a predictor of real economic activity. *The Journal of Finance 46*(2), 555–576.

Faust, J. and J. H. Wright (2013). Forecasting inflation. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Chapter 1, pp. 2–56. Elsevier.

Gamber, E. N. and J. K. Smith (2009). Are the fed's inflation forecasts still superior to the private sector's? *Journal of Macroeconomics 31*(2), 240–251.

Giacomini, R. and B. Rossi (2006). How stable is the forecasting performance of the yield curve for output growth? *Oxford Bulletin of Economics and Statistics 68*(s1), 783–795.

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

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

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

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676.

Groen, J. J. J., R. Paap, and F. Ravazzolo (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics 31*(1), 29–44.

Guidolin, M. and A. Timmermann (2006). An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns. *Journal of Applied Econometrics 21*(1), 1–22.

Hansen, B. E. (2000). Testing for structural change in conditional models. *Journal of Econometrics 97*(1), 93–115.

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

Harvey, D. I., S. J. Leybourne, and E. J. Whitehouse (2017). Forecast evaluation tests and negative long-run variance estimates in small samples. *International Journal of Forecasting 33*(4), 833–847.

Honaker, J., G. King, and M. Blackwell (2011). Amelia II: A program for missing data. *Journal of Statistical Software 45*(7), 1–47.

Horowitz, J. L. and N. Savin (2000). Empirically relevant critical values for hypothesis tests: A bootstrap approach. *Journal of Econometrics 95*(2), 375–389.

Inoue, A. and B. Rossi (2005). Recursive Predictability Tests for Real-Time Data. *Journal of Business & Economic Statistics 23*, 336–345.

Justiniano, A. and G. Primiceri (2008). The time-varying volatility of macroeconomic fluctuations. *American Economic Review 98*(3), 604–641.

Kiefer, N. M. and T. J. Vogelsang (2002a). Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica 70*(5), 2093–2095.

Kiefer, N. M. and T. J. Vogelsang (2002b). Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory 18*(6), 1350–1366.

Kiefer, N. M. and T. J. Vogelsang (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory 21*(6), 1130–1164.

Kiefer, N. M., T. J. Vogelsang, and H. Bunzel (2000). Simple robust testing of regression hypotheses. *Econometrica 68*(3), 695–714.

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

Liebermann, J. (2014). Real-time nowcasting of GDP: A factor model vs. professional forecasters. *Oxford Bulletin of Economics and Statistics 76*(6), 783–811.

Lindner, A. M. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p,q)–processes. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, and T. Mikosch (Eds.), *Handbook of financial time series*, pp. 43–69. Springer.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics 21*, 255–285.

Müller, U. K. (2014). HAC corrections for strongly autocorrelated time series. *Journal of Business & Economic Statistics 32*(3), 311–322.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Perron, P. and Y. Yamamoto (2019). Testing for changes in forecasting performance. *Journal of Business & Economic Statistics*, Advance online publication. https://doi.org/10.1080/07350015.2019.1641410.

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

Rapach, D. E. and J. K. Strauss (2008). Structural breaks and GARCH models of exchange rate volatility. *Journal of Applied Econometrics 23*(1), 65–90.

Romer, C. D. and D. H. Romer (2000). Federal reserve information and the behavior of interest rates. *American Economic Review 90*(3), 429–457.

Rossi, B. (2013). Advances in forecasting under instability. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Chapter 21, pp. 1203–1324. Elsevier.

Rossi, B. and T. Sekhposyan (2016). Forecast rationality tests in the presence of instabilities, with applications to federal reserve and survey forecasts. *Journal of Applied Econometrics 31*(3), 507–532.

Sensier, M. and D. van Dijk (2004). Testing for volatility changes in U.S. macroeconomic time series. *The Review of Economics and Statistics 86*(3), 833–839.

Smeekes, S. and J.-P. Urbain (2014). A multivariate invariance principle for modified wild bootstrap methods with an application to unit root testing. *Maastricht University GSBE Research Memoranda RM/14/008.*

Stark, T. (2010). Realistic evaluation of real-time forecasts in the survey of professional forecasters. *Federal Reserve Bank of Philadelphia, Research Department* (Special Report), 726–740.

Stock, J. and M. Watson (1999). Forecasting inflation. *Journal of Monetary Economics 44*(2), 293–335.

Stock, J. H. and M. W. Watson (2002). Has the business cycle changed and why? *NBER Macroeconomics Annual 17*(1), 159–218.

Stock, J. H. and M. W. Watson (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking 39*(S1), 3–33.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies 21*(4), 1455–1508.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*(5), 1067–1084.

Zarnowitz, V. and P. Braun (1993). Twenty-two years of the NBER-ASA quarterly economic outlook surveys: aspects and comparisons of forecasting performance. In *Business cycles, indicators and forecasting*, pp. 11–94. University of Chicago Press.

Appendices—Not for publication

A Bootstrap implementation for linear regression forecasts

Here, we work out the corresponding wild bootstrap algorithm for the simple, but important case of a regression-based prediction using two different sets of predictors, $\boldsymbol{x}_{1,t}$ and $\boldsymbol{x}_{2,t}$. Let us consider the following linear predictive models

$$z_{t+h} = \boldsymbol{\theta}_i' \boldsymbol{x}_{i,t} + \epsilon_{i,t}\,, \quad t = 1, \ldots, R+P-1, \qquad i = 1, 2, \tag{11}$$

which we estimate by OLS in an either recursive or rolling scheme. The theoretical forecasts for z_{t+h} are given by

$$f_{i,t} = \boldsymbol{\theta}_i' \boldsymbol{x}_{i,t}, \qquad i = 1, 2,$$

at each step $t = R, \ldots, R+P-1$. The row version gradient of $f_{i,t}$ is given by $\boldsymbol{x}_{i,t}'$. Note that $\boldsymbol{x}_{1,t}$ and $\boldsymbol{x}_{2,t}$ (and thus $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$) need not have the same dimensionality.

Then, at each t, the forecasts are generated as $\hat{f}_{i,t} = \hat{\boldsymbol{\theta}}_{i,t}' \boldsymbol{x}_{i,t}$ with $\hat{\boldsymbol{\theta}}_{i,t}$ computed recursively, $\hat{\boldsymbol{\theta}}_{i,t} = \hat{\boldsymbol{\theta}}_{i,t}^{rec}$, or in a rolling fashion, $\hat{\boldsymbol{\theta}}_{i,t} = \hat{\boldsymbol{\theta}}_{i,t}^{rol}$. We use a constant quadratic loss, $\mathcal{L}_t = \mathcal{L}(u_1, u_2) = (u_1 - u_2)^2$ with partial derivative $\partial\mathcal{L}/\partial u_2 = -2(u_1 - u_2)$. In the linear regression case, the estimation effect depends on

$$\mathbf{C}_{i,t} = \boldsymbol{x}_{i,t}\boldsymbol{x}_{i,t}' \quad \text{and} \quad \boldsymbol{a}_{i,t,\boldsymbol{\theta}_i} = \boldsymbol{x}_{i,t}\left(z_{t+h} - \boldsymbol{\theta}_i'\boldsymbol{x}_{i,t}\right) = \boldsymbol{x}_{i,t}\epsilon_{i,t}.$$

To account for the estimation effect, one needs to replicate the behavior of partial sums of $(\boldsymbol{a}_{i,t}, y_t)'$; to do so, we need estimates of these quantities since they are not observed directly. While $\hat{y}_t^{\mathrm{r}}$ is the natural estimator for y_t for both the recursive and the rolling cases, computing estimates $\hat{\boldsymbol{a}}_{i,t}^{\mathrm{r}}$ requires a set of residuals, say $\hat{\epsilon}_{i,t}^{\mathrm{r}}$.

For the recursive setup, estimate for each $t = 1, \ldots, R+P-1$ the LS regression

$$z_{j+h} = \hat{\boldsymbol{\theta}}_{i,t}^{rec,\prime} \boldsymbol{x}_{i,j} + \hat{e}_{i,j,t}^{rec}\,, \quad j = 1, \ldots, t, \tag{12}$$

where the additional index t in equation (12) indicates the dependence of the estimates on the time point at which estimation is conducted. Moreover, residuals are denoted by $\hat{e}_{i,j,t}^{rec}$ to emphasize that a full set of residuals is computed at each time t in a recursive manner (and we do not have one single set of residuals which we could call $\hat{\epsilon}_{i,t}$). Then, we use

$$\hat{\epsilon}_{i,t}^{rec} = \hat{e}_{i,t,t}^{rec} \tag{13}$$

for all $t = 1, \ldots, R+P-1$. That is, the tth entry, $t = 1, \ldots, R+P-1$, into the residual vector used in the bootstrap algorithm (see below) is computed as the last element of the residual vector resulting from the regression (12), which uses the first t observations.

We employ the following bootstrap algorithm for the recursive case:

Algorithm 2

1. Compute $\hat{y}_t^{rec}$ from (2) (recall that $\hat{y}_t^{rec}$ for $1 \leq t \leq R-1$, do not enter the test statistic and may be freely chosen)

2. For all $t = 1, \ldots, R+P-1$, compute $\mathbf{C}_{i,t} = \boldsymbol{x}_{i,t}\boldsymbol{x}_{i,t}'$ and $\hat{\boldsymbol{a}}_{i,t}^{rec} = \boldsymbol{x}_{i,t}\hat{\epsilon}_{i,t}^{rec}$ with $\hat{\epsilon}_{i,t}^{rec}$ from (13).

3. Generate r_t^* wild bootstrap draws, $t = 1, \ldots, R+P-1$.

4. Construct $\left(\boldsymbol{a}_{1,t}^{*,\prime}, \boldsymbol{a}_{2,t}^{*,\prime}, y_t^*\right)'$ as $\left(\hat{\boldsymbol{a}}_{1,t}^{rec,\prime}, \hat{\boldsymbol{a}}_{2,t}^{rec,\prime}, \hat{y}_t^{rec}\right)' r_t^*$ for $t = 1, \ldots, R+P-1$.

5. Compute for $t = R, \ldots, R+P-1$
$$\hat{\boldsymbol{\theta}}_{i,t}^{rec,*} = \left(\sum_{j=1}^{t} \mathbf{C}_{i,j}\right)^{-1} \sum_{j=1}^{t} \boldsymbol{a}_{i,j}^* + \hat{\boldsymbol{\theta}}_{i,R+P-1}^{rec}.$$

6. Compute for $t = R, \ldots, R+P-1$
$$\begin{aligned}\hat{y}_t^{rec,*} &= y_t^* - 2\left(z_{t+h} - \hat{\boldsymbol{\theta}}_{1,t}^{rec,*\prime}\boldsymbol{x}_{1,t}\right)\boldsymbol{x}_{1,t}'\left(\hat{\boldsymbol{\theta}}_{1,t}^{rec,*} - \hat{\boldsymbol{\theta}}_{1,R+P}^{rec}\right) \\ &\quad + 2\left(z_{t+h} - \hat{\boldsymbol{\theta}}_{2,t}^{rec,*\prime}\boldsymbol{x}_{2,t}\right)\boldsymbol{x}_{2,t}' \cdot \left(\hat{\boldsymbol{\theta}}_{2,t}^{rec,*} - \hat{\boldsymbol{\theta}}_{2,R+P}^{rec}\right).\end{aligned}$$

7. Compute the test statistics using the bootstrap sample $\hat{y}_t^{rec,*}$, $t = R, \ldots, R+P-1$.

8. Repeat the steps M times and obtain the desired quantile(s).

The wild bootstrap provides asymptotically pivotal inference if $\sup_{t=1,\ldots,R+P-1} \mathrm{E}\,\|\boldsymbol{x}_{i,t}\|^4 < \infty$, which suffices to verify Assumption 5.(i) in the linear case.

The procedure is similar for rolling window estimation. For each $t = R, \ldots, R+P-1$,
$$z_{j+h} = \hat{\boldsymbol{\theta}}_{i,t}^{rol,\prime}\boldsymbol{x}_{i,j} + \hat{\epsilon}_{i,j,t}^{rol}, \quad j = t-R+1, \ldots, t, \tag{14}$$
At each time t, the forecasts are generated as $\hat{f}_{i,t} = \hat{\boldsymbol{\theta}}_{i,t}^{rol,\prime}\boldsymbol{x}_{i,t}$.

The bootstrap algorithm for the rolling windows case is very similar, but takes into account that we only resort to estimates from the current window at each t. The biggest change is how we get the residuals $\hat{\epsilon}_{i,t}^{rol}$ entering $\hat{\boldsymbol{a}}_{i,t}^{rol}$ (the rolling version estimate of $\boldsymbol{a}_{i,t}$). Again, we have multiple variants to choose $\hat{\epsilon}_{i,t}^{rol}$, given the multitude of computed residuals $\hat{\epsilon}_{i,j,t}^{rol}$. The natural choice is for the rolling window scheme to take
$$\hat{\epsilon}_{i,t}^{rol} = \begin{cases} \hat{\epsilon}_{i,t,R}^{rol} & t = 1, \ldots, R \\ \hat{\epsilon}_{i,t,t}^{rol} & t = R+1, \ldots, R+P-1. \end{cases} \tag{15}$$
That is, the last residual from each window is added to the series of residuals as the window rolls on and the first R are the residuals from the first window. The changes in the bootstrap algorithm are as follows. First, compute $\hat{\boldsymbol{a}}_{i,t}^{rol} = \boldsymbol{x}_{i,t}\hat{\epsilon}_{i,t}^{rol}$ for all $t = 1, \ldots, R+P$, with $\hat{\epsilon}_{i,t}^{rol}$ from (15). Second, generate the bootstrap sample analogously, and compute for $t = R, \ldots, R+P-1$
$$\hat{\boldsymbol{\theta}}_{i,t}^{rol,*} = \left(\sum_{j=t-R+1}^{t} \mathbf{C}_{i,j}\right)^{-1} \sum_{j=t-R+1}^{t} \boldsymbol{a}_{i,j}^* + \hat{\boldsymbol{\theta}}_{i,R+P-1}^{rol}.$$

Finally, compute

$$\begin{aligned}\hat{y}_t^{rol,*} &= y_t^* - 2\Big(z_{t+h} - \Big(\hat{\boldsymbol{\theta}}_{1,t}^{rol,*}\Big)'\boldsymbol{x}_{1,t}\Big)\boldsymbol{x}_{1,t}'(\hat{\boldsymbol{\theta}}_{1,t}^{rol,*} - \hat{\boldsymbol{\theta}}_{1,t}^{rol}) \\ &\quad + 2\Big(z_{t+h} - \Big(\hat{\boldsymbol{\theta}}_{2,t}^{rol,*}\Big)'\boldsymbol{x}_{2,t}\Big)\boldsymbol{x}_{2,t}' \cdot (\hat{\boldsymbol{\theta}}_{2,t}^{rol,*} - \hat{\boldsymbol{\theta}}_{2,t}^{rol}) \qquad \text{for } t = R, \ldots, R+P-1\end{aligned}$$

and proceed as before.

B Proofs

Proof of Lemma 2

Consider $\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}}$ for either $\mathrm{r} = rec$ or $\mathrm{r} = rol$. Then, given the smoothness of the loss function and the forecast functions, there exist $\tilde{\boldsymbol{\theta}}_{i,t}$ between $\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}}$ and $\boldsymbol{\theta}_i$ such that

$$\begin{aligned}\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[sP]-1}\hat{y}_t^{\mathrm{r}} &= \frac{1}{\sqrt{P}}\sum_{t=R}^{R+[sP]-1} y_t + \frac{1}{P}\sum_{t=R}^{R+[sP]-1}\boldsymbol{d}_1'(f_{1,t},\boldsymbol{\theta}_1)\cdot\sqrt{P}\left(\hat{\boldsymbol{\theta}}_{1,t}^{\mathrm{r}} - \boldsymbol{\theta}_1\right) \\ &\quad -\frac{1}{P}\sum_{t=R}^{R+[sP]-1}\boldsymbol{d}_2'(f_{2,t},\boldsymbol{\theta}_2)\cdot\sqrt{P}\left(\hat{\boldsymbol{\theta}}_{2,t}^{\mathrm{r}} - \boldsymbol{\theta}_2\right) + Q_{s,P}^{\mathrm{r}}\end{aligned}$$

where, with $\tilde{f}_{i,t} = f_i\big(\boldsymbol{x}_{i,t},\tilde{\boldsymbol{\theta}}_{i,t}\big)$

$$Q_{s,P}^{\mathrm{r}} = \sum_{i=1}^{2}(-1)^i\,\frac{1}{P}\sum_{t=R}^{R+[sP]-1}\Big(\boldsymbol{d}_i'(\tilde{f}_{i,t},\tilde{\boldsymbol{\theta}}_{i,t}) - \boldsymbol{d}_i'(f_{i,t},\boldsymbol{\theta}_i)\Big)\sqrt{P}\left(\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} - \boldsymbol{\theta}_i\right),$$

such that, for $R \leq t \leq P+R-1$,

$$\left|Q_{s,P}^{\mathrm{r}}\right| \leq 2\sup_{i,t,\tilde{\boldsymbol{\theta}}_i}\left\|\boldsymbol{d}_i(\tilde{f}_{i,t},\tilde{\boldsymbol{\theta}}_{i,t}) - \boldsymbol{d}_i(f_{i,t},\boldsymbol{\theta}_i)\right\|\sup_{i,t}\sqrt{P}\left\|\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} - \boldsymbol{\theta}_i\right\|.$$

Furthermore, it follows from Assumption 1 that

$$\begin{aligned}\sqrt{R}\left(\hat{\boldsymbol{\theta}}_{i,[uR]}^{rol} - \boldsymbol{\theta}_i\right) &\Rightarrow \left((\mathbf{C}_i(u) - \mathbf{C}_i(u-1))'\,\mathbf{W}_{i,\boldsymbol{\theta}_i}\,(\mathbf{C}_i(u) - \mathbf{C}_i(u-1))\right)^{-1} \\ &\quad\ \times(\mathbf{C}_i(u) - \mathbf{C}_i(u-1))'\,\mathbf{W}_{i,\boldsymbol{\theta}_i}\,(\boldsymbol{A}_i(u) - \boldsymbol{A}_i(u-1)) \\ \sqrt{R}\left(\hat{\boldsymbol{\theta}}_{i,[uR]}^{rec} - \boldsymbol{\theta}_i\right) &\Rightarrow \left(\mathbf{C}_i'(u)\,\mathbf{W}_{i,\boldsymbol{\theta}_i}\mathbf{C}_i(u)\right)^{-1}\mathbf{C}_i'(u)\,\mathbf{W}_{i,\boldsymbol{\theta}_i}\boldsymbol{A}_i(u)\end{aligned}$$

hold on $[1, 1+\pi]$. Hence, with $P/R \to \pi > 0$ and $t > R$, $\sqrt{P}\|\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} - \boldsymbol{\theta}_i\|$ is uniformly bounded in probability, such that $\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} \in \Phi_P$, and therefore $\tilde{\boldsymbol{\theta}}_{i,t} \in \Phi_P$, for all t w.p.1, and Assumption 2 ensures that $\sup_{s\in[0,1]}\left|Q_{s,P}^{\mathrm{r}}\right| \xrightarrow{p} 0$. Since the limit vector processes $\boldsymbol{d}_i(\cdot)$ are Lipschitz-continuous and deterministic, the result follows with the continuous mapping theorem [CMT] and the change of variable $u = 1 + s\pi$.

Proof of Proposition 1

After using Equation (10) and the uniformity of the $o_p(1)$ term, the arguments in the proof of Theorem 2 in Kiefer and Vogelsang (2005) indicate that

$$\hat{\Omega} = -\frac{1}{P^2}\sum_{i=1}^{P-1}\sum_{j=1}^{P-1}\frac{P^2}{B^2}k''\left(\frac{i-j}{B}\right)\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+i-1}\left(\hat{y}_t^{\mathrm{r}}-\overline{\hat{y}^{\mathrm{r}}}\right)\right)\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+j-1}\left(\hat{y}_t^{\mathrm{r}}-\overline{\hat{y}^{\mathrm{r}}}\right)\right)+o_p(1)$$

for kernels with smooth derivatives, where $\bar{\cdot}$ denotes the sample average $\bar{w} = P^{-1}\sum_{t=R}^{R+P-1} w_t$ for any choice of w_t. We also have

$$\begin{aligned}\hat{\Omega} &= \frac{2}{bP}\sum_{i=1}^{P}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+i-1}\left(\hat{y}_t^{\mathrm{r}}-\overline{\hat{y}^{\mathrm{r}}}\right)\right)^2 \\ &\quad -\frac{2}{bP}\sum_{i=1}^{[(1-b)P]}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+i-1}\left(\hat{y}_t^{\mathrm{r}}-\overline{\hat{y}^{\mathrm{r}}}\right)\right)\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+i+[bP]-1}\left(\hat{y}_t^{\mathrm{r}}-\overline{\hat{y}^{\mathrm{r}}}\right)\right)+o_p(1)\end{aligned}$$

for the Bartlett kernel. Lemma 2 then implies, for $\mathrm{r} = \{rec, rol\}$,

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[sP]-1}\left(\hat{y}_t^{\mathrm{r}}-\overline{\hat{y}^{\mathrm{r}}}\right) = \frac{1}{\sqrt{P}}\sum_{t=R}^{R+[sP]-1}\hat{y}_t^{\mathrm{r}} - \frac{[sP]}{P}\frac{1}{\sqrt{P}}\sum_{t=R}^{R+P-1}\hat{y}_t^{\mathrm{r}} \Rightarrow B^{\mathrm{r}}_{\mathbf{G},\pi}(s) - sB^{\mathrm{r}}_{\mathbf{G},\pi}(1);$$

the CMT then leads to the desired limiting null distributions.

Proof of Proposition 2

To establish the desired result, it suffices that the bootstrap statistics $\mathcal{T}^{x,*}$ converge weakly to the corresponding distributions of $\mathcal{T}^{x}$, $x \in \{DM, F, Q, C\}$.

To this end, we first show that, under the null and local alternatives of the form $\mu_t = R^{-1/2}\mu(t/R)$ for a piecewise Lipschitz function μ, we have on $[0, 1+\pi]$

$$\frac{1}{\sqrt{P}}\sum_{t=1}^{[uR]}\begin{pmatrix}\boldsymbol{a}^*_{1,t}\\ \boldsymbol{a}^*_{2,t}\\ y^*_t\end{pmatrix} \overset{p}{\Rightarrow} \sqrt{c}\begin{pmatrix}\boldsymbol{A}_1(u)\\ \boldsymbol{A}_2(u)\\ A_y(u)\end{pmatrix}$$

with "$\overset{p}{\Rightarrow}$" standing for weak convergence in probability. Write

$$\frac{1}{\sqrt{P}}\sum_{t=1}^{[uR]}\begin{pmatrix}\boldsymbol{a}^*_{1,t}\\ \boldsymbol{a}^*_{2,t}\\ y^*_t\end{pmatrix} = \frac{1}{\sqrt{P}}\sum_{t=1}^{[uR]}\begin{pmatrix}\hat{\boldsymbol{a}}^{\mathrm{r}}_{1,t}-\boldsymbol{a}_{1,t,\boldsymbol{\theta}_1}\\ \hat{\boldsymbol{a}}^{\mathrm{r}}_{2,t}-\boldsymbol{a}_{2,t,\boldsymbol{\theta}_2}\\ \hat{y}^{\mathrm{r}}_t - y_t + \mu_t\end{pmatrix} r^*_t + \frac{1}{\sqrt{P}}\sum_{t=1}^{[uR]}\begin{pmatrix}\boldsymbol{a}_{1,t,\boldsymbol{\theta}_1}\\ \boldsymbol{a}_{2,t,\boldsymbol{\theta}_2}\\ y_t - \mu_t\end{pmatrix} r^*_t \tag{16}$$

for $u \leq 1+\pi$. The condition $\sup_t \|\boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i)\| = O_p\left(P^{1/2-\gamma}\right)$ from Assumption 5 implies for $\gamma > 0$ that $\sup_t |\hat{y}^{\mathrm{r}}_t - y_t| \xrightarrow{p} 0$ – given the behavior of $\hat{\boldsymbol{\theta}}^{rol}_{i,R+P-1}$ and $\hat{\boldsymbol{\theta}}^{rec}_{i,R+P-1}$ from the proof of Lemma 2, such that

$$\sup_t \left\| \begin{pmatrix} \hat{\boldsymbol{a}}^{\mathrm{r}}_{i,t} - \boldsymbol{a}_{i,t,\boldsymbol{\theta}_i} \\ \hat{y}^{\mathrm{r}}_t - y_t + \mu_t \end{pmatrix} \right\| \xrightarrow{p} 0,$$

which implies in turn that, as required for the 1st summand of the r.h.s. of (16) to vanish,

$$\frac{1}{P} \sum_{t=R}^{R+[sP]-1} \left\| \begin{pmatrix} \hat{\boldsymbol{a}}^{\mathrm{r}}_{i,t} - \boldsymbol{a}_{i,t,\boldsymbol{\theta}_i} \\ \hat{y}^{\mathrm{r}}_t - y_t + \mu_t \end{pmatrix} r^*_t \right\|^2 \overset{p}{\Rightarrow} 0 \tag{17}$$

since $\sup_t |r^*_t| = o^* \left(P^{-\gamma}\right)$ for any $\gamma > 0$ whenever r^*_t has finite moments of any order, and the first summand on the r.h.s. of (16) vanishes uniformly in u. (Here, $o^*_p(1)$ stands for a quantity vanishing in probability w.r.t. the bootstrap measure.)

Let us now study the 2nd summand of the r.h.s. of (16). We first examine the case where the bootstrap multipliers r^*_t are standard normal. Let $\boldsymbol{S}^*_P(u)$ denote the normalized partial sums

$$\boldsymbol{S}^*_P(u) = \frac{1}{\sqrt{P}} \sum_{t=1}^{[uR]} \begin{pmatrix} \boldsymbol{a}_{1,t,\boldsymbol{\theta}_1} \\ \boldsymbol{a}_{2,t,\boldsymbol{\theta}_2} \\ y_t - \mu_t \end{pmatrix} r^*_t := \frac{1}{\sqrt{P}} \sum_{t=1}^{[uR]} \boldsymbol{\xi}^*_t = \frac{1}{\sqrt{P}} \sum_{t=1}^{[uR]} \boldsymbol{\xi}_t r^*_t,$$

which, conditional on the sample, is a Gaussian process with independent increments. Its covariance kernel, conditional on the data, is given by

$$\mathrm{Cov}^* \left(\boldsymbol{S}^*_P(s), \boldsymbol{S}^*_P(r)\right) = \frac{1}{P} \sum_{t=1}^{[\min\{s,r\}R]} \boldsymbol{\xi}_t \boldsymbol{\xi}'_t \, \mathrm{E}\left(\left(r^*_t\right)^2\right) = \frac{1}{P} \sum_{t=1}^{[\min\{s,r\}R]} \boldsymbol{\xi}_t \boldsymbol{\xi}'_t.$$

Note that, under Assumption 4, we obtain pointwise in s

$$\frac{1}{P} \sum_{j=1}^{[uR]} \boldsymbol{\xi}_t \boldsymbol{\xi}'_t \xrightarrow{p} \int_0^u \mathbf{G}(r) \, \mathrm{E}\left(\tilde{\boldsymbol{v}}_t \tilde{\boldsymbol{v}}'_t\right) \mathbf{G}'(r) \mathrm{d}r = c \int_0^u \mathbf{G}(r) \mathbf{G}'(r) \mathrm{d}r \tag{18}$$

via a Law of Large Numbers for strong mixing processes (see Davidson, 1994, Section 20.6).

Recall that the quadratic covariation process of the desired weak limit, $\sqrt{c} \int_0^u \mathbf{G}(r) \mathrm{d}\boldsymbol{W}(r)$, is given by $c \int_0^u \mathbf{G}(r) \mathbf{G}'(r) \mathrm{d}r$. Then, like in the proof of Lemma A.5 in Cavaliere et al. (2010), weak convergence in probability of the bootstrap partial sums to a Gaussian process with independent increments and quadratic covariation process $c \int_0^u \mathbf{G}(r) \mathbf{G}'(r) \mathrm{d}r$ follows from uniformity of the convergence in (18).

Uniformity is indeed given, since the increments of the limit $c \int_0^u \mathbf{G}(r) \mathbf{G}'(r) \mathrm{d}r$ are positive semidefinite by construction, so any quadratic form thereof would be a continuous, nondecreasing function, hence leading to uniform convergence of the corresponding quadratic forms of the l.h.s. of (18). Given such univariate uniform convergence of any quadratic form, it follows that convergence in probability in (18) must be uniform itself.

In the case where the bootstrap multipliers r^*_t are not standard normal but follow the Mammen distribution, say, $\boldsymbol{S}^*_P(s)$ is not Gaussian, but weak convergence to a Gaussian process holds conditional on the sample (see, e.g., Davidson, 1994, Corollary 29.14, with r^*_t being iid and having finite moments of any order). The result follows along the lines of the Gaussian argument above. Finally, $\sup_t \left\| \hat{\mathbf{C}}^{\mathrm{r}}_{i,t} - \mathbf{C}_{i,t,\boldsymbol{\theta}_i} \right\| \xrightarrow{p} 0$ implies for either rolling or recursive estimation that

$$\frac{1}{R} \sum_{t=1}^{[uR]} \hat{\mathbf{C}}^{\mathrm{r}}_{i,j} \Rightarrow \mathbf{C}_i\left(u\right)$$

such that, with $\mathbf{W}_{i,\cdot}$ continuous, we have

$$\begin{aligned}
\sqrt{R}\left(\hat{\boldsymbol{\theta}}^{rol,*}_{i,[uR]} - \hat{\boldsymbol{\theta}}^{rol}_{i,R+P}\right) &\overset{p}{\Rightarrow} \left(\left(\mathbf{C}_i(u) - \mathbf{C}_i(u-1)\right)' \mathbf{W}_{i,\boldsymbol{\theta}_i} \left(\mathbf{C}_i(u) - \mathbf{C}_i(u-1)\right)\right)^{-1} \\
&\quad \times \left(\mathbf{C}_i(u) - \mathbf{C}_i(u-1)\right)' \mathbf{W}_{i,\boldsymbol{\theta}_i} \left(\boldsymbol{A}_i(u) - \boldsymbol{A}_i(u-1)\right) \\
\sqrt{R}\left(\hat{\boldsymbol{\theta}}^{rec,*}_{i,[uR]} - \hat{\boldsymbol{\theta}}^{rec}_{i,R+P}\right) &\overset{p}{\Rightarrow} \left(\mathbf{C}_i'\left(u\right)\mathbf{W}_{i,\boldsymbol{\theta}_i}\mathbf{C}_i\left(u\right)\right)^{-1} \mathbf{C}_i'\left(u\right)\mathbf{W}_{i,\boldsymbol{\theta}_i}\boldsymbol{A}_i\left(u\right)
\end{aligned}$$

on $[1, 1+\pi]$. For either estimation scheme, given the behavior of $\hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,R+P-1}$ from the proof of Lemma 2, this implies that $\sup_{R\le t\le R+P-1} \left\|\hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{i,t} - \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,R+P-1}\right\| = O_p\left(R^{-1/2}\right)$, so both $\hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,R+P-1}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{i,t}$ belong w.p.1 to the set Φ_P. The arguments from the proof of Lemma 2 then apply, and, together with the change of variable $s = \left(u-1\right)/\pi$ for $1 \le u \le \pi$, the CMT implies weak convergence in probability of the partial sums of $\hat{y}^{\mathrm{r},*}_t$,

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[sP]-1} \hat{y}^{rol,*}_t \overset{p}{\Rightarrow} \sqrt{c}B^{rol}_{\mathbf{G},\pi}\left(s\right) \qquad \text{and} \qquad \frac{1}{\sqrt{P}} \sum_{t=R}^{R+[sP]-1} \hat{y}^{rec,*}_t \overset{p}{\Rightarrow} \sqrt{c}B^{rec}_{\mathbf{G},\pi}\left(s\right). \tag{19}$$

To complete the result, it is tedious, yet straightforward to obtain a representation of the bootstrap long-run covariance estimator parallelling to the one in the proof of Proposition 1,

$$\hat{\Omega}^* = -\frac{1}{P^2} \sum_{i=1}^{P-1} \sum_{j=1}^{P-1} \frac{P^2}{B^2} k''\left(\frac{i-j}{B}\right) \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+i-1} \left(\hat{y}^{\mathrm{r},*}_t - \overline{\hat{y}^{\mathrm{r},*}}\right)\right) \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+j-1} \left(\hat{y}^{\mathrm{r},*}_t - \overline{\hat{y}^{\mathrm{r},*}}\right)\right) + o^*_p(1)$$

for kernels with smooth derivatives, or

$$\begin{aligned}
\hat{\Omega}^* = \frac{2}{bP} \sum_{i=1}^{P} \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+i-1} \left(\hat{y}^{\mathrm{r},*}_t - \overline{\hat{y}^{\mathrm{r},*}}\right)\right)^2 \\
- \frac{2}{bP} \sum_{i=1}^{[(1-b)P]} \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+i-1} \left(\hat{y}^{\mathrm{r},*}_t - \overline{\hat{y}^{\mathrm{r},*}}\right)\right) \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+i+[bP]-1} \left(\hat{y}^{\mathrm{r},*}_t - \overline{\hat{y}^{\mathrm{r},*}}\right)\right) + o^*_p(1)
\end{aligned}$$

for the Bartlett kernel. Note that the scale factor $\sqrt{c}$ cancels out from all four bootstrap statistics $\mathcal{T}^{x,*}$, $x \in \{DM, F, Q, C\}$, and convergence in (19) together with the CMT then implies weak convergence in probability of $\mathcal{T}^{x,*}$ to the same distributions as $\mathcal{T}^{x}$ as required for the result.

C The multivariate case

This appendix sketches the modifications arising in our procedure when the analyst wishes to study several loss differentials.[18] The proofs are trivial generalizations of those presented in Appendix B and omitted.

A leading scenario of this type, following Giacomini and White (2006), arises in tests of equal conditional predictive ability.[19] Here, the observed loss differentials are leveraged with a vector $\boldsymbol{w}_t$ of K test functions (which are measurable w.r.t. the relevant information set; see Giacomini and White, 2006). The forecast losses are still given by $\mathcal{L}_t\big(z_{t+h}, \hat{f}_{i,t}\big) = \mathcal{L}_t\big(z_{t+h}, f_i\big(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}_{i,t}\big)\big)$, so one uses

$$\hat{\boldsymbol{y}}_t = \boldsymbol{w}_t\big(\mathcal{L}_t\big(z_{t+h}, \hat{f}_{1,t}\big) - \mathcal{L}_t\big(z_{t+h}, \hat{f}_{2,t}\big)\big), \qquad t = R, \ldots, R+P-1, \tag{20}$$

for testing. The null is correspondingly that of $\mathrm{E}\left(\boldsymbol{y}_t\right) = \mathbf{0}$, to be tested using the feasible $\hat{\boldsymbol{y}}_t$.

As is usual in such multivariate settings, we consider two-sided tests. Let $\boldsymbol{S}_a^b \equiv \sum_{t=a}^{b} \hat{\boldsymbol{y}}_t$ and $Q_a^b \equiv \big(\boldsymbol{S}_a^b\big)' \hat{\boldsymbol{\Omega}}^{-1} \boldsymbol{S}_a^b$, with covariance matrix estimator $\hat{\boldsymbol{\Omega}} = \sum_{j=-P+1}^{P-1} k\left(j/B\right) \hat{\boldsymbol{\Gamma}}_j$, where $\hat{\boldsymbol{\Gamma}}_{|j|} = P^{-1} \sum_{t=|j|+R}^{R+P-1} \left(\hat{\boldsymbol{y}}_t - \bar{\hat{\boldsymbol{y}}}\right) \left(\hat{\boldsymbol{y}}_{t-|j|} - \bar{\hat{\boldsymbol{y}}}\right)'$ and $\hat{\boldsymbol{\Gamma}}_{-|j|} = \hat{\boldsymbol{\Gamma}}'_{|j|}$. The Diebold and Mariano (1995) statistic can then be written as

$$\mathcal{T}_K^{DM} = \frac{1}{P} Q_R^{R+P-1}. \tag{21}$$

In a K-variate setup, the fluctuation test of Giacomini and Rossi (2010) requires to compute for each $t = R + [S/2], \ldots, P + R - [S/2]$ a moving-window based version of (21),

$$F_{t,S} = \frac{1}{S} Q_{t-[S/2]}^{t+[S/2]-1}$$

and

$$\mathcal{T}_K^{F} = \max_{t \in \{R+[S/2], \ldots, R+P-[S/2]\}} F_{t,S}, \quad S = \lfloor \nu P \rfloor \quad \text{with} \quad \nu \in (0,1)\,. \tag{22}$$

The CUSUM and Cramér-von Mises statistics can be written as

$$\mathcal{T}_K^{Q} = \max_{R \le t \le R+P-1} \sqrt{Q_R^t / P} \qquad \text{and} \qquad \mathcal{T}_K^{C} = \frac{1}{P^2} \sum_{t=R}^{R+P-1} Q_R^t. \tag{23}$$

The multivariate random fixed-b limits of the long-run covariance matrix become

$$\boldsymbol{\Lambda}_{k,b}\left(\boldsymbol{X}\right) \equiv \begin{cases} -\frac{1}{b^2} \int_0^1 \int_0^1 k''\left(\frac{r-s}{b}\right) \bar{\boldsymbol{X}}(r) \bar{\boldsymbol{X}}(s)' \,\mathrm{d}r \mathrm{d}s \\ \frac{1}{b}\left(2 \int_0^1 \bar{\boldsymbol{X}}(r) \bar{\boldsymbol{X}}(r)' \mathrm{d}r - \int_0^{1-b} \bar{\boldsymbol{X}}(r+b) \bar{\boldsymbol{X}}(r)' \,\mathrm{d}r - \int_0^{1-b} \bar{\boldsymbol{X}}(r) \bar{\boldsymbol{X}}(r+b)' \,\mathrm{d}r\right) \end{cases}$$

for kernels with smooth derivatives and the Bartlett kernel, respectively.

The moment conditions are taken to obey, jointly with $\boldsymbol{y}_t$, a "$N_1 + N_2 + K$-dimensional" version of Assumption 4, which then implies a multivariate analog of Lemma 2, where $\boldsymbol{A}_{\boldsymbol{y}}$ is K-variate.

[18] We focus on recursive estimation for brevity, all modifications to the rolling case are straightforward.

[19] It may also be conceivable to study several loss functions simultaneously, but we do not provide the details here.

Moreover, let the multivariate analog of $\boldsymbol{d}_i$ be given as

$$\mathbf{D}_i(f,\boldsymbol{t}) = \boldsymbol{w}_t \cdot \left.\frac{\partial \mathcal{L}_t}{\partial u_2}\right|_{\substack{u_1=z_{t+h}\\ u_2=f}} \left.\frac{\partial f_i}{\partial \boldsymbol{\theta}'}\right|_{\substack{\boldsymbol{x}_{i,t}\\ \boldsymbol{\theta}=\boldsymbol{t}}},$$

which is smooth in the same sense as $\boldsymbol{d}_i$ in Assumption 2. Assume analogously that, for $R, P \to \infty$ with $P/R \to \pi$,

$$\frac{1}{P}\sum_{t=R}^{R+[sP]-1} \mathbf{D}_i(f_{i,t},\boldsymbol{\theta}_i) \Rightarrow \mathbf{H}_i(s),$$

where $\mathbf{H}_i$ is deterministic and Lipschitz-continuous.

Lemma 3. *Let $\boldsymbol{\mathcal{A}}(s) \equiv \left(\boldsymbol{A}_{\boldsymbol{y}}\left(1+s\pi\right) - \boldsymbol{A}_{\boldsymbol{y}}(1)\right)/\sqrt{\pi}$. For* $\mathrm{r} \in \{rec, rol\}$, *it holds that*

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[sP]-1} \hat{y}_t^{\mathrm{r}} \Rightarrow \boldsymbol{\mathcal{A}}(s) + \sqrt{\pi}\sum_{i=1}^{2}(-1)^{i+1}\left(\int_0^s \boldsymbol{N}_i^{\mathrm{r}\prime}(r)(\mathbf{M}_i^{\mathrm{r}})^{-1}(r)\mathrm{d}\mathbf{H}_i'(r)\right)' \equiv B^{\mathrm{r}}_{\mathbf{G},\pi}(s),$$

on $[0,1]$, *where* $\mathbf{M}_i^{\mathrm{r}}(s)$ *and* $\boldsymbol{N}_i^{\mathrm{r}}(s)$ *are defined in Lemma 2.*

In the multivariate case, the statistics are based on quadratic forms. For some vector process $\boldsymbol{X}$ and (stochastic) matrix $\mathbf{T}$, (a.s.) invertible, define therefore the limiting functionals of the multivariate test statistics as

$$\mathcal{F}(\boldsymbol{X},\mathbf{T}) = \sup_{s\in[\nu/2;1-\nu/2]} \frac{1}{\nu}\left(\boldsymbol{X}\left(s+\frac{\nu}{2}\right) - \boldsymbol{X}\left(s-\frac{\nu}{2}\right)\right)'\mathbf{T}^{-1}\left(\boldsymbol{X}\left(s+\frac{\nu}{2}\right) - \boldsymbol{X}\left(s-\frac{\nu}{2}\right)\right),$$

$$\mathcal{Q}(\boldsymbol{X},\mathbf{T}) = \sup_{s\in[0,1]} \sqrt{\boldsymbol{X}'(s)\,\mathbf{T}^{-1}\boldsymbol{X}(s)} \qquad \text{and} \qquad \mathcal{C}(\boldsymbol{X},\mathbf{T}) = \int_0^1 \boldsymbol{X}'(s)\,\mathbf{T}^{-1}\boldsymbol{X}(s)\,\mathrm{d}s.$$

Proposition 3. *Under the assumptions of Lemma 3 and the null, for $x \in \{F, Q, C\}$ and $\mathcal{X} \in \{\mathcal{F}, \mathcal{Q}, \mathcal{C}\}$ we have for either* $\mathrm{r} \in \{rec, rol\}$ *that*

$$\mathcal{T}_K^{DM} \xrightarrow{d} \boldsymbol{B}^{\mathrm{r},\prime}_{\mathbf{G},\pi}(1)\,\boldsymbol{\Lambda}^{-1}_{k,b}(\boldsymbol{B}^{\mathrm{r}}_{\mathbf{G},\pi})\,\boldsymbol{B}^{\mathrm{r}}_{\mathbf{G},\pi}(1) \quad \text{and} \quad \mathcal{T}_K^{x} \Rightarrow \mathcal{X}\left(\boldsymbol{B}^{\mathrm{r}}_{\mathbf{G},\pi}, \boldsymbol{\Lambda}_{k,b}(\boldsymbol{B}^{\mathrm{r}}_{\mathbf{G},\pi})\right).$$

In the multivariate case, the bootstrap algorithm restoring pivotality only requires modification of the first step to account for multivariate $\hat{\boldsymbol{y}}_t$ and reads as follows:

Algorithm 3

1. Compute $\hat{\boldsymbol{y}}_t^{\mathrm{r}}$ from (20) (recall that, for $t = 1, \ldots, R-1$, $\hat{\boldsymbol{y}}_t^{\mathrm{r}}$ does not enter the statistic so it may be freely chosen); for $t = 1, \ldots, R+P-1$, compute $\hat{\mathbf{C}}^{\mathrm{r}}_{i,t}$, $\hat{\mathbf{W}}^{\mathrm{r}}_{i,t}$ and $\hat{\boldsymbol{a}}^{\mathrm{r}}_{i,t}$ as in Algorithm 1.
2. For $t = 1, \ldots, R+P-1$, draw multipliers r_t^* and construct $\left(\boldsymbol{a}^{*\prime}_{1,t}, \boldsymbol{a}^{*\prime}_{2,t}, \boldsymbol{y}^{*\prime}_t\right)'$ as $\left(\hat{\boldsymbol{a}}^{\mathrm{r},\prime}_{1,t}, \hat{\boldsymbol{a}}^{\mathrm{r},\prime}_{2,t}, \hat{\boldsymbol{y}}^{\mathrm{r},\prime}_t\right) r_t^*$.
3. Compute $\hat{\boldsymbol{\theta}}^{rol,*}_{i,t}$ or $\hat{\boldsymbol{\theta}}^{rec,*}_{i,t}$ for $t = R+1, \ldots, R+P-1$ like in the algorithm in the main text.
4. With $\hat{f}^{\mathrm{r},*}_{i,t} = f_i\left(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{i,t}\right)$, compute for either $\mathrm{r} = rec$ or $\mathrm{r} = rol$

$$\hat{\boldsymbol{y}}^{\mathrm{r},*}_t = \boldsymbol{y}^*_t + \mathbf{D}_1(\hat{f}^{\mathrm{r},*}_{1,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{1,t}) \cdot \left(\hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{1,t} - \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{1,R+P-1}\right) - \mathbf{D}_2(\hat{f}^{\mathrm{r},*}_{2,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{2,t}) \cdot \left(\hat{\boldsymbol{\theta}}^{\mathrm{r},*}_{2,t} - \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{2,R+P-1}\right)$$

for $t = R, \ldots, R+P-1$.

5. Compute the test statistics of interest using the bootstrap sample $\hat{\boldsymbol{y}}_t^{\mathrm{r},*}$, $t = R, \ldots, R+P-1$.
6. Repeat steps 2–5 M times and obtain the desired quantile(s).

D Numerical evidence

This section investigates the finite-sample properties of the different statistics, in view of the asymptotic arguments from Section 2.2. We consider both potential time-varying forecasting ability and estimation uncertainty (cf. Proposition 1). For concreteness, we shall investigate the simple and widely relevant case of regression-based prediction through competing univariate predictors. Algorithm 2 in Appendix A summarizes the corresponding bootstrap procedure for replicating the non-pivotal distributions from Proposition 1.

Our main DGP is as follows. We aim to predict an ARMA(1,1)-process $z_t = 0.4z_{t-1} + \epsilon_t + 0.3\epsilon_{t-1}$, $t = 1, \ldots, R + P$, through two competing AR(1)-processes $x_{i,t} = 0.5x_{i,t-1} + u_{i,t}$, $i = 1, 2$. Let $\boldsymbol{u}_t = (\epsilon_t, u_{1,t}, u_{2,t})'$, generated from a multivariate normal distribution with correlation matrix $\boldsymbol{\Upsilon}_t$ specified further below. The predictions of z_t via the $x_{i,t}$, and hence loss differentials to be used for all the test statistics, are—unless indicated otherwise—obtained by simple (recursive) OLS as in (12), taking $h = 0$ for simplicity.

We study the fluctuation ($\mathcal{T}^F$), CUSUM ($\mathcal{T}^Q$) and Cramér-von Mises ($\mathcal{T}^C$) statistics discussed in (4), (5) and (6).[20]

We also investigate "asymptotic" fixed-b tests for completeness, abbreviated as "asy" in the figures, as opposed to "bs" for the bootstrap versions. In these, the fixed-b versions of the $\mathcal{T}^Q$, $\mathcal{T}^C$ and $\mathcal{T}^F$ statistics are compared against critical values derived from fixed-b limiting distributions which, unlike those reported in Proposition 1, do not account for time-varying variances and thus suffer from non-pivotality under heteroskedasticity. Concretely and directly extending the approach of Kiefer and Vogelsang (2005), we obtain fixed-b critical values for these tests from simulating the distributions from Proposition 1 under the homoskedasticity assumption that $\mathbf{G}(s) = \mathbf{I}$.[21]

We furthermore follow up on the suggestion of the editor to explore the effectiveness of allowing for a time-varying version of the estimate $\hat{\Omega}$ from (7) (based on $\hat{y}_t$ throughout), generically denoted $\hat{\Omega}_t$, for all the above statistics. Recursive estimates of $\hat{\Omega}_t$ estimate the autocovariances, for $\ell = 1, \ldots, P$, via $\hat{\gamma}_{j,\ell} = \ell^{-1} \sum_{t=|j|+R+1}^{R+\ell} \left(\hat{y}_t - \bar{\hat{y}}\right) \left(\hat{y}_{t-|j|} - \bar{\hat{y}}\right)$, and accordingly, for fixed-b, use a bandwidth $B = b \cdot \ell$.[22] (For small-b, refer to footnote 24, with $P = \ell$.) The rolling estimates, with a window size chosen as $E = [0.3 \cdot P]$ and $\ell = E + 1, \ldots, P$, are based on $\hat{\gamma}_{j,\ell} = E^{-1} \sum_{t=|j|+R+1+\ell-E}^{R+\ell} \left(\hat{y}_t - \bar{\hat{y}}\right) \left(\hat{y}_{t-|j|} - \bar{\hat{y}}\right)$ and $B = b \cdot E$.[23]

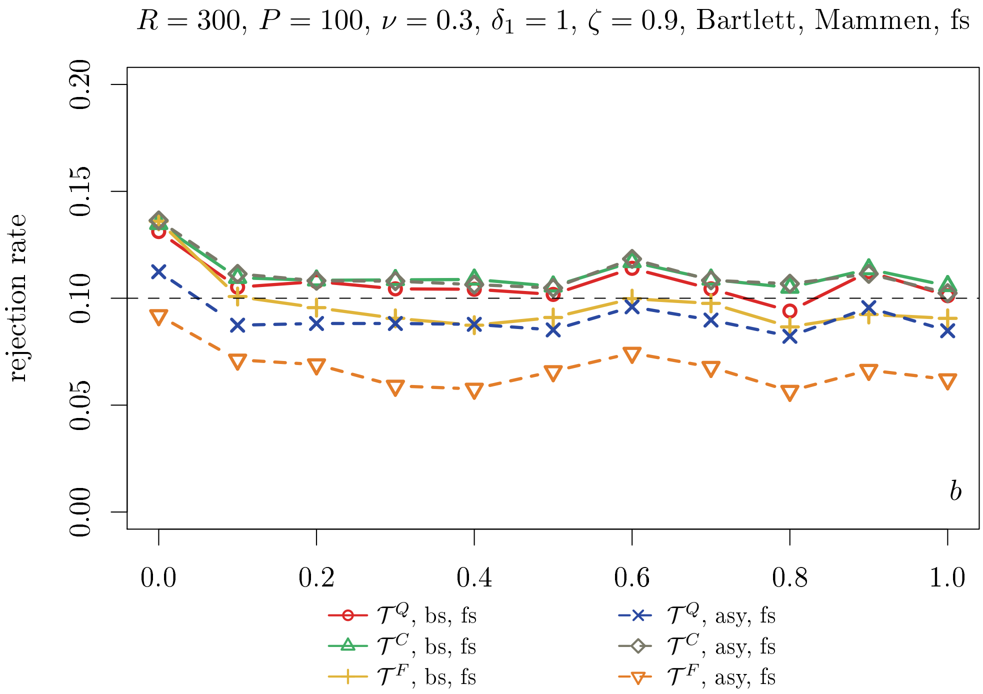
Under time-varying variances, such an estimate is potentially more effective in capturing the behavior of the statistics under consideration over time. At the same time, both small- and fixed-b approaches rely on asymptotics assuming the long-run variance is estimated using the entire sample. It is an open issue explored in our simulations to what extent asymptotic and bootstrap critical values accurately reflect the behavior of the test statistics when computed based on such a time-varying estimate.

[20] We additionally considered the one-time reversal (OTR) statistic QLR_P^* put forward by Giacomini and Rossi (2010, Prop. 2). We found the statistic to perform well in situation for which it was designed (viz. small-b and homoskedasticity), and to have a wild bootstrap version constructed analogously to the statistics analyzed here to perform similarly to the above tests. Results are not reported for brevity.

[21] Table 5 in the Appendix E reports these critical values.

[22] We also impose an initial window size of $P/5$ to avoid instable long-run variance estimates that might otherwise arise in very small samples.

[23] Similar to the recursive setup, we take $\hat{\gamma}_{j,\ell} = \hat{\gamma}_{j,E+1}$ for the initial period $\ell = 1, \ldots, E$.



See (4), (5) and (6) for the fluctuation ($\mathcal{T}^F$), CUSUM ($\mathcal{T}^Q$) and Cramér-von Mises ($\mathcal{T}^C$) statistics. All statistics are computed based on recursive OLS. “bs” abbreviates the bootstrap versions, cf. Algorithm 2. “asy” uses standard non-robust fixed- or small-b critical values. “fs” denotes test statistics computed based on a full-sample estimate of the long-run variance matrix, while “tv” (in later figures) refers to a time-varying variance estimate. R denotes the estimation sample, P the prediction sample, ν the relative window width of $\mathcal{T}^F$ (see (4)), δ_1 the post-break variance and ζ the breakfraction. Bartlett denotes the Bartlett kernel, QS is short for quadratic spectral (in later figures). Mammen and Normal (in the additional simulation results) indicates the bootstrap distribution used in Step 2 of Algorithm 2. See the main text for further details.

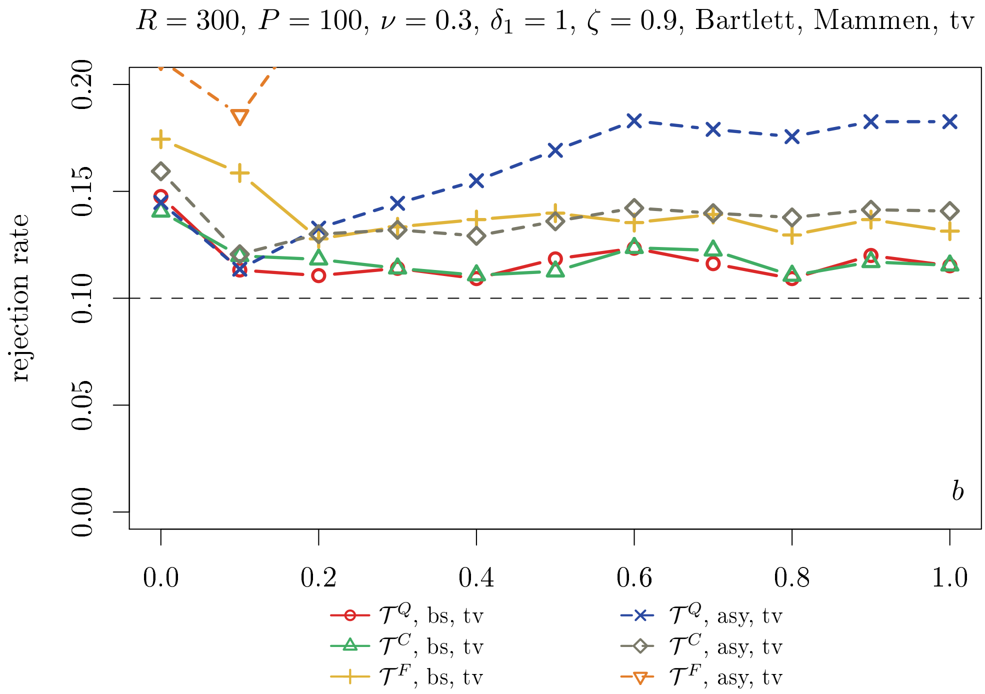
Figure 3: Size under homoskedasticity, asymptotic and bootstrap tests, full-sample covariance estimate

The statistics computed with such a time-varying estimate are denoted with “tv” in the figures, as opposed to those based on the full-sample estimate $\hat{\Omega}$, denoted by “fs”. More specifically and in line with the logic of the corresponding test statistics, we employ recursive estimates (i.e., using an increasing window of observations) $\hat{\Omega}_t$ for $\mathcal{T}^Q$ and $\mathcal{T}^C$ and rolling estimates $\hat{\Omega}_t$ for $\mathcal{T}^F$. For the bootstrap versions of the tests, we also compute corresponding bootstrap estimates of the time-varying covariances matrix estimates for each point in time $R+1,\dots,R+P$.

Bootstrapping such long-run variance estimates at each point in time in a Monte Carlo study on a large parameter grid is computationally rather expensive. The size and power experiments hence consider and report a (representative) subset of cases from the grid $R \in \{50, 100, 200, 300\}$, $P \in \{50, 100, 200, 500, 1000\}$, both the Bartlett and Quadratic Spectral (QS) kernels, $b \in \{0, 0.1, 0.2, \dots, 1\}$[24] and $\sigma_u \in \{-0.4, -0.2, 0, 0.2, 0.4, 0.5, 0.6\}$. Moreover, time-varying variance is introduced by generating a structural break in the covariance matrix by scaling $\boldsymbol{\Upsilon}_t$ by $\delta_1 \in \{1/3, 1, 3\}$, at break dates specified further below.

In step 4 of Algorithm 2, we draw r_t^* from the Mammen (1993) two-point distribution, but also report

[24] For $b = 0$, we use the automatic estimator for B, $\hat{B} = d(4\hat{\rho}^2(1-\hat{\rho})^{-4}P)^{1/g}$ with $\hat{\rho}$ from an approximating $AR(1)$ model for the series (see Andrews, 1991, eqs. (6.2) and (6.4)). Here, $d = 1.1447$, $g = 3$ for the Bartlett and $d = 1.3221$, $g = 5$ for the QS kernel.



See notes to Figure 3.

Figure 4: Size under homoskedasticity, asymptotic and bootstrap tests, time-varying covariance estimate

robustness checks for normal errors in a supplementary appendix.[25] These choices are common in the literature. In view of the findings of Giacomini and Rossi (2010, Table II), we choose a relative window size of $\nu = 0.3$ for $\mathcal{T}^F$. We test against two-sided alternatives at a nominal level of $\alpha = 0.1$ and use $M = 500$ bootstrap replications for the wild bootstrap tests.

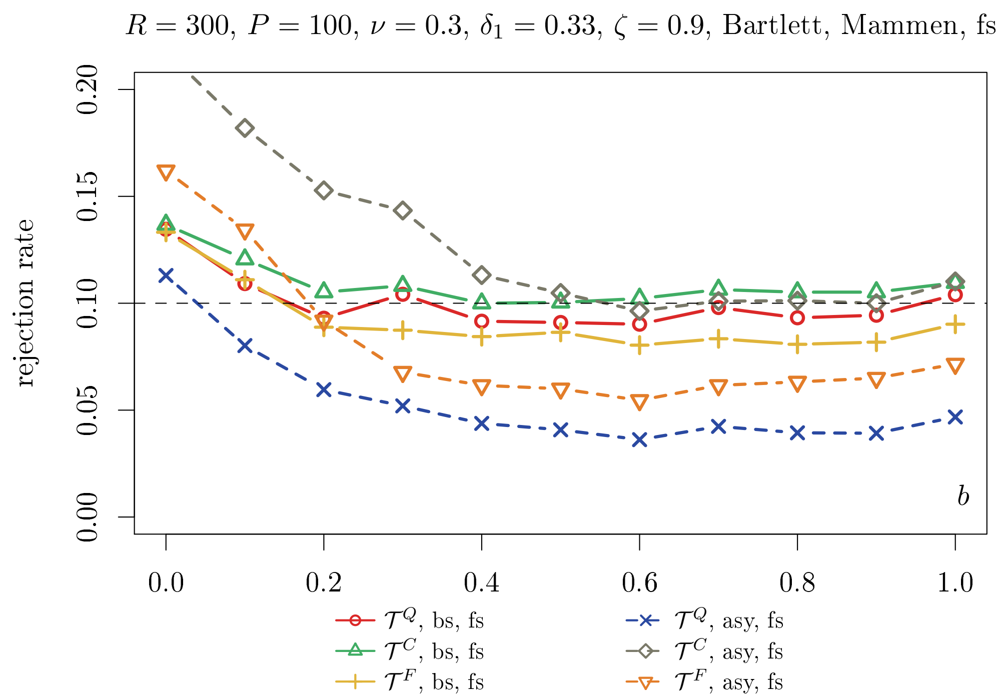
D.1 Size

The size experiments, based on 5,000 replications, take $\boldsymbol{\Upsilon}_t = \boldsymbol{\Upsilon}$ to be an equicorrelation matrix with identical off-diagonal elements $\sigma_u = 0.5$. This yields a scenario in which $x_{1,t}$ and $x_{2,t}$ have equal predictive ability for z_t so that the null hypothesis of the tests is true. Here, we scale by δ_1 at times $[\zeta \cdot (R + P)]$, where $\zeta \in \{0.3, 0.6, 0.9\}$. For instance, $\delta_1 = 1/3$ and $\zeta = 0.9$ yield a late downward break in variance.

First, Figure 3 shows that the full-sample covariance estimate based versions of $\mathcal{T}^Q$ and $\mathcal{T}^C$ fixed-b tests perform well across b under the benchmark case of homoskedasticity ($\delta_1 = 1$), as does the bootstrap $\mathcal{T}^F$ test. The asymptotic fluctuation test is slightly undersized for $b > 0$. Unreported simulations for larger P reveal this to be, as expected, a small-sample phenomenon. It is worth stressing that Giacomini and Rossi (2010) focus on $b = 0$, a value for which $\mathcal{T}^F$ performs very well, so that our findings do not contradict theirs. Also, the results are reminiscent of Kiefer and Vogelsang (2005)—while fixed-b tests yield good finite-sample size, there are finite-sample size distortions for the small-b versions (Newey and West, 1987; Andrews, 1991), i.e., for $b = 0$.

The bootstrap appears effective in controlling size for both full-sample $\hat{\Omega}$ (Figure 3) and time-varying

[25] FOR THE REFEREES: please refer to Appendix F.



See notes to Figure 3.

Figure 5: Size under heteroskedasticity (late downward break), asymptotic and bootstrap tests, full-sample

$\hat{\Omega}_t$ (Figure 4) for all statistics. Under asymptotic c.v.s, the use of $\hat{\Omega}_t$—which would not have been necessary given $\delta_1 = 1$—however generally leads to somewhat inferior performance for all tests here, possibly reflecting the above-mentioned mismatch between the assumptions and practice of estimating the long-run variance. No clear ranking emerges for the proposals of the present paper.

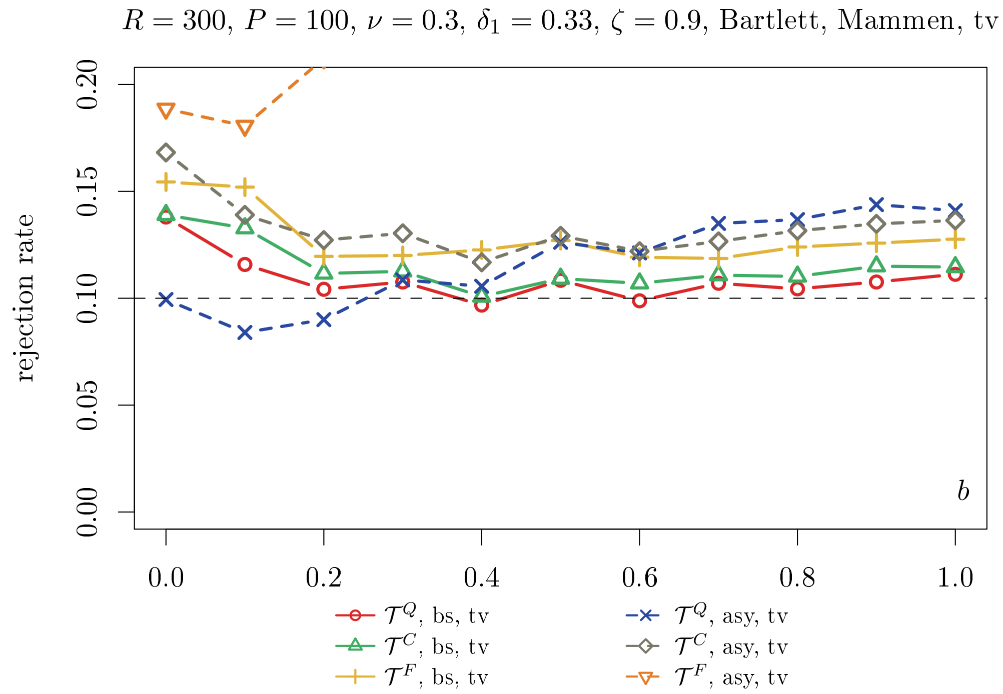
The non-pivotality of the asymptotic tests under heteroskedasticity becomes apparent in Figures 5-8. Here, the break occurs at observation $[0.9 \cdot (300 + 100)] = 360$. The first $R = 300$ preliminary observations have been used for parameter estimation. In particular, the full-sample asymptotic tests are distorted as soon as b takes moderate or large values. That is, fixed-b versions of the tests, as predicted by Proposition 1, no longer provide accurate finite-sample size in the presence of time-varying variance, although a time-varying estimate appears as an effective remedy for some asymptotic tests in this case (cf. Figures 6 and 8).

In turn and as a result of Proposition 2, the bootstrap fixed-b versions maintain good size. Again, they are slightly less successful at correcting the well-known small-sample small-b size distortions. Also, the fixed-b approximations work slightly less well for smaller b, still being close to the standard small-b case ($b = 0$), which is in line with Kiefer and Vogelsang (2005). For about $b > 0.2$, the bootstrap tests generally perform very well.[26]

Focussing on the robust bootstrap tests, Figures 9 reveals that there is little to choose between the Bartlett and QS kernel in terms of size. Both perform similarly well.

Figure 10 demonstrates, for $\mathcal{T}^Q$, that P has a minor effect on both versions of the bootstrap tests. Unlike for the time-varying covariance matrix based version, size appears to improve for the full-

[26] There is some small and unsystematic variation in the empirical sizes when varying b. We consider this to be due to simulation variability given the relatively small number of Monte Carlo replications for each case.



See notes to Figure 3.

Figure 6: Size under heteroskedasticity (late downward break), asymptotic and bootstrap tests, time-varying

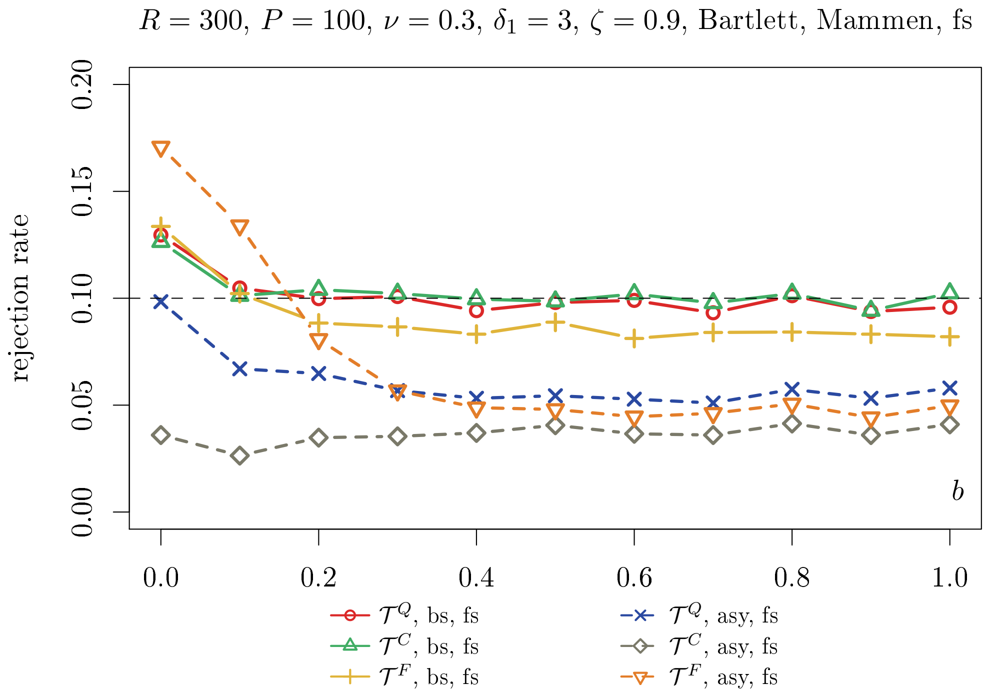
sample covariance asymptotic tests, but this finding is not robust with respect to other ζ and δ_1.

We complete the discussion of the size results by an additional experiment in which there is no orthogonality between residuals and regressors, with corresponding effects on the relevance of correcting for estimation error. Concretely, instead of via recursive OLS, we construct the predictions for z_t via the $x_{i,t}$ via recursive instrumental variables (IV) estimation, using the lag of the $x_{i,t}$, $x_{i,t-1}$, as instrument for the predictors. Given the predictors' autoregressive structure, such instruments satisfy the relevance condition of a valid IV. The only necessary modifications in Algorithm 2 presented in Appendix A are (see also the discussion below Assumption 1) that, now, $\mathbf{C}_{i,t} = \boldsymbol{x}_{i,t}\boldsymbol{x}'_{i,t-1}$ and $\hat{\boldsymbol{a}}_{i,t} = \boldsymbol{x}_{i,t-1}\hat{\epsilon}_{i,t}$, with $\hat{\epsilon}_{i,t}$ now denoting the IV residuals.

Figures 11 and 12 present results. These confirm the capability of the robust bootstrap tests to also provide accurate inference in IV-based setups, with the asymptotic tests again exhibiting expected size distortions under heteroskedasticity. We do not observe relevant qualitative differences to the OLS-based results obtained with the same parameters of the DGP (cf. Figures 5 and 6), suggesting minor importance of the particular estimation scheme at least in the experiments considered here.

D.2 Power

In our power experiments, based on 2,500 replications, we specify two distinct scenarios. First, we consider a time-invariant so-called "Toeplitz" structure for $\boldsymbol{\Upsilon}_t$. More specifically, both ϵ_t and $u_{1,t}$ as well as $u_{1,t}$ and $u_{2,t}$ are correlated (with a correlation coefficient of σ_u) while ϵ_t and $u_{2,t}$ are uncorrelated. Thus, $x_{2,t}$ is independent of z_t and therefore has no predictive power, in contrast to $x_{1,t}$. Time-varying variance is generated as in the size experiments.



See notes to Figure 3.

Figure 7: Size under heteroskedasticity (late upward break), asymptotic and bootstrap tests, full-sample

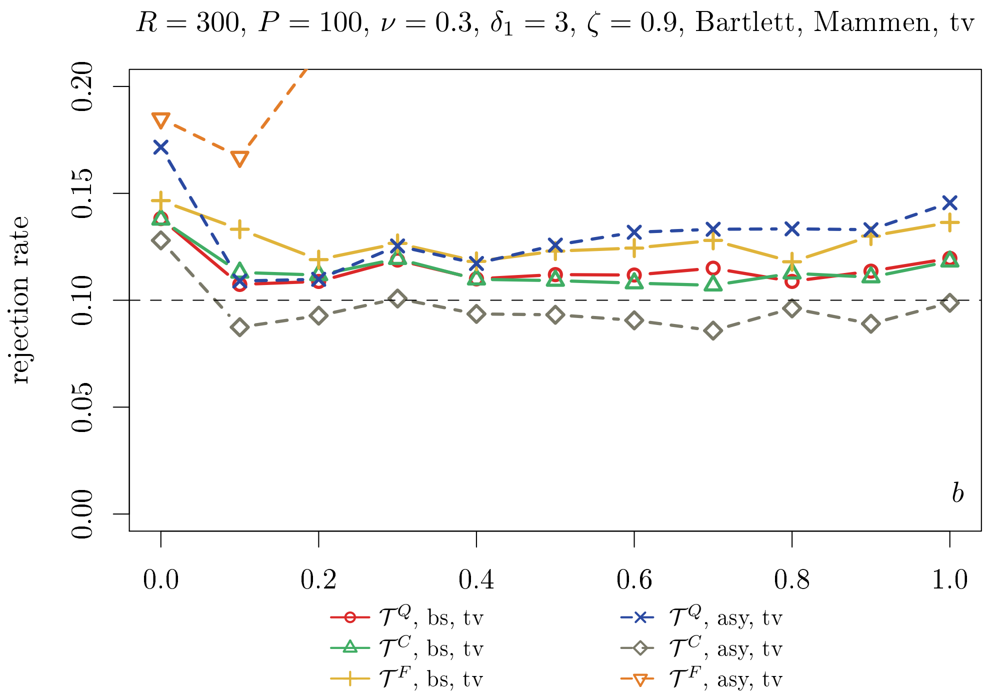
In order to generate time-varying forecasting ability, we specify a simple switch from an equicorrelated matrix to a "Toeplitz" matrix at time $\tau := [R + P/4]$. Thus, the structural break in predictive power emerges from a time-varying correlation matrix $\boldsymbol{\Upsilon}_t$. The break date hence is located in the first quarter of the prediction sample and renders the DGP practically relevant.[27] Here, the variance break also occurs at τ.

We first consider results from the "Toeplitz" experiments. First, Figures 13 and 14 show that the power of both bootstrap and asymptotic tests increases in P both for full-sample and time-varying estimation of the long-run covariance matrix. Second, observing that this power experiment corresponds to the size study reported in Figures 5 and 6, it comes as no surprise that the power ranking is strongly affected by whether a test accurately exhausts or even exceeds nominal size.

For example, the asymptotic full-sample CUSUM statistic $\mathcal{T}^Q$ is fairly undersized for $b = 0.4$, negatively affecting its power. The asymptotic Cramér-von Mises statistic $\mathcal{T}^C$ is slightly oversized, with corresponding positive impact on power. Recall that Section D.1 revealed that the bootstrap tests generally effectively exhaust nominal size, implying that their power is either better than that of the asymptotic tests when the latter are undersized, or more credible when the latter are oversized. Third, $\mathcal{T}^Q$ and $\mathcal{T}^C$ perform well and quite similarly in terms of power.

Upward size distortions are even more pronounced for the asymptotic tests using time-varying variance estimates (Figure 6), so that their power as reported in Figure 14 is not associated with the nominal type-I error. Figure 15 therefore reports size-adjusted power, where size-adjusted critical values—bearing in mind that, as argued forcefully by Horowitz and Savin (2000), the particular

[27] We also experimented with later values of τ. Of course, a smaller sample in which the predictors' forecasting ability differs translates into lower power, but the general qualitative conclusions of our study remain the same.



See notes to Figure 3.

Figure 8: Size under heteroskedasticity (late upward break), asymptotic and bootstrap tests, time-varying

point chosen in composite space of points satisfying the null so as to be able to obtain such critical values and hence adjusted power is always somewhat arbitrary—are computed from a simulation in which $\sigma_u = 0$ for both predictors.[28]

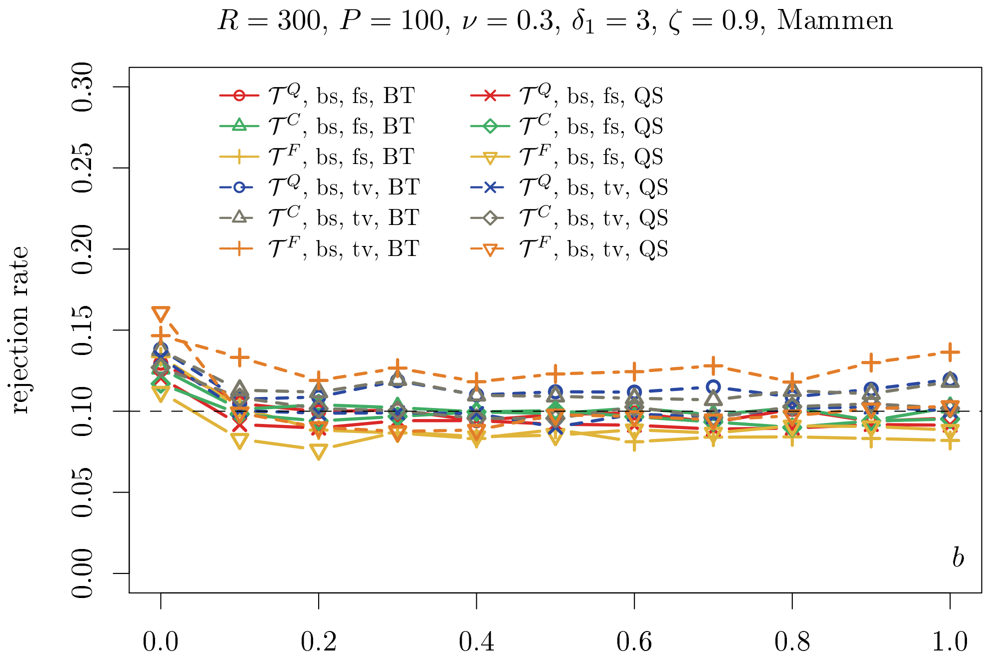
As expected, the test statistics that are somewhat undersized (cf. again Figures 5 and 6) do relatively better when computing their power from size-adjusted critical values, and the high rejection rates of the tests that were oversized against unadjusted critical values drop accordingly. When benchmarked against adjusted critical values, we still find the versions of $\mathcal{T}^Q$ and $\mathcal{T}^C$ test to perform best under this specific DGP. The fluctuation tests' $\mathcal{T}^F, asy$ and $\mathcal{T}^F, tv$ performances come next.

That said, as adjusted critical values are of course not available in practice, it is not clear how to exploit this finding in applications. We therefore agree with Horowitz and Savin (2000) that the bootstrap as, for example, discussed in this paper is a reasonable way to obtain tests with good size and hence credible power properties in the presence of asymptotically non-pivotal statistics.

In any case, the above power study with constant relative forecasting ability may of course be somewhat more geared towards statistics such as $\mathcal{T}^Q$ and $\mathcal{T}^C$ that take more of a full-sample perspective than, e.g., $\mathcal{T}^F$. We therefore now present some of the results for the time-varying forecasting ability case in which the predictive power of $x_{2,t}$ for z_t is identical to that of $x_{1,t}$ until τ. After τ, the predictive power of $x_{2,t}$ for z_t vanishes. Figures 16 (Bartlett) and 17 (QS) compare the power for the two kernels in the bootstrap case. Here, we plot power against σ_u. First and as expected, the power of all tests increases in $|\sigma_u|$. This is because the predictive power of $x_{1,t}$ for z_t then is larger relative to that of $x_{2,t}$ after τ.

Comparing the entries of Figure 16 with the corresponding one of Figure 17 reveals that the Bartlett

[28] Alternatively, it would have been no less plausible to set, for example, $\sigma_u = 0.4$ for both predictors.

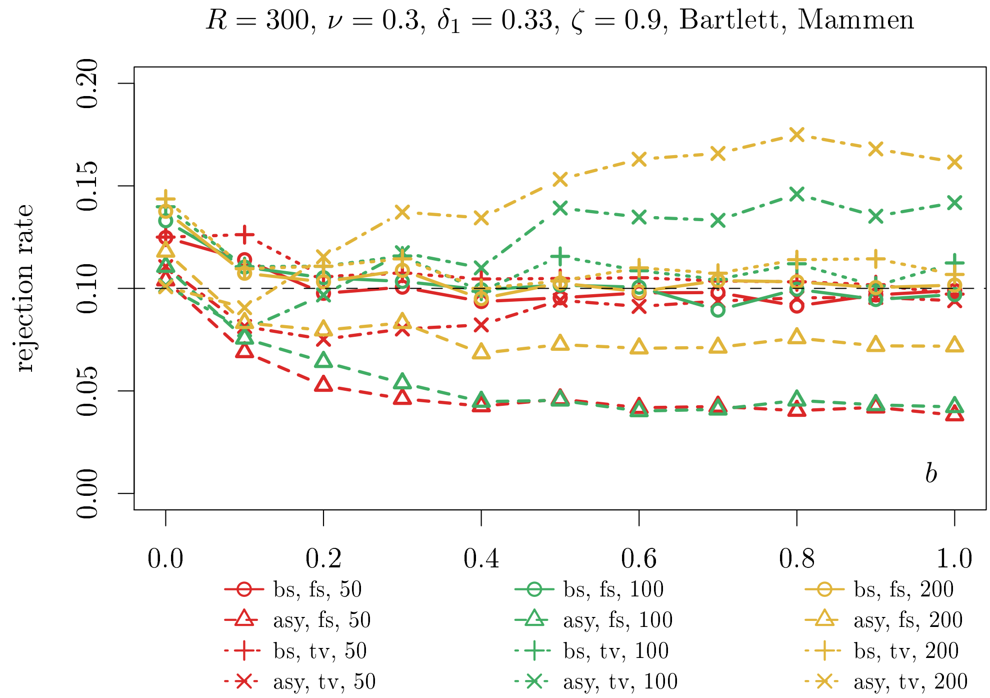


See notes to Figure 3.

Figure 9: Bootstrap size under heteroskedasticity, different kernels

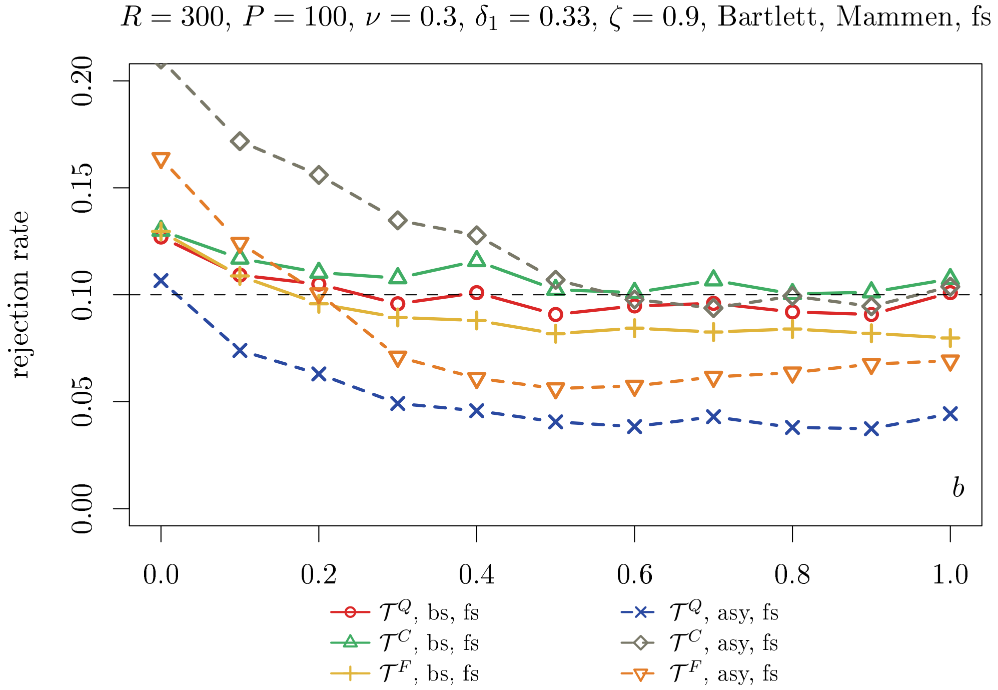
kernel leads to more powerful tests. This is noteworthy, as both variants fairly effectively exhaust nominal size (cf. the entries at $\sigma_u = 0$; see also Figure 9). Figures 16 and 17 also reveal that, relative to the constant relative forecasting ability case, the different variants of $\mathcal{T}^F$ become more attractive compared to $\mathcal{T}^Q$ and $\mathcal{T}^C$. In particular, the bootstrapped version of $\mathcal{T}^F$ with a time-varying covariance matrix estimate performs very well. This is intuitive in that the fluctuation test may be expected to work relatively better in cases in which the relative forecasting ability is time-varying. Moreover, the use of a time-varying variance estimate seems to further improve performance in such a setup. In view of the small size distortions of the bootstrap tests, we waive to report size-adjusted power in this case.

Figures 18-20 demonstrate (for the Bartlett kernel) that power increases in P also under time-varying relative forecasting ability, the reason being that the time span during which a change in forecasting ability can be detected also increases. Intuitively, we again find the "rolling" $\mathcal{T}^F$ tests to perform relatively better in such a scenario. More specifically, when assessed against size-adjusted critical values (computed from a scenario in which $\sigma_u = 0$ throughout), Figure 20 reveals $\mathcal{T}^F$ with a time-varying covariance matrix estimate to perform best for $P > 200$, with the full-sample version of $\mathcal{T}^Q$ slightly more powerful for smaller P. A comparison of Figures 15 and 20 however also reveals that power of all tests is generally lower in the time-varying forecasting ability case. This is as expected, as there is a smaller period $R + P - \tau$ during which they may detect differences in forecasting power. There appears to be no clear pattern in this scenario as to whether a time-varying or full-sample covariance estimate leads to higher power.



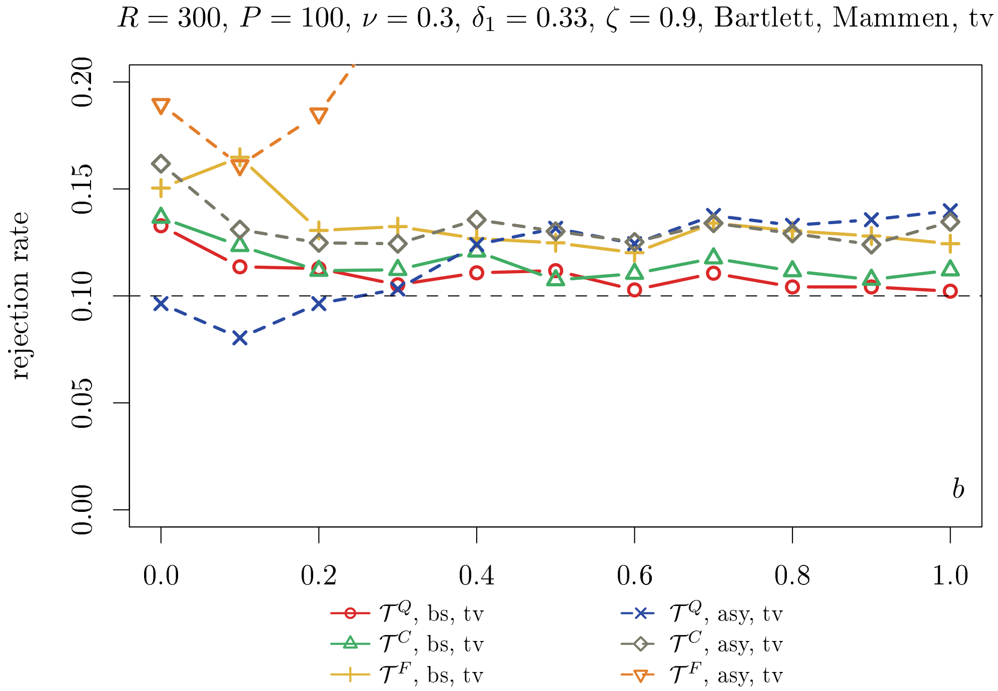
See notes to Figure 3.

Figure 10: Size of $\mathcal{T}^Q$ under heteroskedasticity for different P, asymptotic and bootstrap tests, full-sample and time-varying



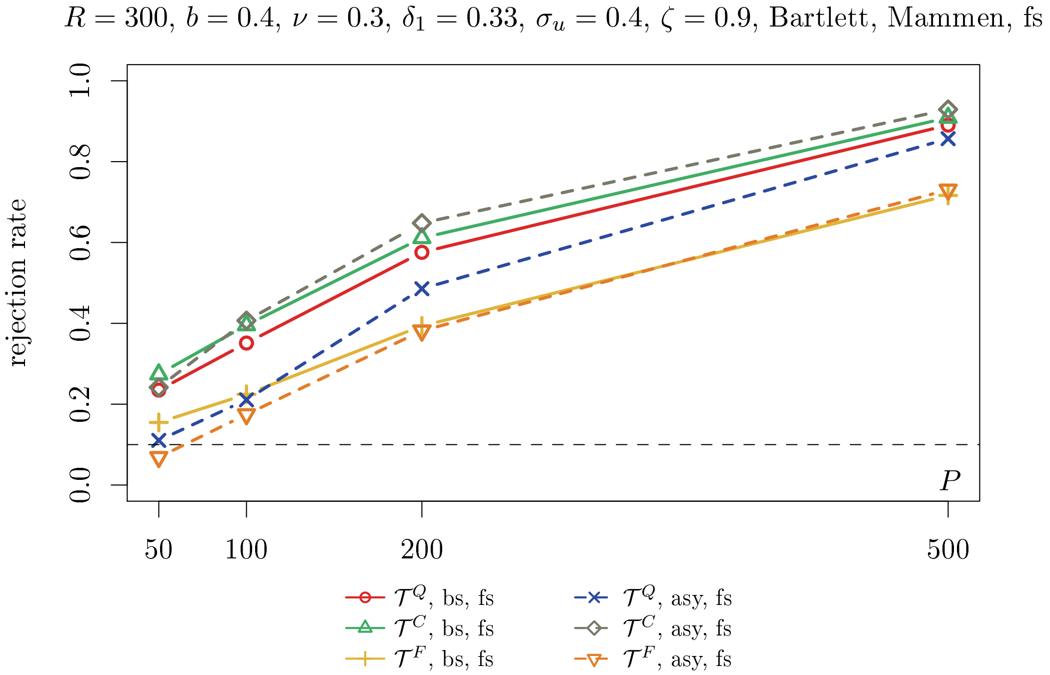
See notes to Figure 3.

Figure 11: IV estimation, size under heteroskedasticity, asymptotic and bootstrap tests, time-varying covariance estimate



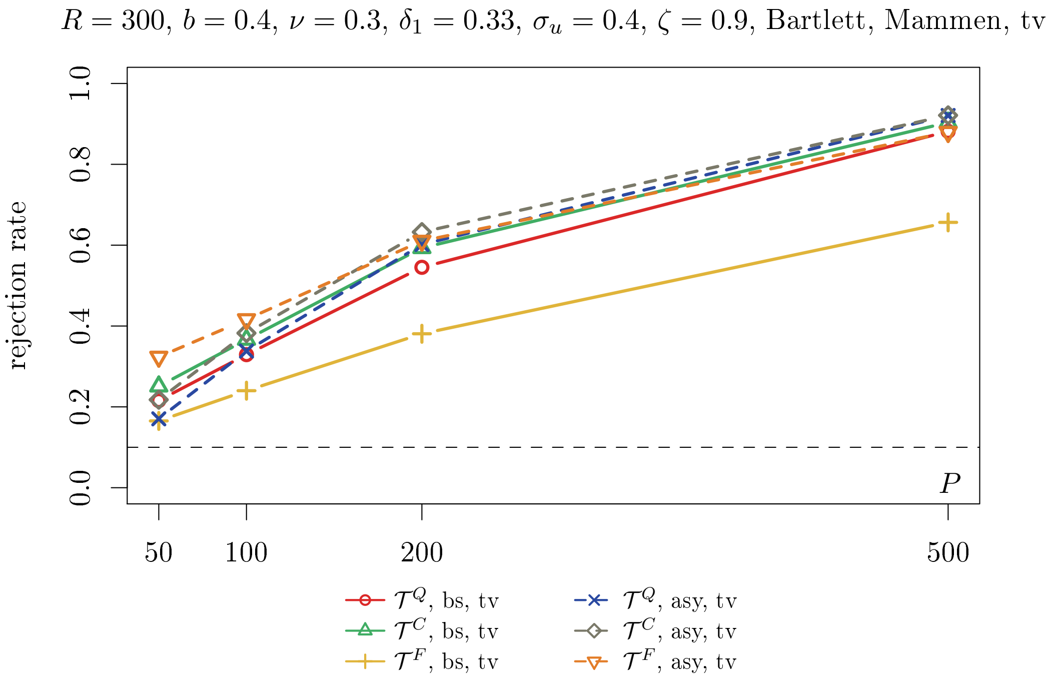
See notes to Figure 3.

Figure 12: IV estimation, size under heteroskedasticity, asymptotic and bootstrap tests, time-varying covariance estimate



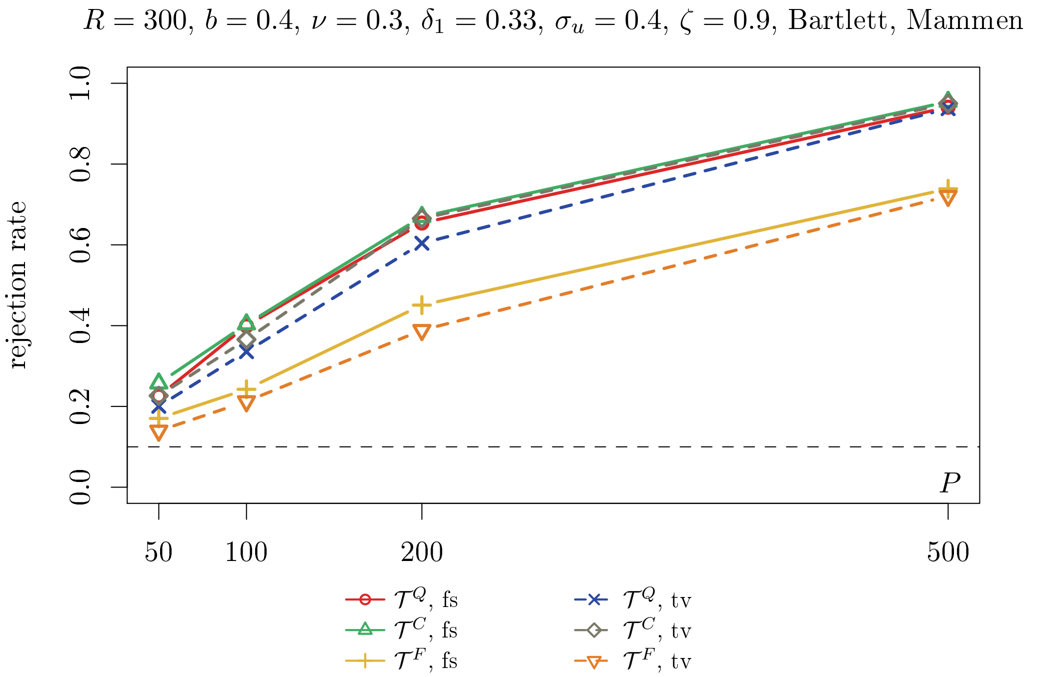
See notes to Figure 3.

Figure 13: Power vs. P, constant relative forecasting ability, asymptotic and bootstrap tests, full-sample



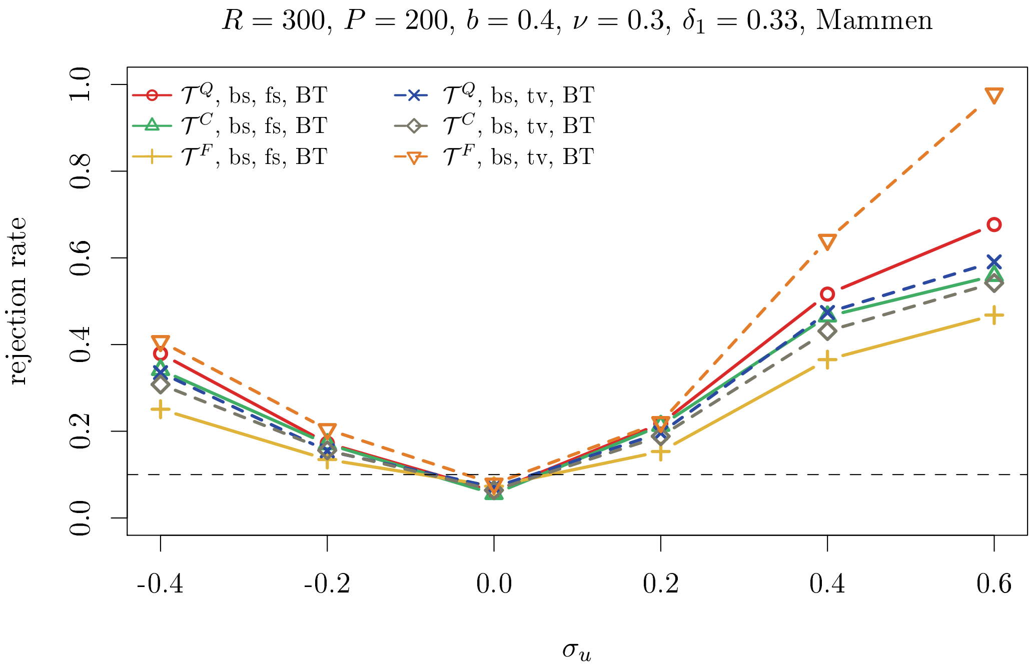
See notes to Figure 3.

Figure 14: Power vs. P, constant relative forecasting ability, asymptotic and bootstrap tests, time-varying



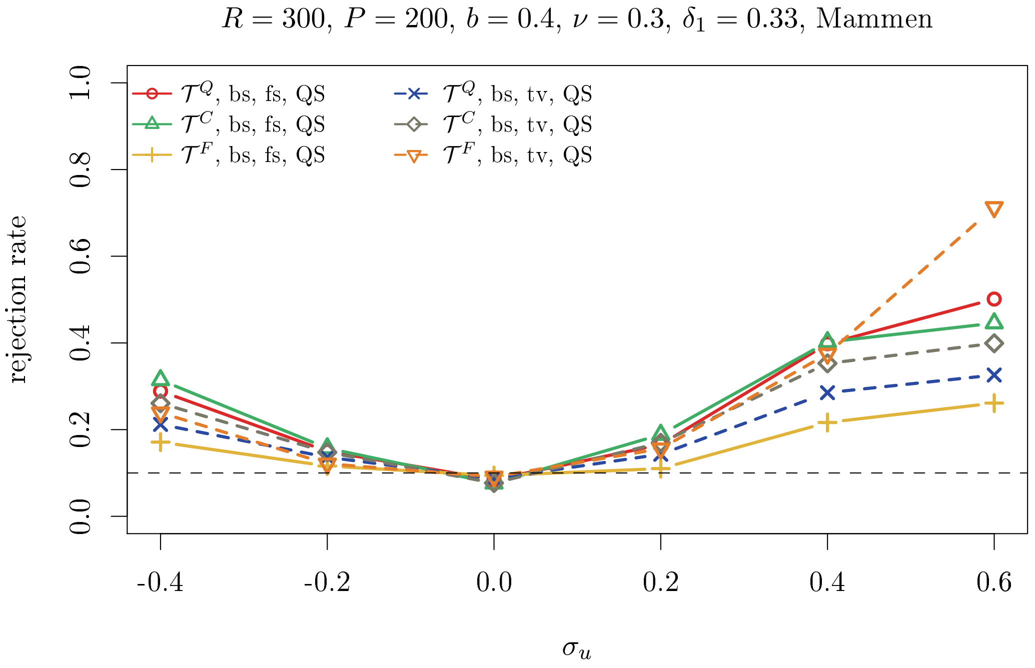
See notes to Figure 3.

Figure 15: "Size-adjusted power" vs. P, constant relative forecasting ability



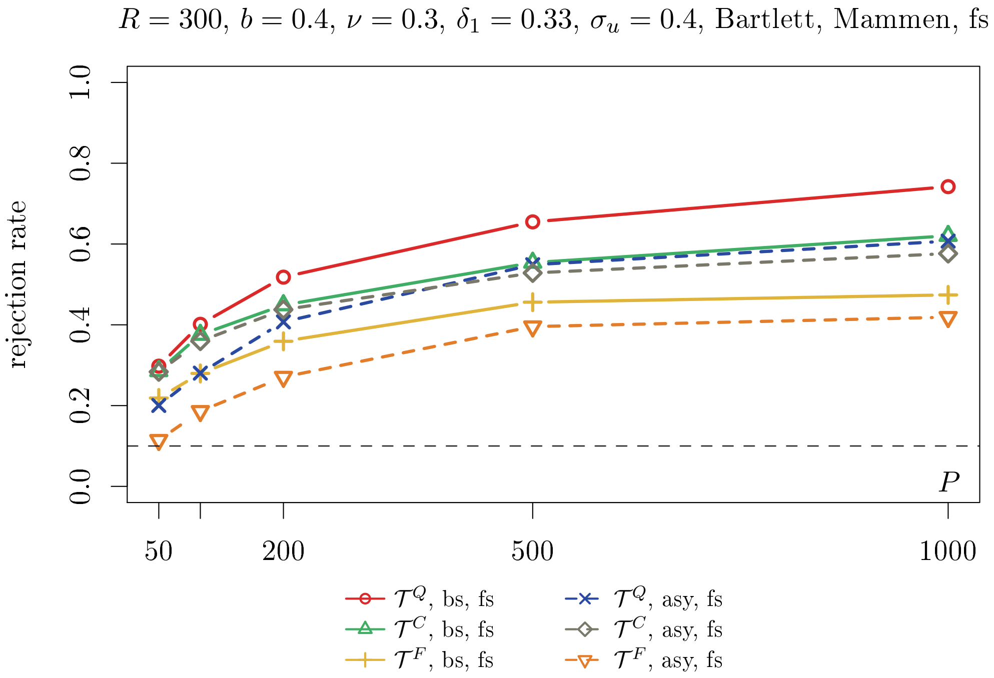
See notes to Figure 3.

Figure 16: Power bootstrap tests vs. σ_u, time-varying relative forecasting ability, Bartlett



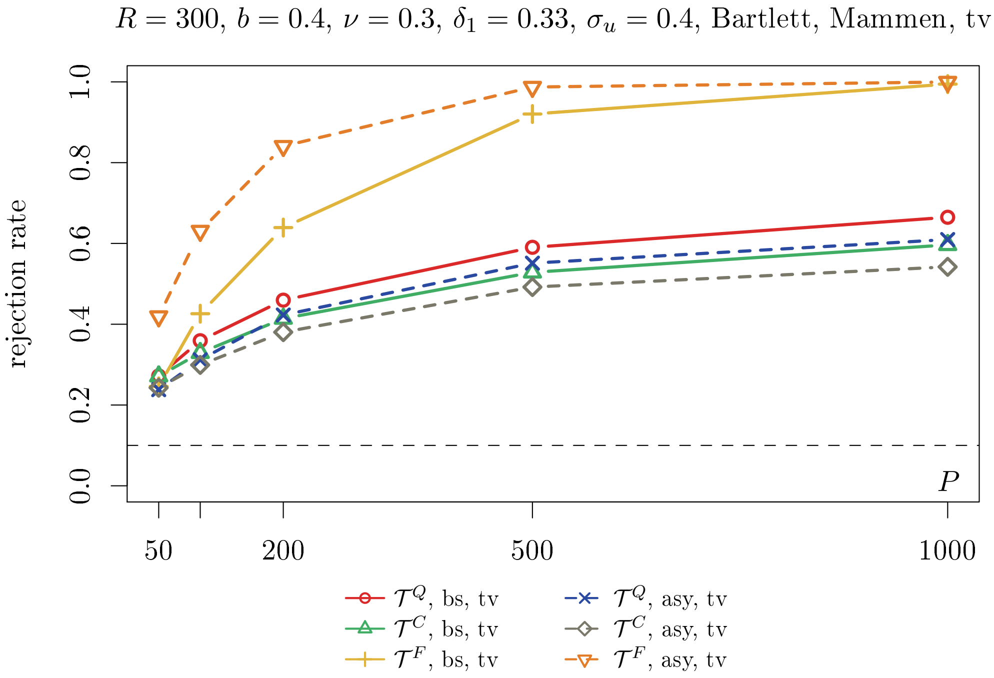
See notes to Figure 3.

Figure 17: Power bootstrap tests vs. σ_u, time-varying relative forecasting ability, Quadratic Spectral



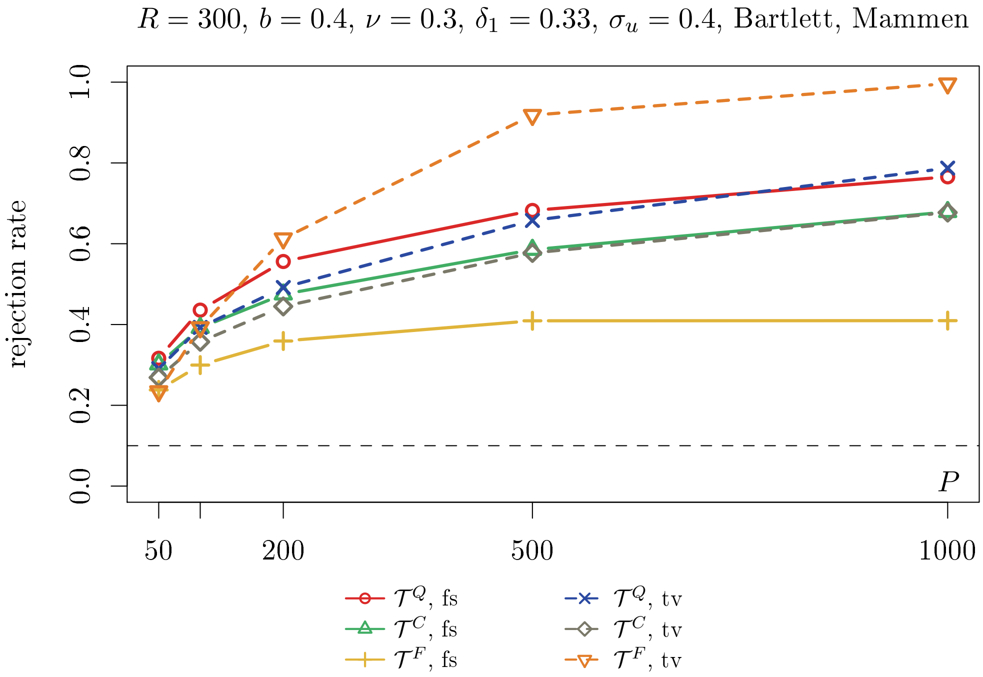
See notes to Figure 3.

Figure 18: Power vs. P, time-varying relative forecasting ability, asymptotic and bootstrap tests, full-sample



See notes to Figure 3.

Figure 19: Power vs. P, time-varying relative forecasting ability, asymptotic and bootstrap tests, time-varying



See notes to Figure 3.

Figure 20: “Size-adjusted power” vs. P, time-varying relative forecasting ability, asymptotic and bootstrap tests

E Asymptotic critical values

b	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						10% critical values					
$\mathcal{T}^Q$											
Bartlett	1.97	2.14	2.37	2.63	2.92	3.19	3.46	3.70	3.92	4.14	4.36
QS	1.97	2.25	2.71	3.30	4.08	4.99	5.97	7.02	8.17	9.38	10.68
$\mathcal{T}^C$											
Bartlett	1.21	1.43	1.71	2.06	2.47	2.91	3.42	3.91	4.35	4.89	5.42
QS	1.22	1.57	2.14	3.08	4.47	6.38	9.09	12.29	16.47	21.75	28.04
$\mathcal{T}^F, \nu = 0.3$											
Bartlett	8.05	8.46	9.87	12.13	15.50	19.38	23.27	27.00	30.46	34.01	37.76
QS	8.05	9.30	13.00	20.79	35.50	56.94	85.95	121.04	164.70	218.95	282.40
$\mathcal{T}^F, \nu = 0.5$											
Bartlett	6.52	7.19	8.55	10.44	12.52	14.86	17.86	20.94	23.83	26.69	29.56
QS	6.53	7.95	11.34	16.57	25.37	39.12	58.27	83.59	115.38	153.14	197.37
$\mathcal{T}^{DM}$											
Bartlett	2.71	3.39	4.20	5.19	6.33	7.59	8.91	10.11	11.40	12.75	14.16
QS	2.71	3.76	5.31	7.83	11.52	16.47	22.92	30.83	41.03	53.50	68.53
						5% critical values					
$\mathcal{T}^Q$											
Bartlett	2.25	2.49	2.81	3.17	3.50	3.87	4.19	4.49	4.76	5.03	5.30
QS	2.25	2.65	3.32	4.21	5.36	6.76	8.29	9.94	11.65	13.44	15.35
$\mathcal{T}^C$											
Bartlett	1.69	2.03	2.46	3.07	3.69	4.44	5.16	5.94	6.67	7.44	8.24
QS	1.69	2.26	3.31	5.00	7.86	11.95	17.58	25.19	34.80	46.25	59.28
$\mathcal{T}^F, \nu = 0.3$											
Bartlett	9.58	9.85	11.79	14.80	19.30	24.53	29.33	33.99	37.96	42.41	47.06
QS	9.59	11.17	16.96	29.94	54.87	94.61	150.47	222.97	317.35	428.42	559.59
$\mathcal{T}^F, \nu = 0.5$											
Bartlett	8.14	8.92	10.87	13.57	16.49	19.52	23.79	28.14	31.83	35.77	39.47
QS	8.14	10.08	15.48	24.40	40.58	68.10	105.25	151.32	212.99	288.50	380.02
$\mathcal{T}^{DM}$											
Bartlett	3.83	4.97	6.45	8.04	9.79	11.90	13.92	15.91	17.96	20.12	22.26
QS	3.83	5.68	8.64	13.38	21.02	31.57	46.04	65.35	89.22	119.31	151.89

Table 5: Asymptotic critical values

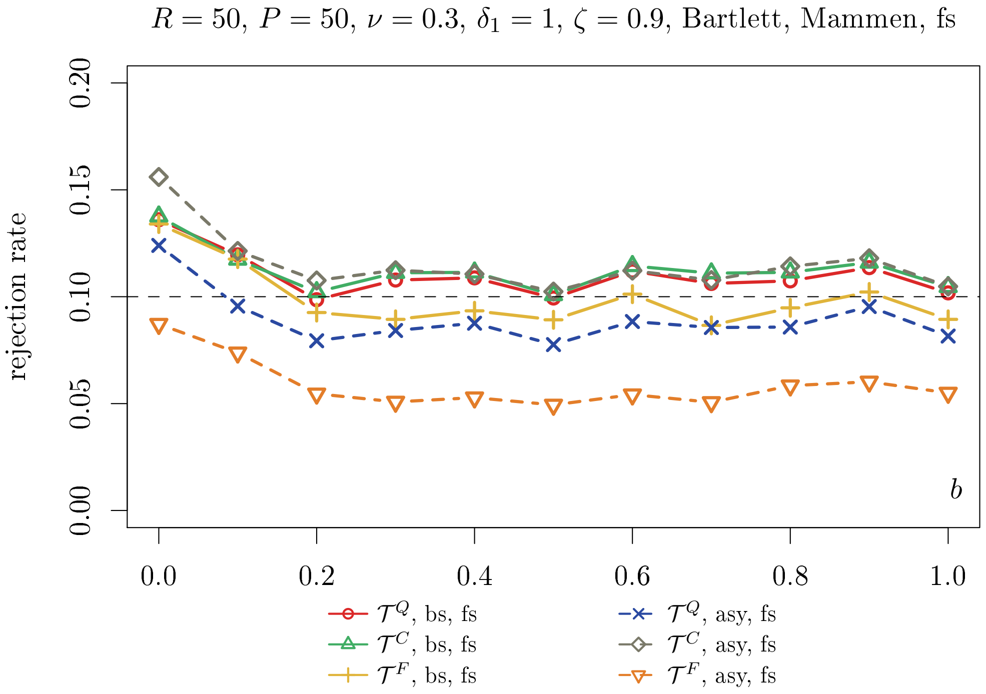
Table 5 reports asymptotic critical values ignoring possible time-varying variance. "Small-b" χ_1^2 quantiles are recovered as special cases for the squared Diebold and Mariano (1995) statistic $\mathcal{T}^{DM}$ for $b = 0$. Also note that, under small-b asymptotics, the critical values are independent of the kernel (up to simulation variability).

F Additional simulation results

This section addresses questions raised by the editor and the referees.

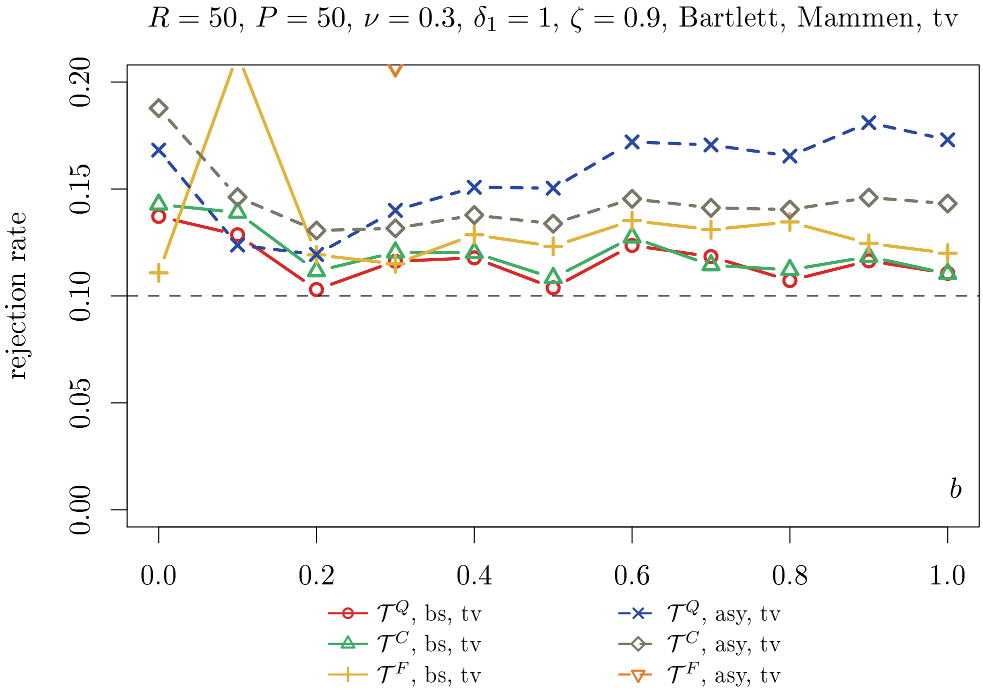
Figures 21 and 22 suggest that our key findings are robust also to smaller sizes of the estimation and prediction samples R and P. Similarly, Figures 23 and 24 demonstrate that choosing R small and P large, i.e., the opposite choices to those in Figures 3 and 4, likewise leaves our conclusions unaffected. Figures 25 to 28—the counterparts to Figures 3 to 6—provide evidence that it does not matter if we choose the Mammen or normal distribution for the bootstrap errors.

Finally, Figures 29-32 consider cases (both for homoskedastic and heteroskedastic series) where the equicorrelated correlation matrix $\boldsymbol{\Upsilon}_t$ additionally undergoes a break from $\sigma_u = 0.5$ to $\sigma_u = 0.2$ at τ. Thus, both regressors' predictive ability decreases after τ, but proportionally so. Hence, the relative predictive performance is unaffected, so that the null hypothesis is true. This design is inspired by a similar size study of a equal relative predictability scenario in Section 4.1 of Giacomini and Rossi (2010). The results demonstrate that the conclusions of the corresponding Figures 3-6 also apply here, in particular regarding the robustness of the wild bootstrap implementations put forward here.



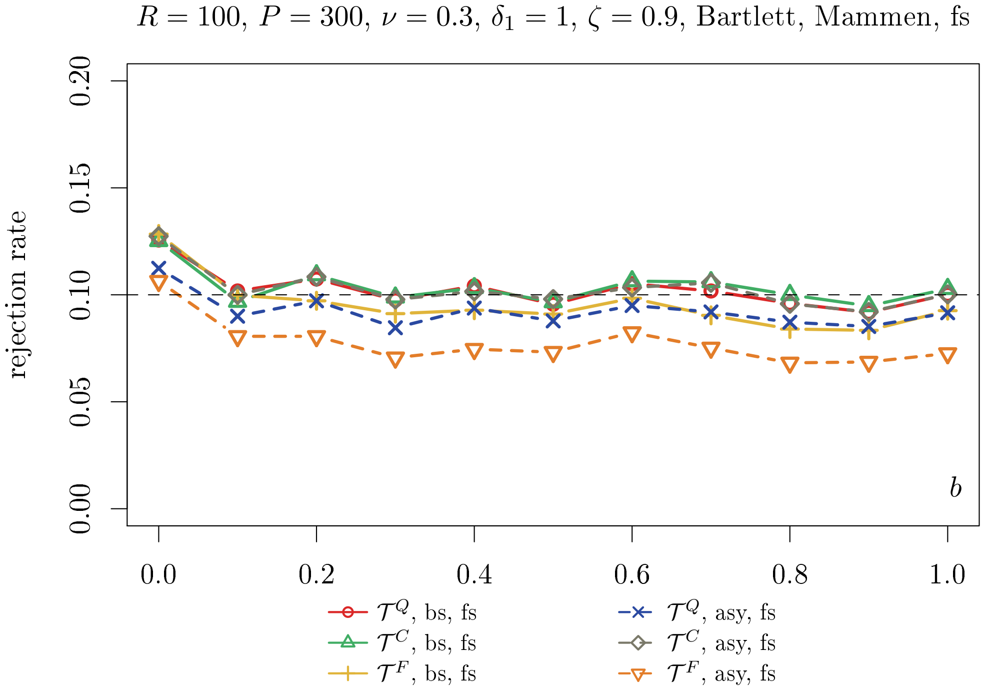
See notes to Figure 3.

Figure 21: Size under homoskedasticity in small samples, asymptotic and bootstrap tests, full sample



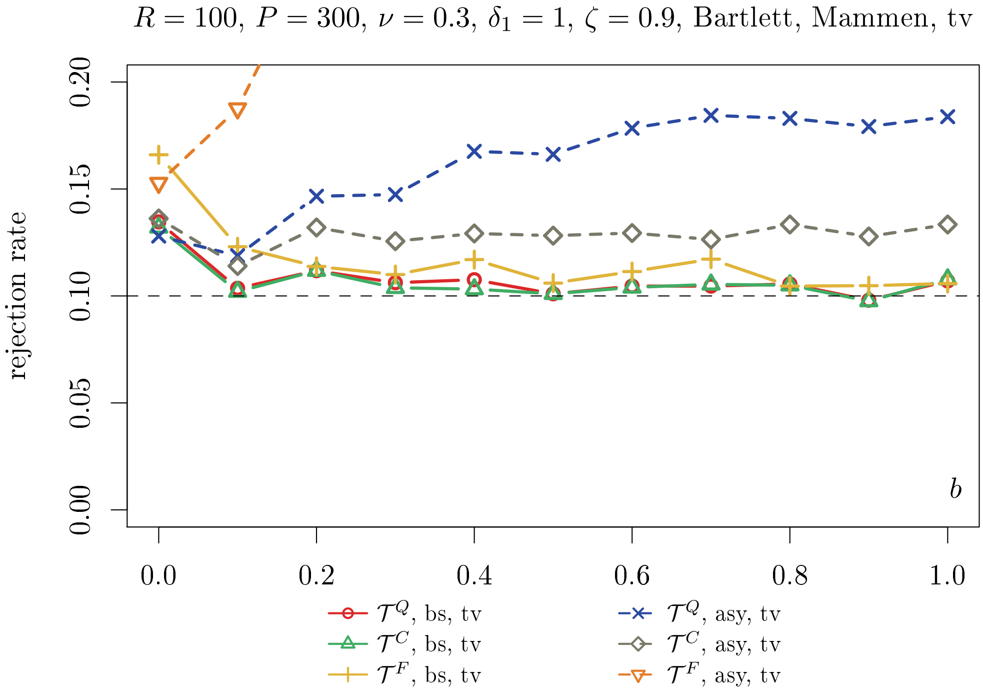
See notes to Figure 3.

Figure 22: Size under homoskedasticity in small samples, asymptotic and bootstrap tests, time-varying



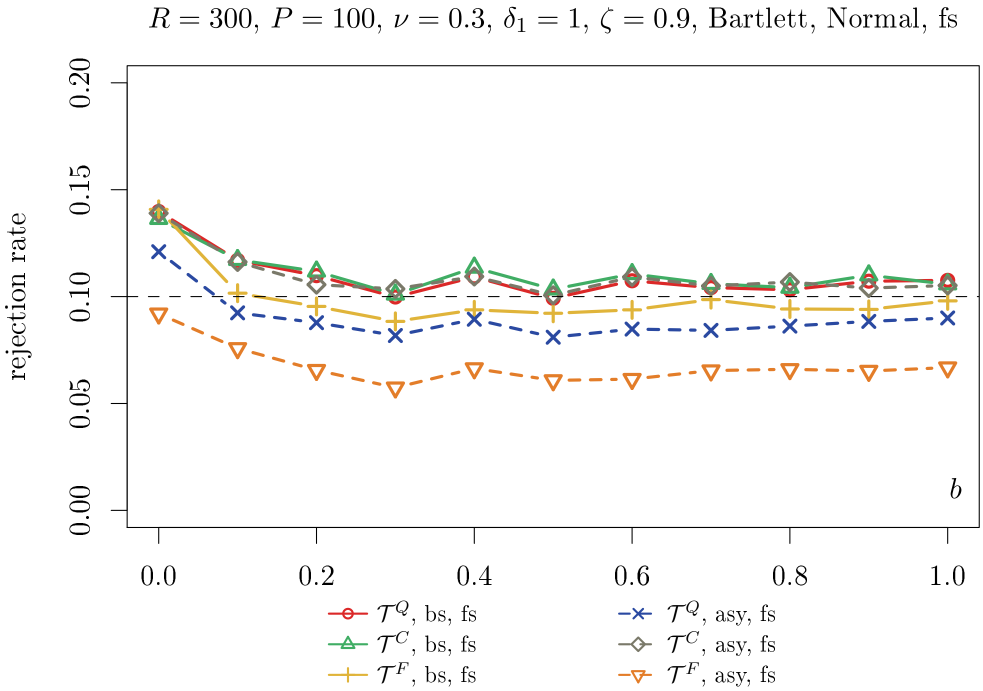
See notes to Figure 3.

Figure 23: Size under homoskedasticity, asymptotic and bootstrap tests, small R, large P, full-sample



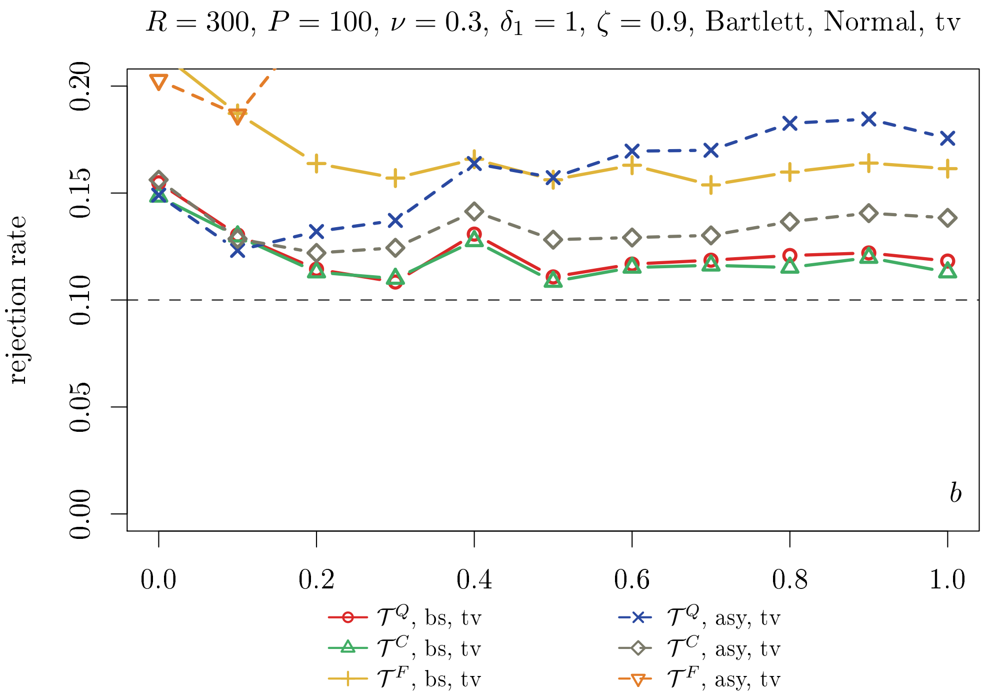
See notes to Figure 3.

Figure 24: Size under homoskedasticity, asymptotic and bootstrap tests, small R, large P, time-varying



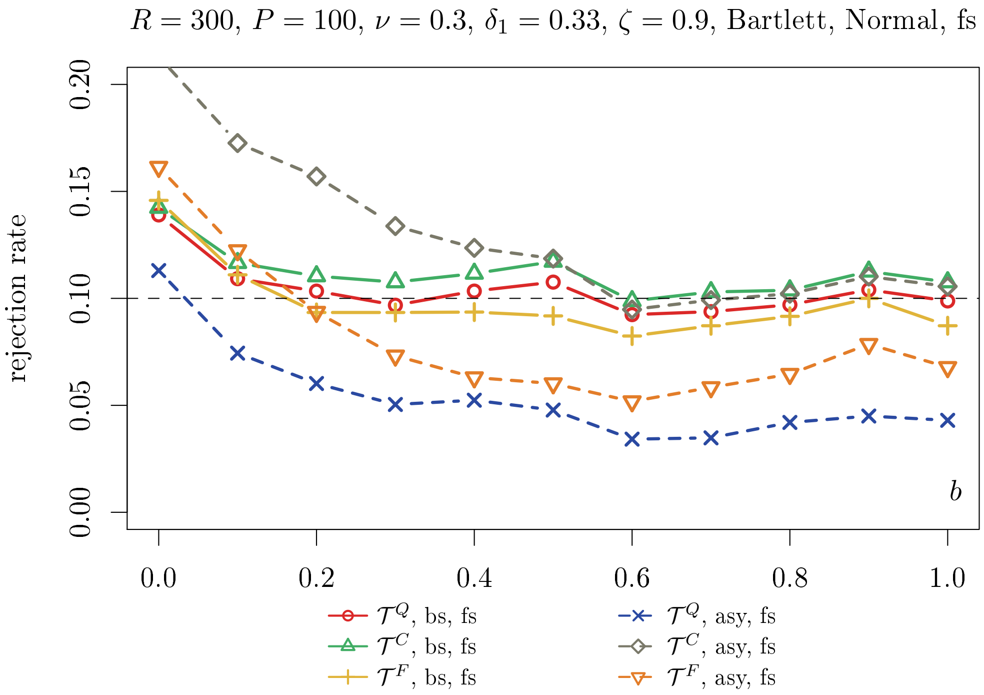
See notes to Figure 3.

Figure 25: Size under homoskedasticity, asymptotic and bootstrap tests, normal bootstrap errors, full-sample



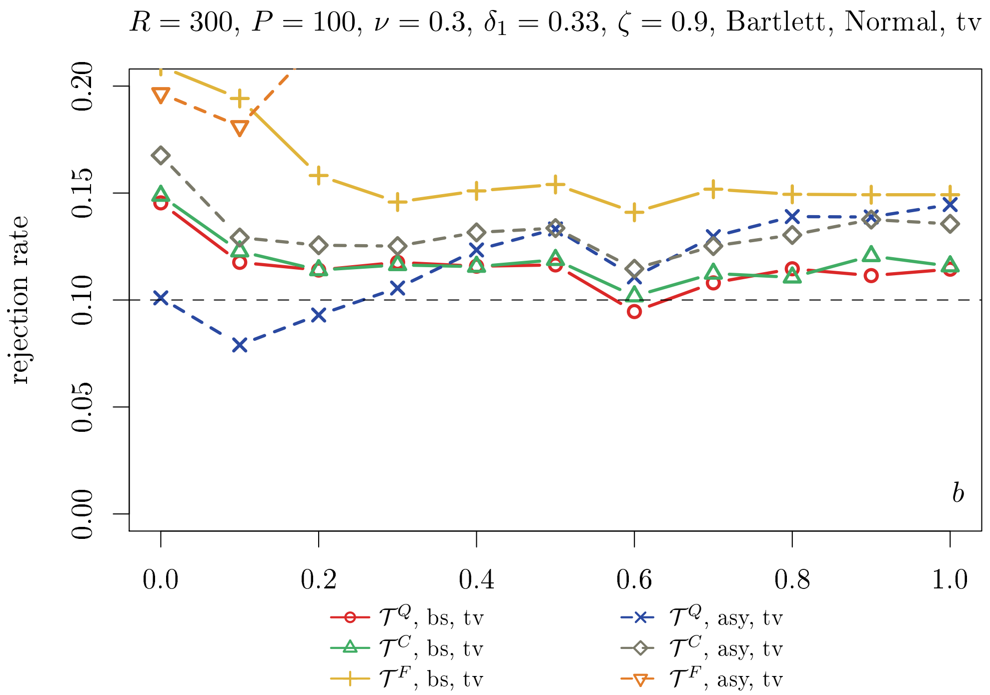
See notes to Figure 3.

Figure 26: Size under homoskedasticity, asymptotic and bootstrap tests, normal bootstrap errors, time-varying



See notes to Figure 3.

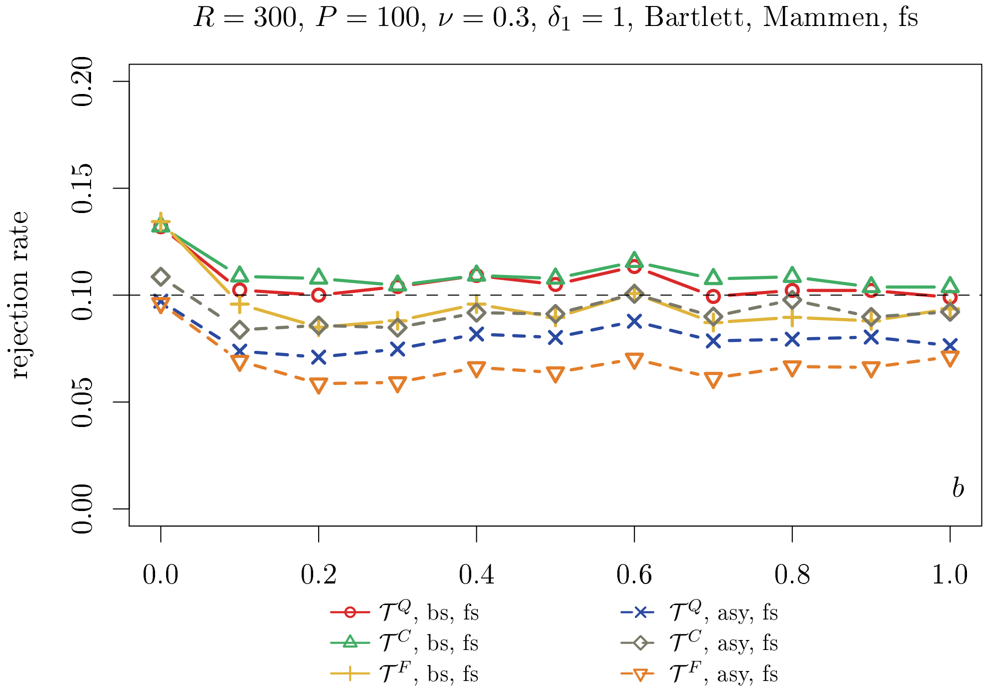
Figure 27: Size under heteroskedasticity, asymptotic and bootstrap tests, normal bootstrap errors, full-sample



$R = 300,\ P = 100,\ \nu = 0.3,\ \delta_1 = 0.33,\ \zeta = 0.9$, Bartlett, Normal, tv

See notes to Figure 3.

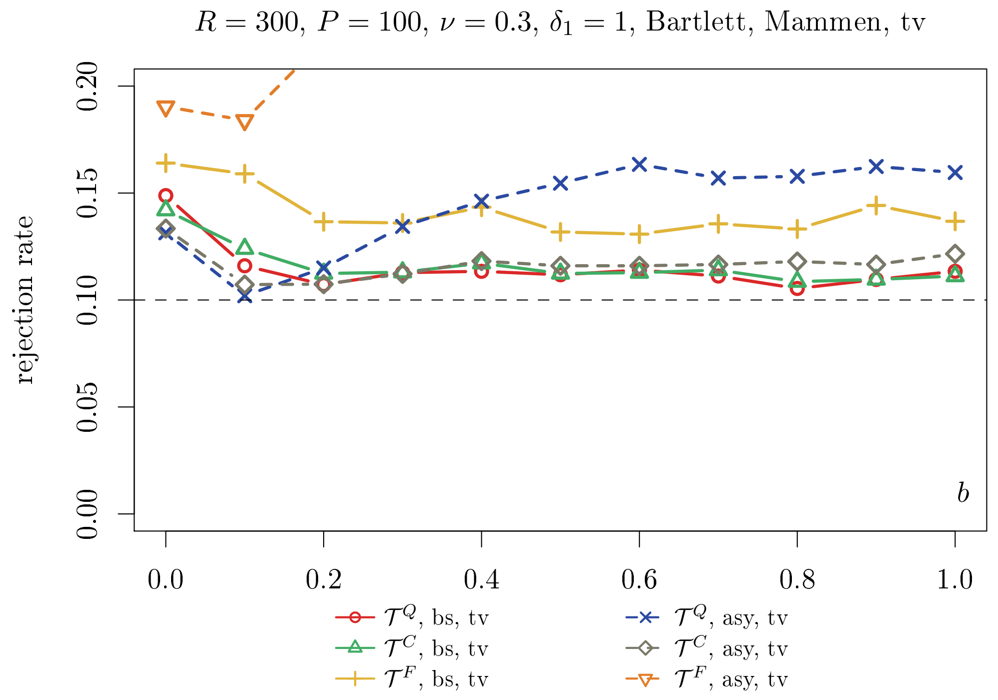
Figure 28: Size under heteroskedasticity, asymptotic and bootstrap tests, normal bootstrap errors, time-varying



$R = 300,\ P = 100,\ \nu = 0.3,\ \delta_1 = 1$, Bartlett, Mammen, fs

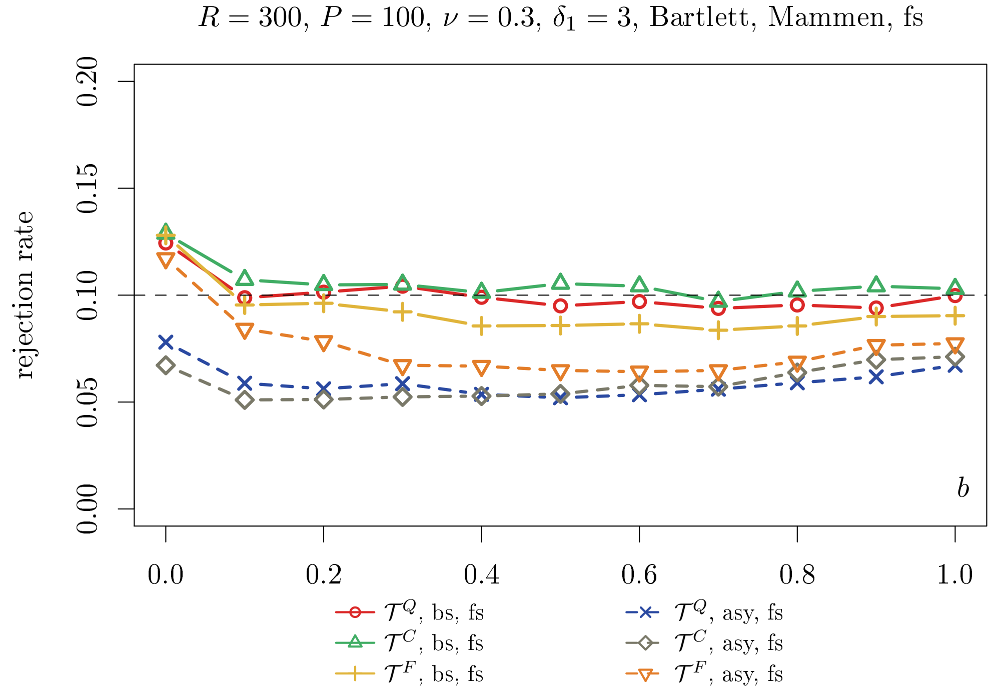
See notes to Figure 3.

Figure 29: Size under homoskedasticity, asymptotic and bootstrap tests, equal relative predictability, full-sample



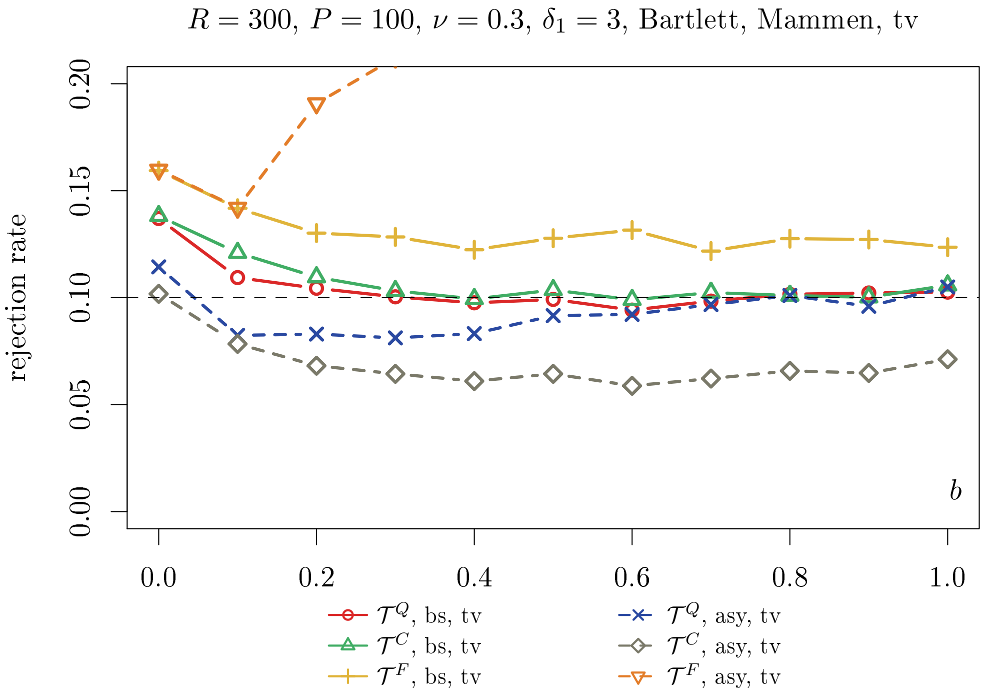
See notes to Figure 3.

Figure 30: Size under homoskedasticity, asymptotic and bootstrap tests, equal relative predictability, time-varying



See notes to Figure 3.

Figure 31: Size under heteroskedasticity, asymptotic and bootstrap tests, equal relative predictability, full-sample



See notes to Figure 3.

Figure 32: Size under heteroskedasticity, asymptotic and bootstrap tests, equal relative predictability, time-varying

G Imputation

This appendix contains details on the imputed values for the missing observations in the SPF data set from the "Forecast Error Statistics for the Survey of Professional Forecasters" obtained from the Federal Reserve Bank of Philadelphia.

A few missing values in the SPF series and the are imputed via a bootstrap based expectation maximization [EM] algorithm, see Honaker et al. (2011). The algorithm makes use of the standard EM algorithm on multiple bootstrapped samples of the original data set (containing missing values) to obtain imputed values. We use 10,000 bootstrap replications for the EM algorithm. The code is written in `R` (by using the `Amelia` package) and available upon request from the authors. Tables 6–7 contain the imputed values (underlined) in connection to neighboring values. The obtained bootstrap averages serve as imputed values which are plausible.

Table 6: Data entries for the first release of RGDP and PGDP series. #MV gives the number of missing values in total. For underlined dates imputed values are obtained from the bootstrap-based EM algorithm. Neighboring values are reported for comparison.

Date	RGDP	PGDP
1995Q3	4.20481	0.58927
1995Q4	2.41452	2.26685
1996Q1	2.80932	2.60573
#MV	1	1

Table 7: Data entries for four-quarters ahead SPF forecasts. #MV gives the number of missing values in total. For underlined dates imputed values are obtained from the bootstrap-based EM algorithm. Neighboring values are reported for comparison.

Date	RGDP	PGDP
1969Q4	4.03701	3.21260
1970Q1	3.55115	3.56122
1970Q2	3.90855	3.56355
1970Q3	4.05961	3.90631
1970Q4	3.10037	3.01866
1971Q1	4.54798	4.15267
1971Q2	4.26233	2.95183
1975Q2	5.40498	3.50332
1975Q3	5.33554	6.52413
1975Q4	5.02638	6.57499
#MV	5	5

H Additional empirical results - rolling window

This appendix contains additional empirical results based on rolling window estimation. First, it reports evaluations against the final release, starting with summary statistics. Next, full-sample and time-variation test results are given. The appendix ends with plots of forecast error loss differentials and graphs for the analysis of time-variation in the relative forecast performance.

Table 8: Summary statistics for output growth (RGDP) and GDP deflator inflation (PGDP) using the final data release. RelLoss denotes the relative root mean squared error loss of the competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve); SD(·) labels the standard deviation of the loss differentials in the subsample I (1969-1984), II (1985-2006) or III (2007-2017). AC(1) denotes the empirical first-order autocorrelation coefficient of the loss differential series.

Statistic		RelLoss	SD(I)	SD(II)	SD(III)	AC(1)
Sample		1969-2017	1969-1984	1985-2006	2007-2017	1969-2017
RGDP - NC/SPF						
	$h = 0$	1.54	44.66	5.83	8.05	0.11
	$h = 1$	1.39	51.28	7.42	10.57	0.21
	$h = 4$	1.41	62.13	10.63	18.70	0.28
RGDP - TMS/SPF						
	$h = 0$	1.34	22.65	5.34	15.45	0.21
	$h = 1$	1.05	18.82	5.63	9.18	0.20
	$h = 4$	1.01	24.03	3.82	2.68	0.03
PGDP - NC/SPF						
	$h = 0$	1.35	4.79	1.44	2.30	0.22
	$h = 1$	1.18	9.04	1.61	2.29	0.14
	$h = 4$	1.08	15.47	2.55	2.12	0.33
PGDP - PC/SPF						
	$h = 0$	1.35	4.46	1.44	1.82	0.09
	$h = 1$	1.22	10.05	1.62	2.12	0.11
	$h = 4$	1.26	20.93	2.48	2.51	0.43

Table 9: Test decisions for the full-sample $\mathcal{T}^{DM}$-statistic for equal predictive ability of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve) - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the final data release. Evaluation sample runs from 1969Q4 to 2017Q2.

	RGDP - NC/SPF						RGDP - TMS/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	***	***	***	***	***				
0.1	***	***	***	***	***	**	***	***				
0.2	***	**	***	**	***	**	***	**				
0.3	***	*	***	*	**	*	**	**				
0.4	***	*	**	*	**	*	**	**				
0.5	***	*	**	*	**	*	**	*				
0.6	***	*	**	*	**	*	**	*				
0.7	**		**		**		**	*				
0.8	**		**		**		**	*				
0.9	**		**		**		**	*				
1	**		**		**		**	*				

	PGDP - NC/SPF						PGDP - PC/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	**	*			***	***	***	**	***	**
0.1	***	***	***	**			***	***	***	**	***	**
0.2	***	***	***	**	*		***	**	***	**	***	*
0.3	***	***	***	**	**	*	***	**	***	**	***	*
0.4	***	***	***	***	**	*	***	*	***	**	***	*
0.5	***	***	***	***	**	*	**	*	***	**	***	*
0.6	***	***	***	***	***	*	**	*	***	**	***	*
0.7	***	***	***	***	***	*	**	*	***	**	***	
0.8	***	***	***	***	***	*	**	*	***	**	***	
0.9	***	***	***	***	***	*	**	*	***	**	***	
1	***	***	***	***	***	*	**	*	***	**	**	*

Table 10: Test decisions for the time-variation $\mathcal{T}^{\{Q,C,F\}}$-statistics for time-variation in the predictive ability of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve) - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the final data release. Evaluation sample runs from 1969Q4 to 2017Q2.

RGDP - NC/SPF

	$h=0$						$h=1$						$h=4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	
0.1	***	**	***	***	***	***	***	**	***	***	***	***	***	**	***	***	***	
0.2	**	*	***	**	**	*	**		***	**	*	*	**	*	**	**	**	
0.3	**		**	*	*		**		**	*			*		**	*		
0.4	**		**	*			*		**	*			*		**	*		
0.5	**		**				*		**						*			
0.6	*		**						**						*			
0.7	*		**						*						*			
0.8	*		**						**									
0.9	*		**						*						*			
1	*		**						*						*			

RGDP - TMS/SPF

	$h=0$						$h=1$						$h=4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***												
0.1	***	**	***	***	***	***												
0.2	**	*	**	**	**	**												
0.3	*		**	**	*													
0.4	*		*	*														
0.5			*	*														
0.6			*	*														
0.7			*	*									*					
0.8			*	*									*					
0.9			*	*									*					
1			*	*									*					

Table 11: continued from Table 10.

PGDP - NC/SPF

	$h=0$						$h=1$						$h=4$					
b	$\mathcal{T}^{Q}_{\text{bs}}$	$\mathcal{T}^{Q}_{\text{asy}}$	$\mathcal{T}^{C}_{\text{bs}}$	$\mathcal{T}^{C}_{\text{asy}}$	$\mathcal{T}^{F}_{\text{bs}}$	$\mathcal{T}^{F}_{\text{asy}}$	$\mathcal{T}^{Q}_{\text{bs}}$	$\mathcal{T}^{Q}_{\text{asy}}$	$\mathcal{T}^{C}_{\text{bs}}$	$\mathcal{T}^{C}_{\text{asy}}$	$\mathcal{T}^{F}_{\text{bs}}$	$\mathcal{T}^{F}_{\text{asy}}$	$\mathcal{T}^{Q}_{\text{bs}}$	$\mathcal{T}^{Q}_{\text{asy}}$	$\mathcal{T}^{C}_{\text{bs}}$	$\mathcal{T}^{C}_{\text{asy}}$	$\mathcal{T}^{F}_{\text{bs}}$	$\mathcal{T}^{F}_{\text{asy}}$
0	***	***	**	***	**	***												
0.1	***	***	***	***	***	***	***	**	*	*	***	***						
0.2	***	***	***	***	***	***	***	**	**	**	***	***						
0.3	***	**	***	***	***	**	***	**	**	**	***	***					*	
0.4	***	**	***	**	**	**	***	**	***	**	***	***					*	
0.5	***	**	***	**	**	**	***	**	***	**	***	***					**	
0.6	***	**	***	**	**	**	***	**	***	**	***	***	*		*		**	
0.7	***	**	***	***	**	**	***	**	***	**	***	***	*		**		**	
0.8	***	**	***	***	**	**	***	**	***	**	***	***			**		**	
0.9	***	**	***	**	**	**	***	**	***	**	***	***	*		**		**	
1	***	**	***	***	***	**	***	**	***	**	***	***	*		**		**	

PGDP - PC/SPF

	$h=0$						$h=1$						$h=4$					
b	$\mathcal{T}^{Q}_{\text{bs}}$	$\mathcal{T}^{Q}_{\text{asy}}$	$\mathcal{T}^{C}_{\text{bs}}$	$\mathcal{T}^{C}_{\text{asy}}$	$\mathcal{T}^{F}_{\text{bs}}$	$\mathcal{T}^{F}_{\text{asy}}$	$\mathcal{T}^{Q}_{\text{bs}}$	$\mathcal{T}^{Q}_{\text{asy}}$	$\mathcal{T}^{C}_{\text{bs}}$	$\mathcal{T}^{C}_{\text{asy}}$	$\mathcal{T}^{F}_{\text{bs}}$	$\mathcal{T}^{F}_{\text{asy}}$	$\mathcal{T}^{Q}_{\text{bs}}$	$\mathcal{T}^{Q}_{\text{asy}}$	$\mathcal{T}^{C}_{\text{bs}}$	$\mathcal{T}^{C}_{\text{asy}}$	$\mathcal{T}^{F}_{\text{bs}}$	$\mathcal{T}^{F}_{\text{asy}}$
0	***	***	***	***	***	***	**			*	**	**	**		*	**		*
0.1	***	***	***	***	***	***	***	**	***	**	***	***	**		**	**		*
0.2	***	**	***	***	***	***	***	*	***	**	***	***	**		***	**		*
0.3	**	*	***	**	**	*	***	*	***	**	**	**	*		**	*		
0.4	**		**	*	*		**		***	*	*	**			**			
0.5	**		**	*			**		***	*	*	*			**			
0.6	*		**	*			**		***	*	*	*			**			
0.7	*		**	*			**		***	*	*	*			**			
0.8	*		**	*			**		***	*	*	**			**			
0.9	*		**	*			**		***	**	*	**			**			
1	*		**	*			**		***	**	*	**			**			

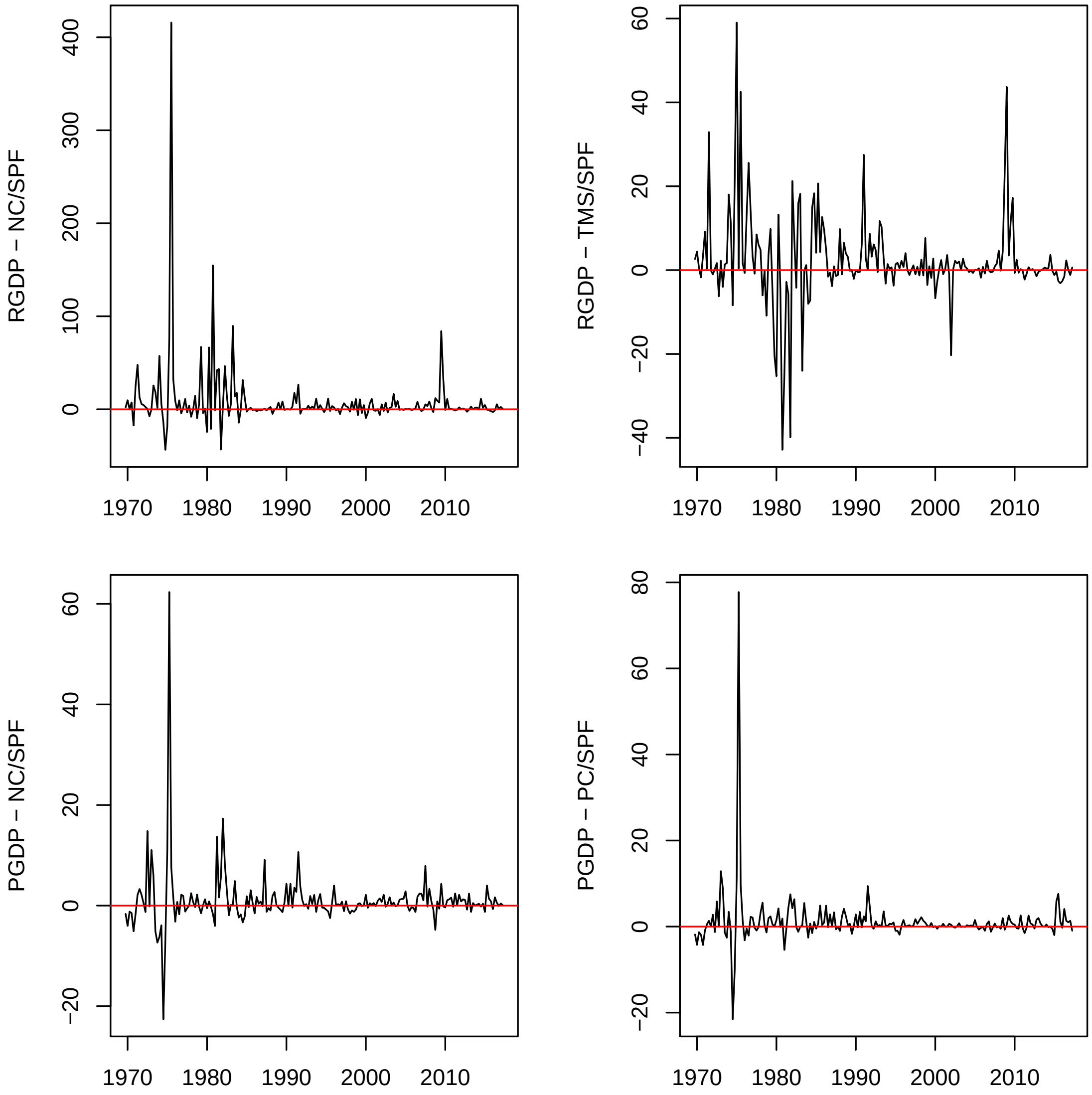


Figure 33: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve). One-quarter ahead forecasts ($h = 1$) are evaluated against the first release.

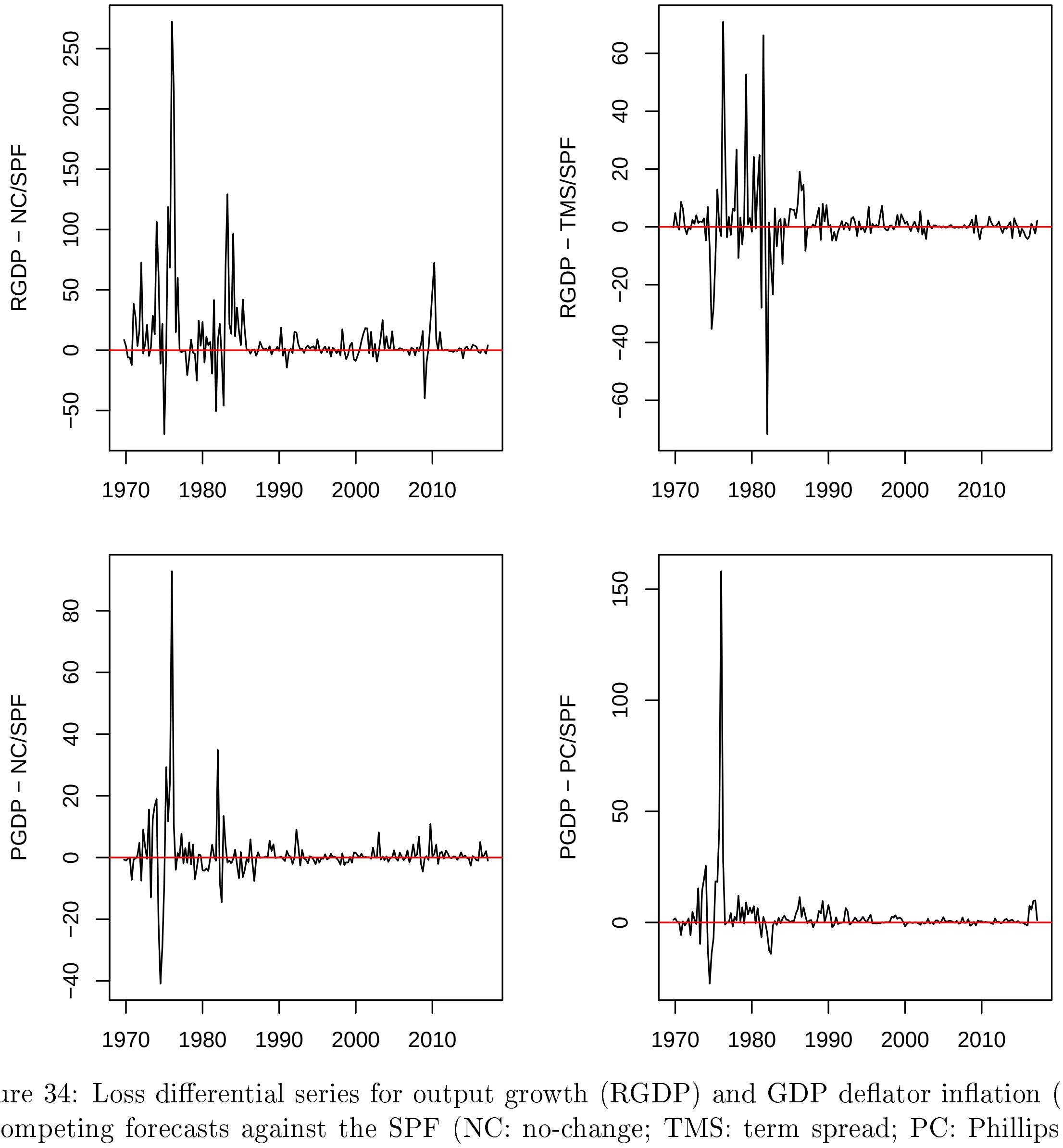


Figure 34: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve). One-year ahead forecasts ($h = 4$) are evaluated against the first release.

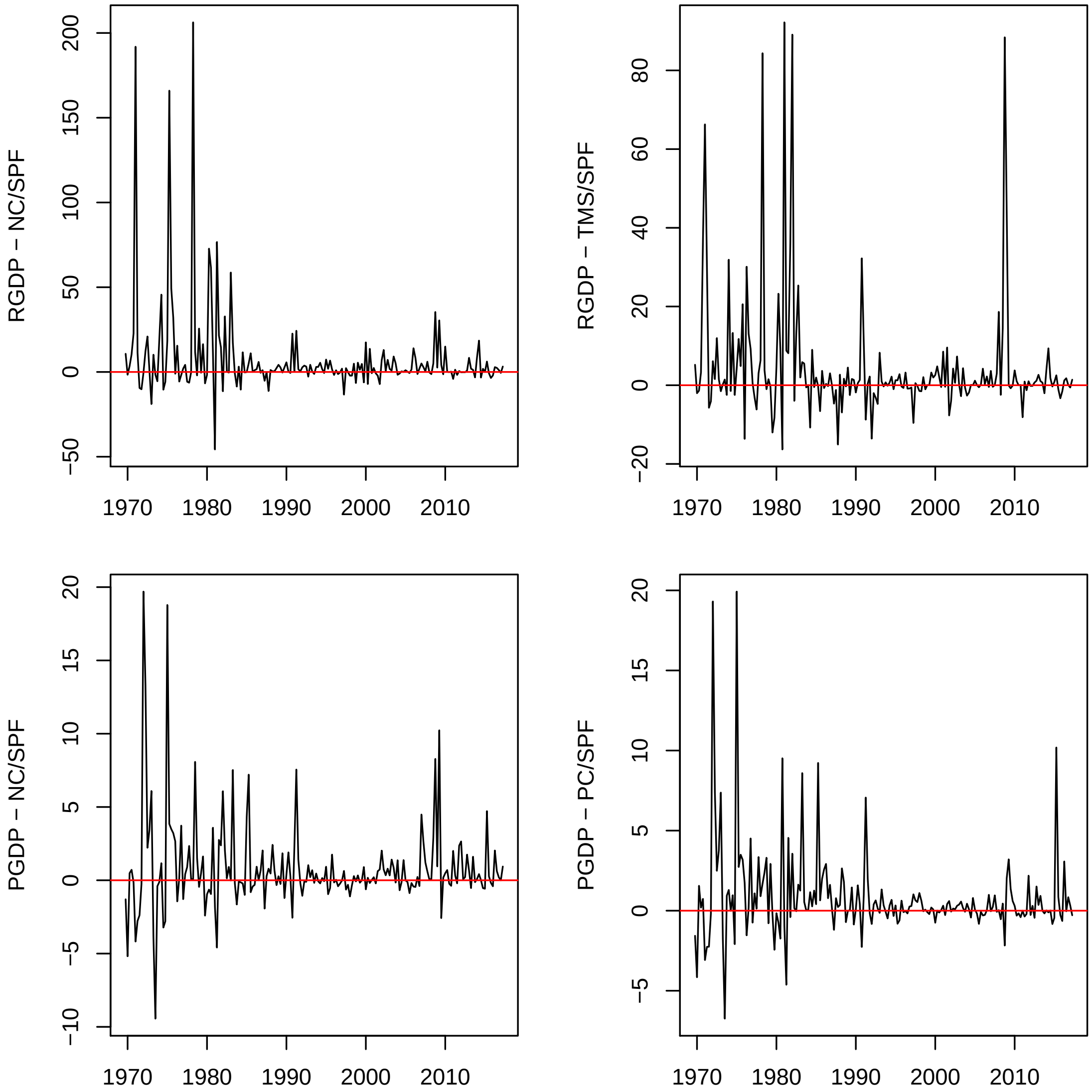


Figure 35: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve). Nowcasts ($h = 0$) are evaluated against the final release.

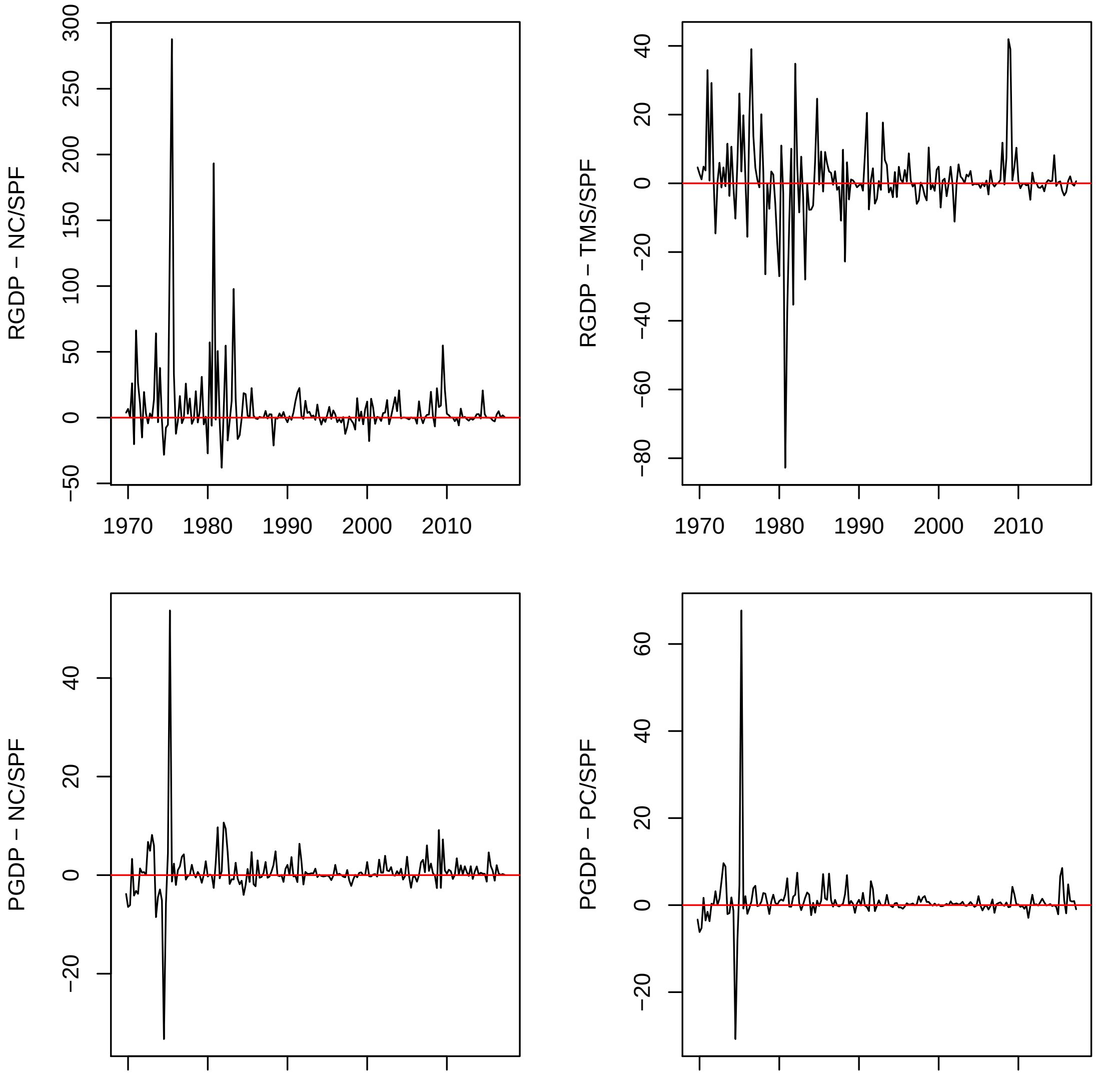


Figure 36: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve). One-quarter ahead forecasts ($h = 1$) are evaluated against the final release.

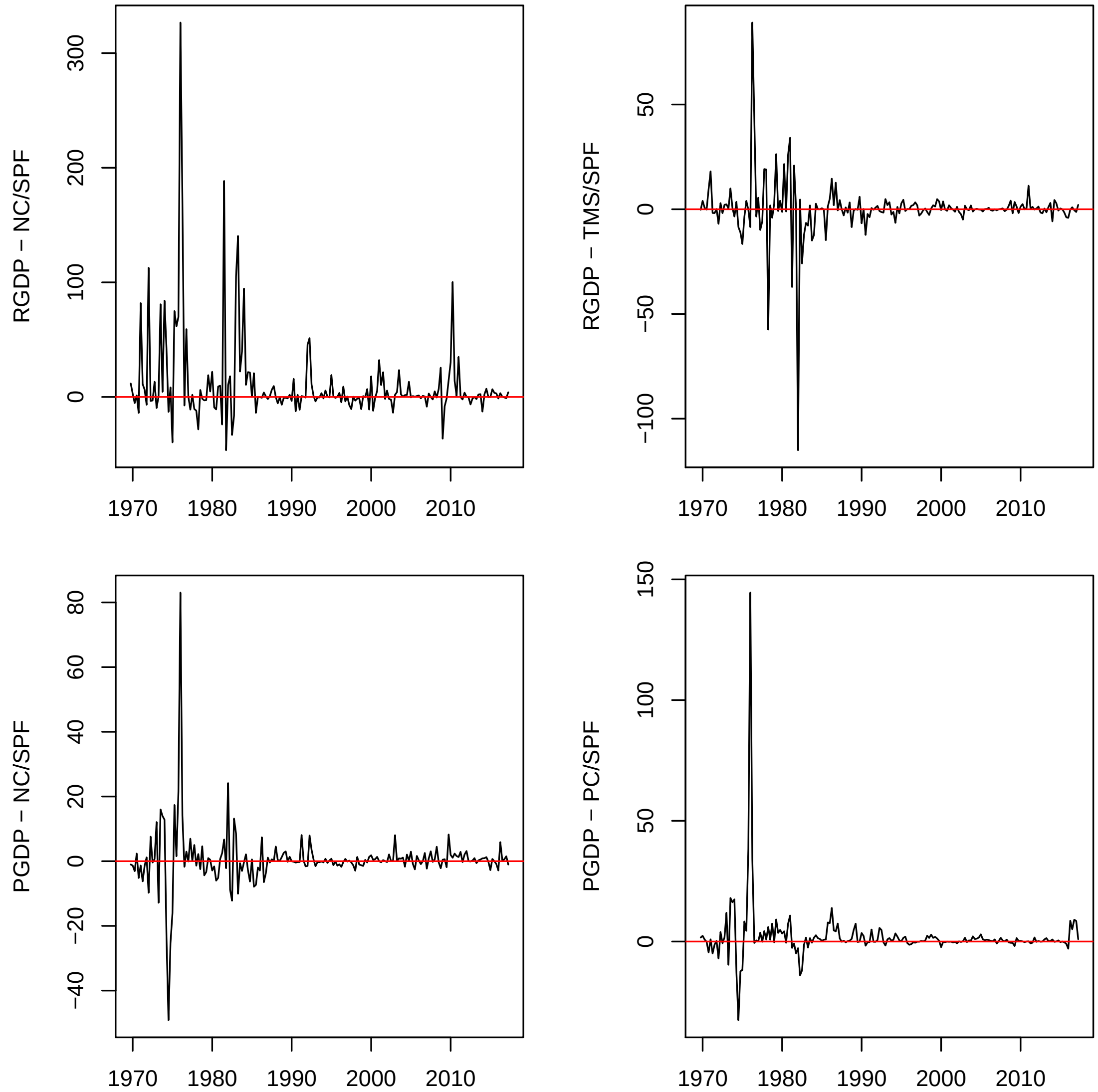


Figure 37: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve). One-year ahead forecasts ($h = 4$) are evaluated against the final release.

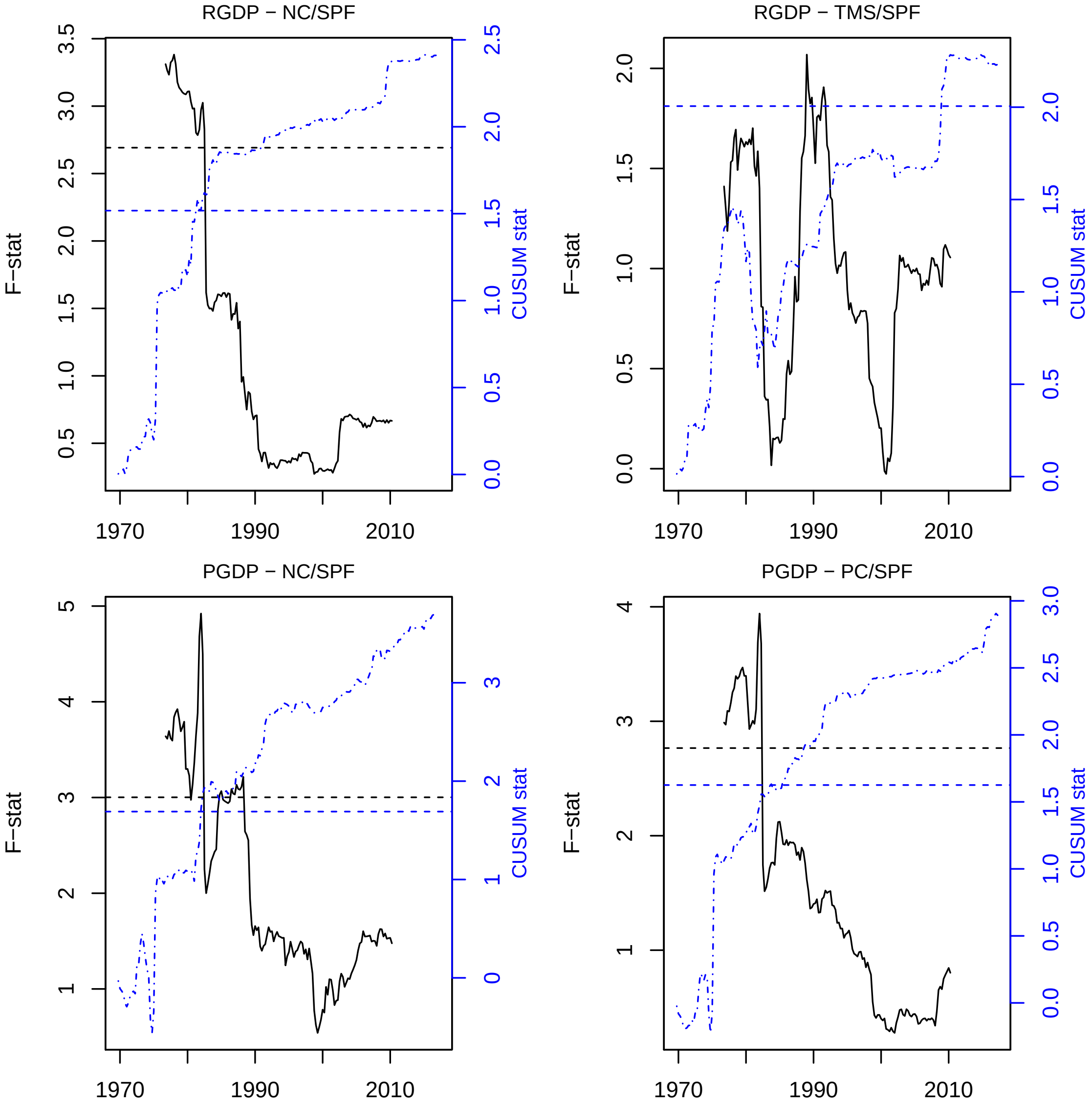


Figure 38: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. One-quarter ahead forecasts ($h = 1$) are evaluated against the first release; $b = 0.2$, $\nu = 0.3$.

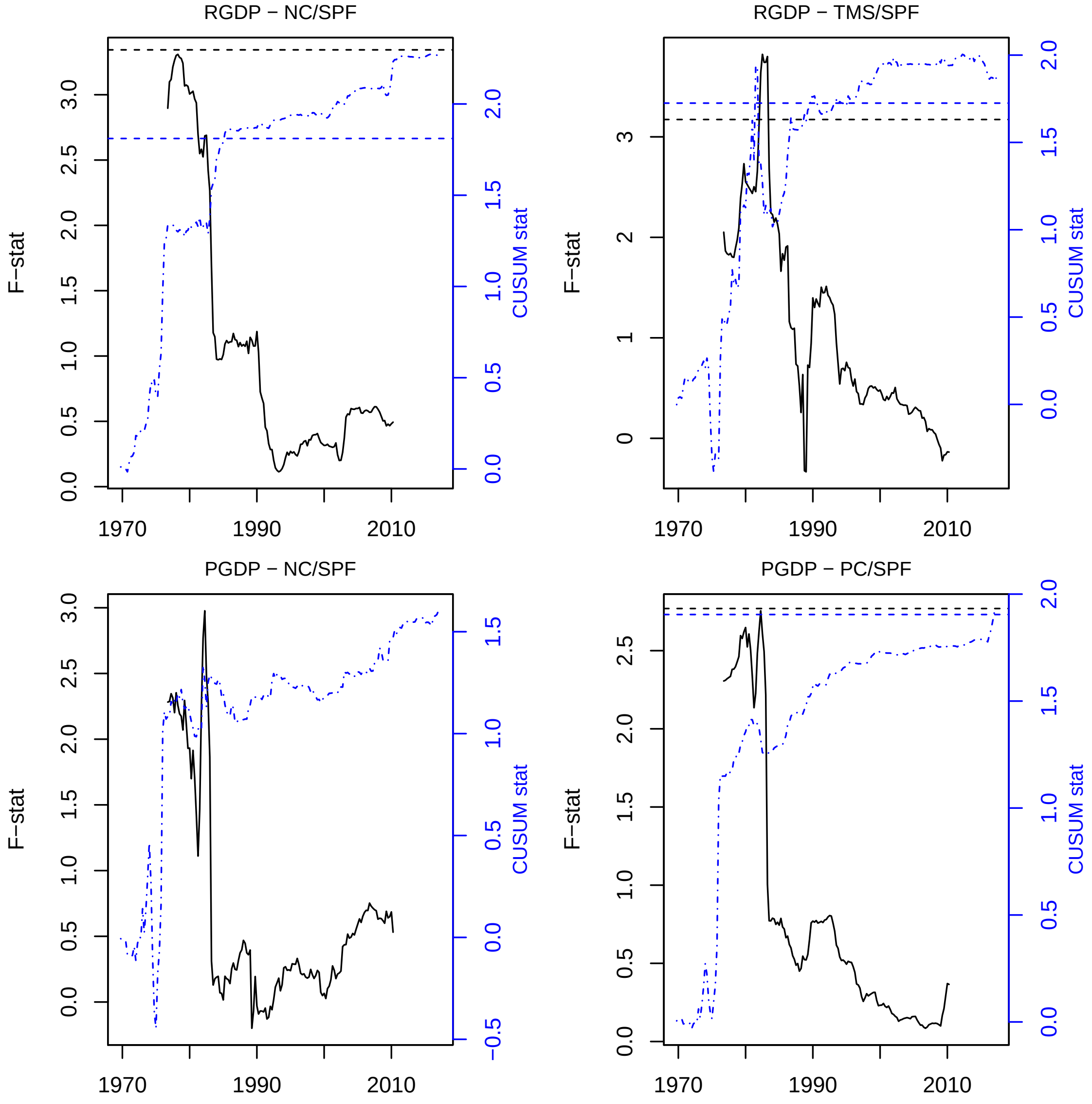


Figure 39: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. One-year ahead forecasts ($h = 4$) are evaluated against the first release; $b = 0.2$, $\nu = 0.3$.

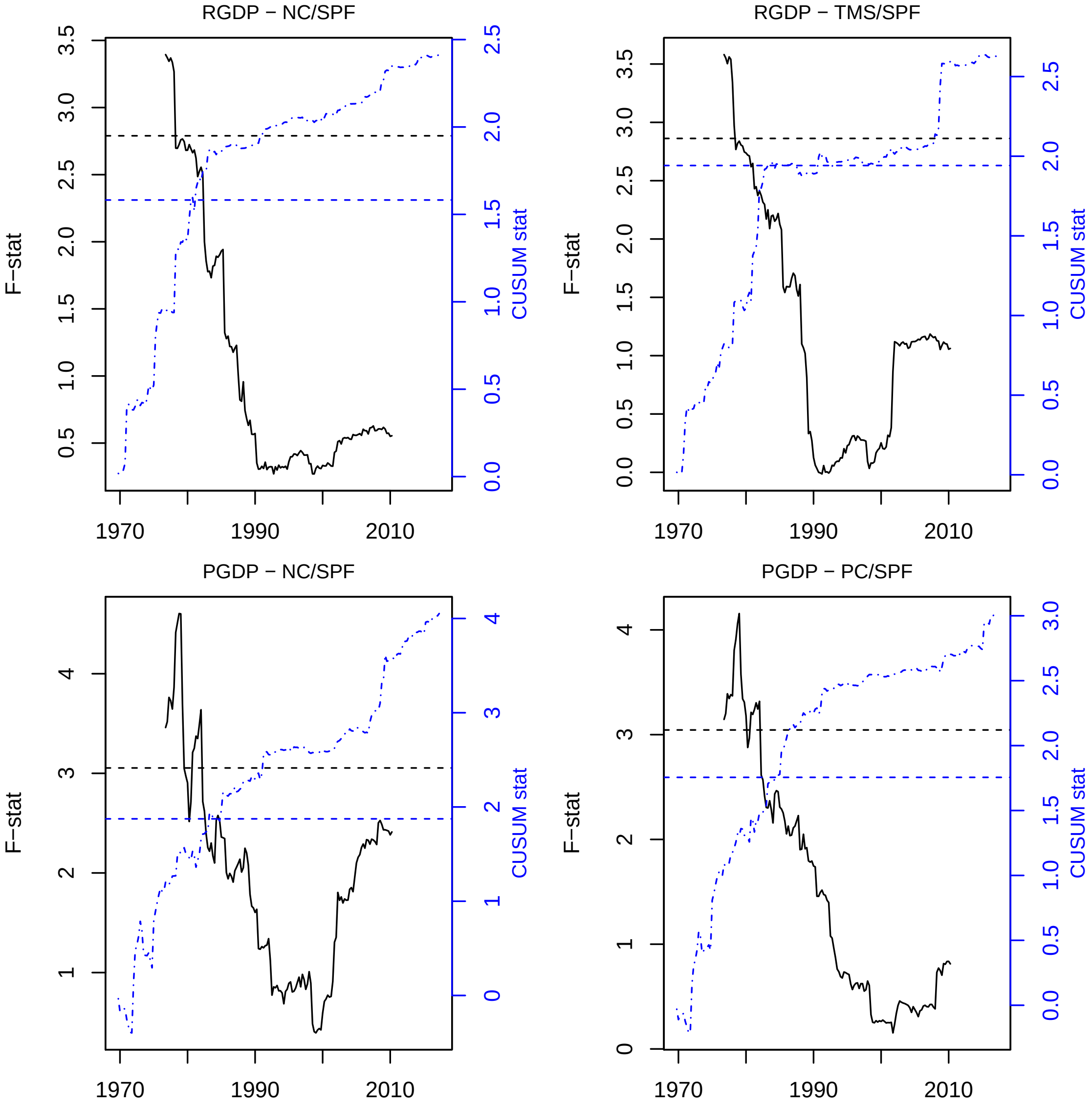


Figure 40: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. Nowcasts ($h = 0$) are evaluated against the final release; $b = 0.2$, $\nu = 0.3$.

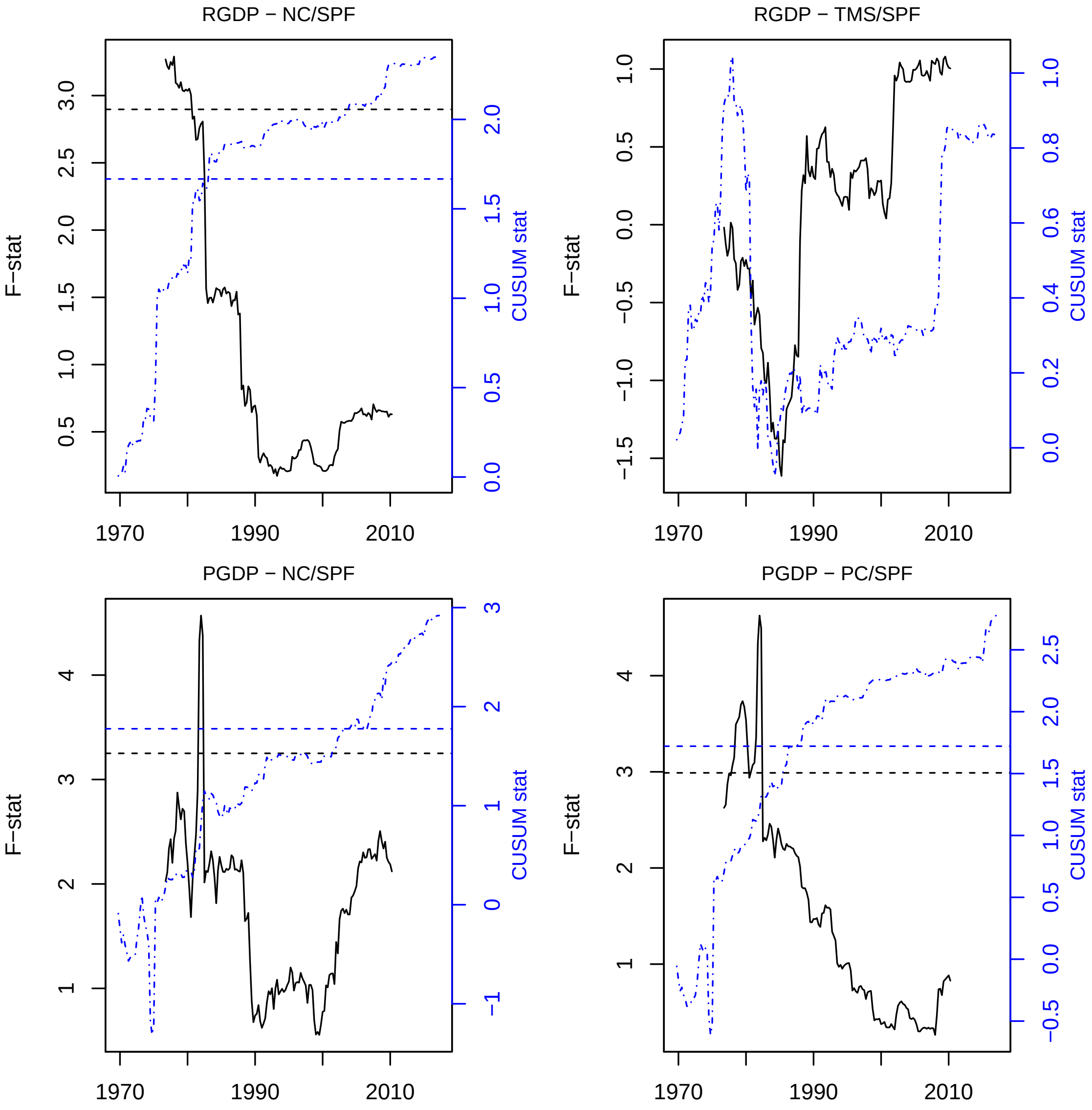


Figure 41: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. One-quarter ahead forecasts ($h = 1$) are evaluated against the final release; $b = 0.2$, $\nu = 0.3$.

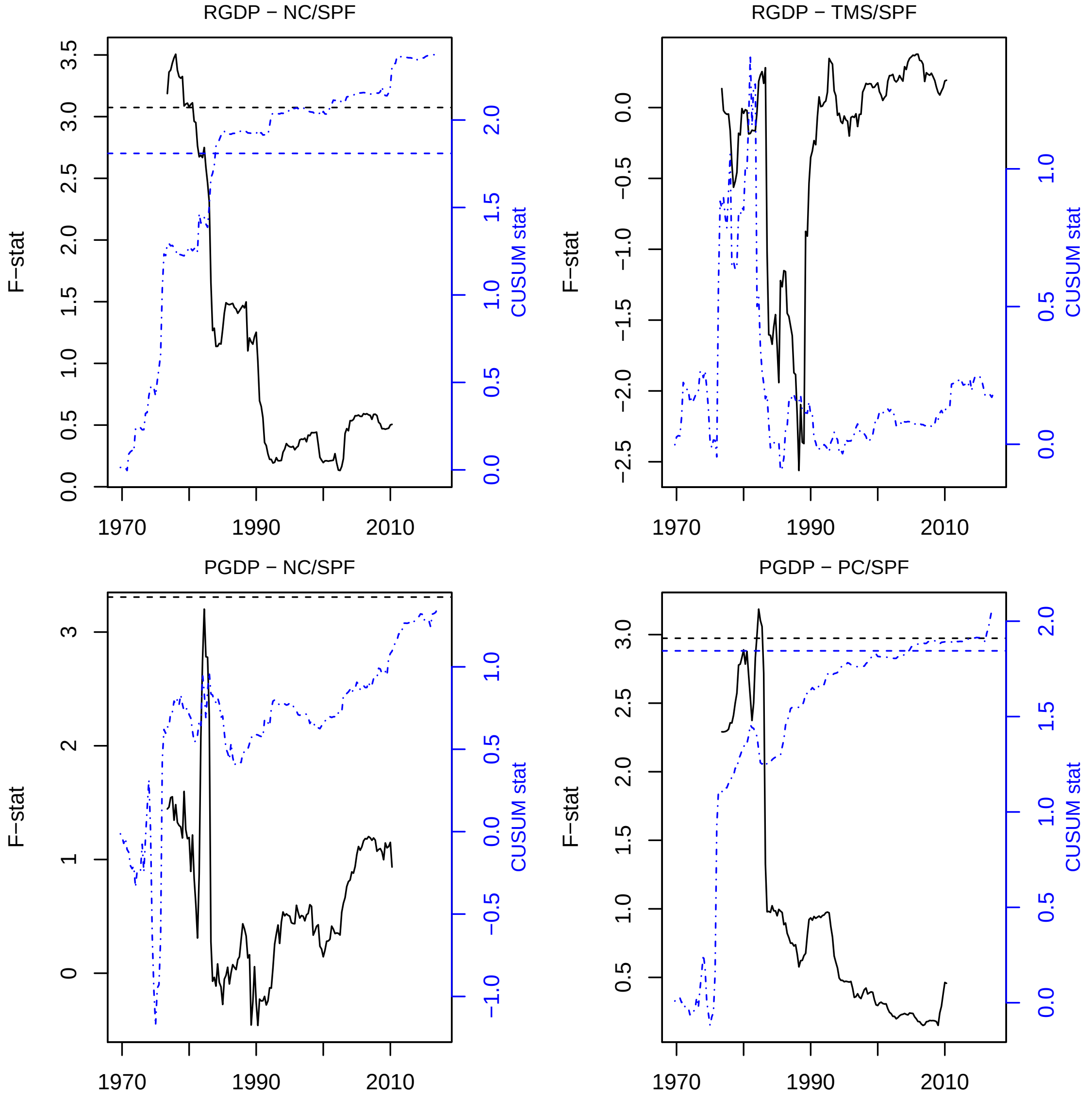


Figure 42: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. One-year ahead forecasts ($h = 4$) are evaluated against the final release; $b = 0.2$, $\nu = 0.3$.

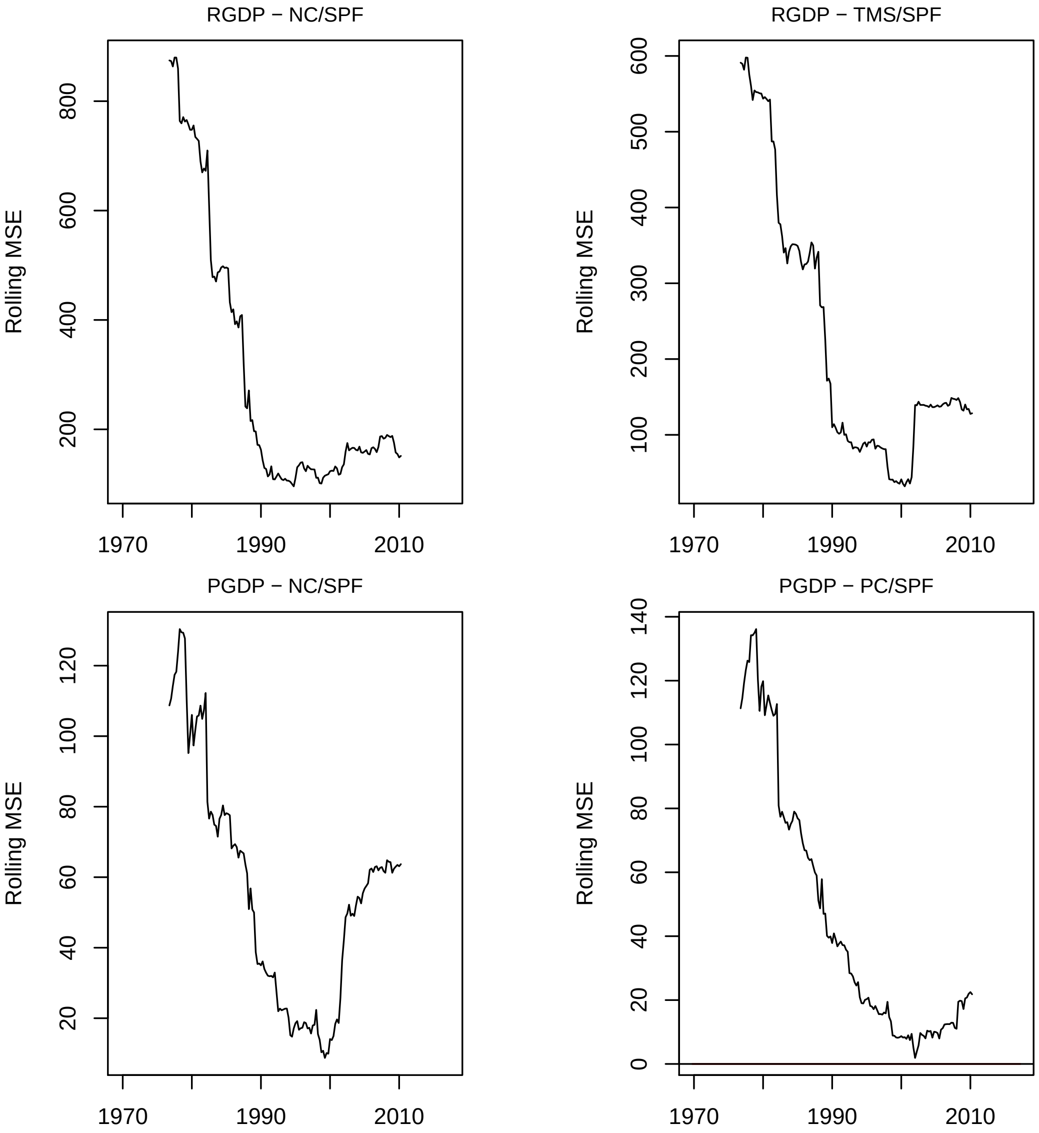


Figure 43: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). Nowcasts ($h = 0$) are evaluated against the first release.

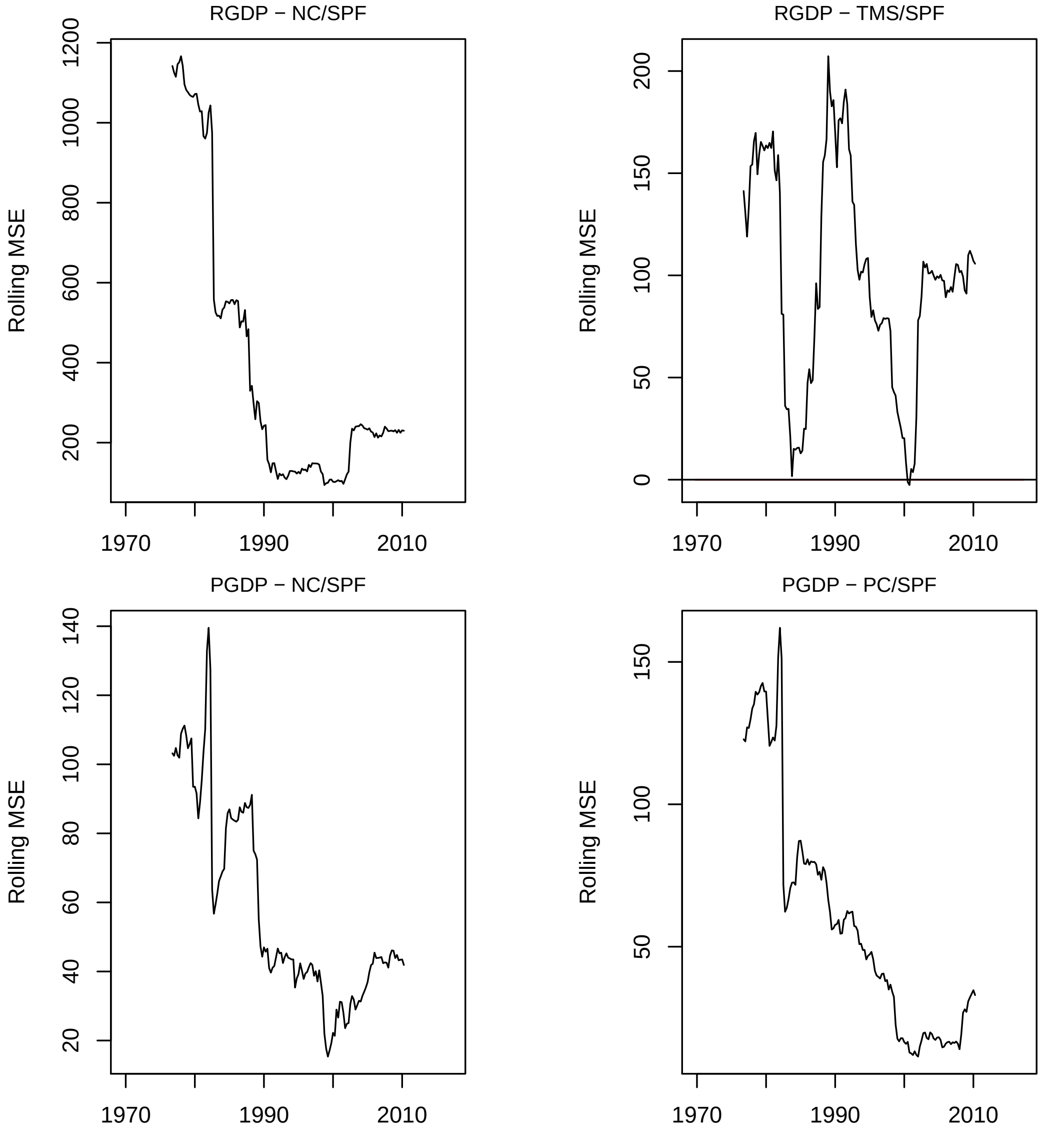


Figure 44: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). One-quarter ahead forecasts ($h = 1$) are evaluated against the first release.

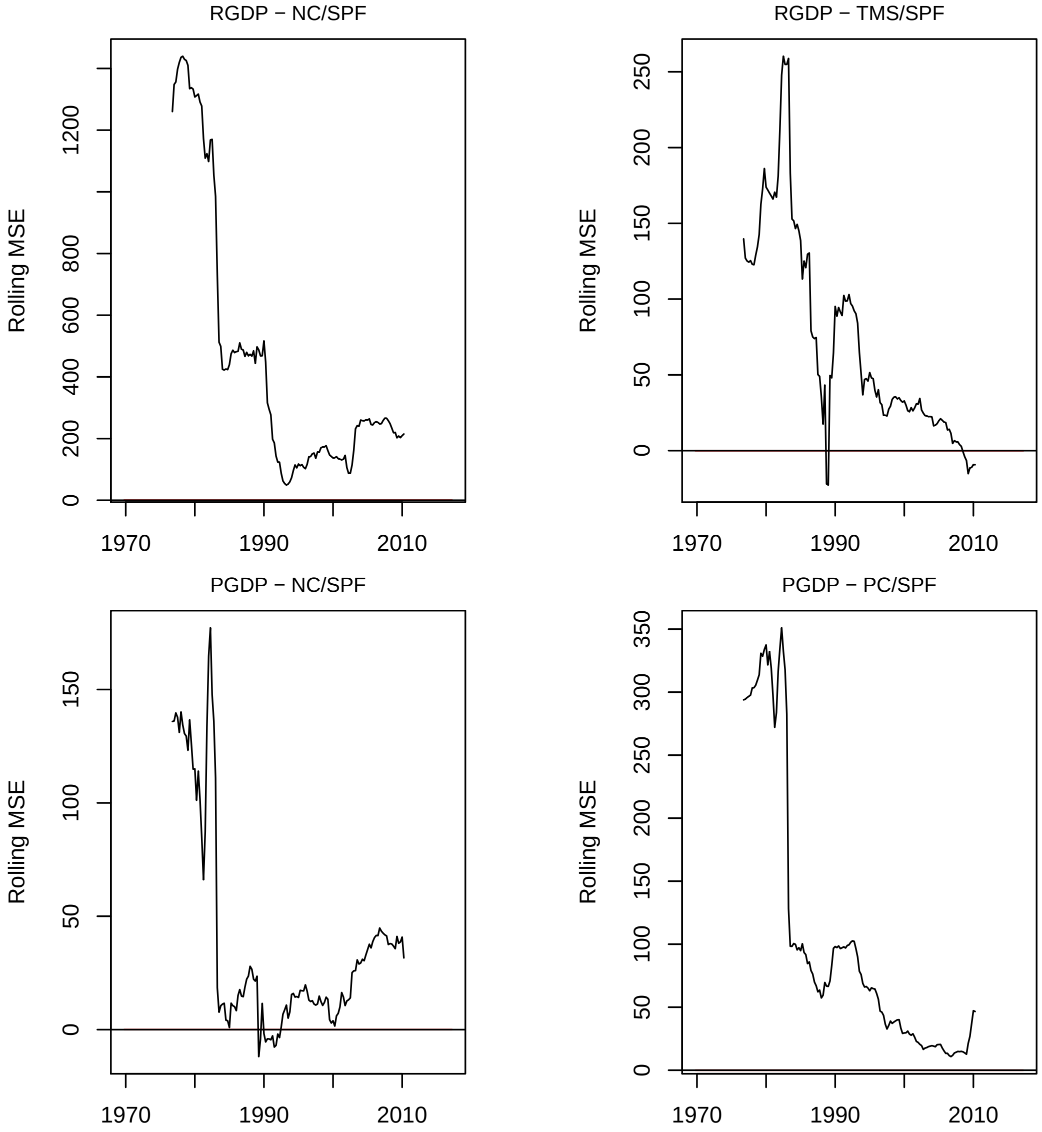


Figure 45: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). One-year ahead forecasts ($h = 4$) are evaluated against the first release.

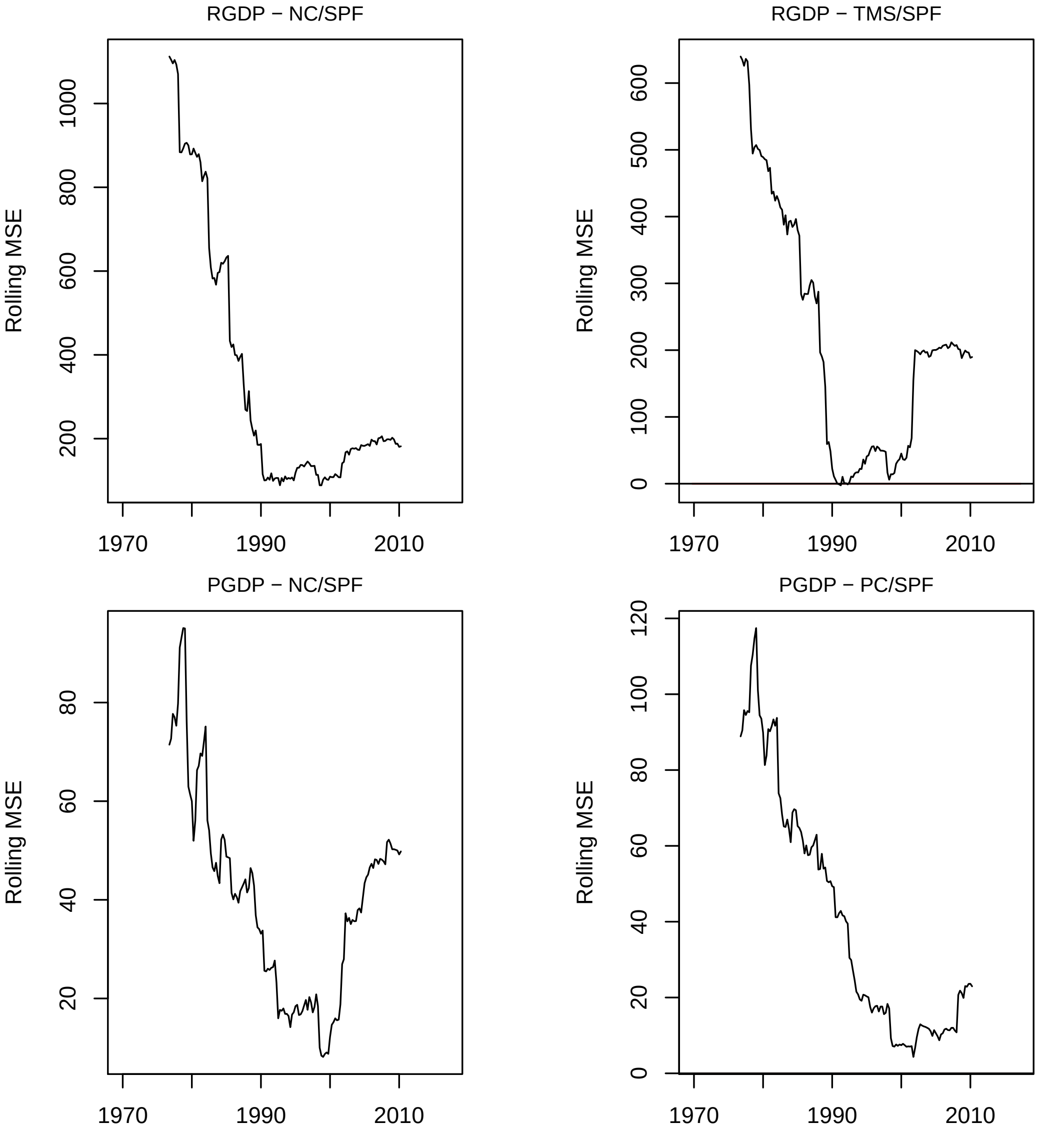


Figure 46: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). Nowcasts ($h = 0$) are evaluated against the final release.

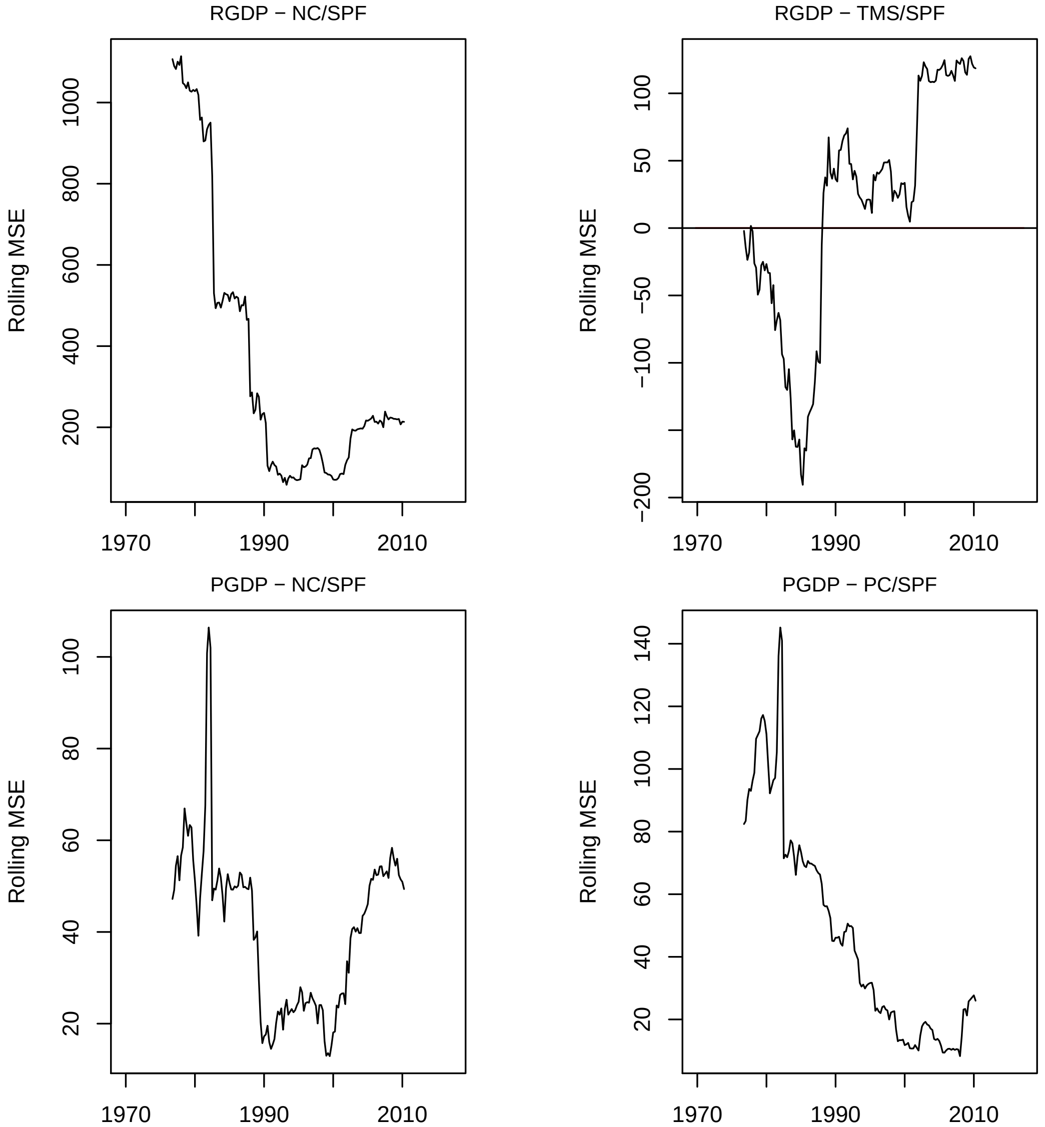


Figure 47: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). One-quarter ahead forecasts ($h = 1$) are evaluated against the final release.

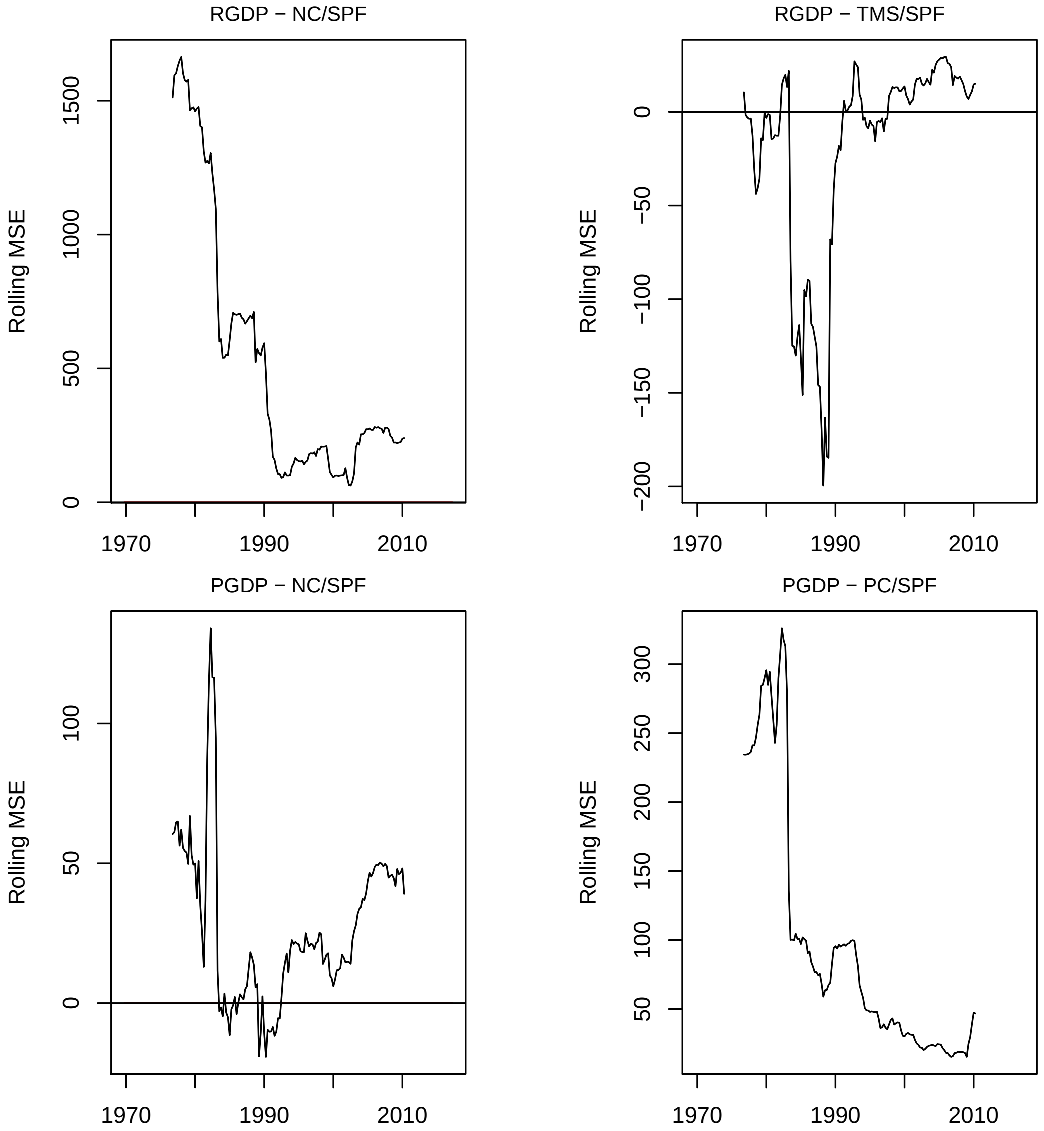


Figure 48: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). One-year ahead forecasts ($h = 4$) are evaluated against the final release.

I Additional empirical results - recursive estimation

This appendix contains additional empirical for recursive estimation results. First, it reports evaluations against the first and final release, starting with summary statistics. Next, full-sample and time-variation test results are given. The appendix ends with plots of forecast error loss differentials and graphs for the analysis of time-variation in the relative forecast performance.

Table 12: Summary statistics for output growth (RGDP) and GDP deflator inflation (PGDP) using the first and final data release. RelLoss denotes the relative root mean squared error loss of the competing term spread (TMS) and Phillips curve (PC) forecasts against the SPF; SD(·) labels the standard deviation of the loss differentials in the subsample I (1969-1984), II (1985-2006) or III (2007-2017). AC(1) denotes the empirical first-order autocorrelation coefficient of the loss differential series.

Statistic		RelLoss	SD(I)	SD(II)	SD(III)	AC(1)
Sample		1969-2017	1969-1984	1985-2006	2007-2017	1969-2017
RGDP (First) - TMS/SPF						
	$h = 0$	1.56	18.08	5.85	11.70	0.21
	$h = 1$	1.22	16.34	7.93	13.43	0.28
	$h = 4$	1.10	14.12	4.41	4.17	0.07
RGDP (Final) - TMS/SPF						
	$h = 0$	1.34	21.12	6.89	17.10	0.26
	$h = 1$	1.09	18.63	7.84	15.04	0.30
	$h = 4$	1.03	15.01	4.89	4.18	0.05
PGDP (First) - PC/SPF						
	$h = 0$	1.36	5.16	1.38	2.24	0.01
	$h = 1$	1.23	9.64	1.52	1.80	0.19
	$h = 4$	1.18	11.80	1.77	2.39	0.13
PGDP (Final) - PC/SPF						
	$h = 0$	1.34	4.08	1.21	2.23	0.10
	$h = 1$	1.19	8.77	1.35	2.13	0.06
	$h = 4$	1.16	10.52	1.94	1.87	0.17

Table 13: Test decisions for the full-sample $\mathcal{T}^{DM}$-statistic for equal predictive ability of competing term spread (TMS) and Phillips curve (PC) forecasts against the SPF - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the first and final data release. Evaluation sample runs from 1969Q4 to 2017Q2.

	RGDP (First) - TMS/SPF						RGDP (Final) - TMS/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	***	***	***	***	***				
0.1	***	***	***	**	***	***	***	***				
0.2	***	***	***	**	***	***	***	**				
0.3	***	**	***	***	***	***	***	**				
0.4	**	**	***	***	***	***	**	**				
0.5	**	**	***	***	***	**	**	**				
0.6	**	**	***	***	***	**	**	**				
0.7	**	**	***	***	***	**	**	**				
0.8	**	**	***	***	***	**	**	**				
0.9	**	**	***	***	***	**	**	**				
1	**	**	***	***	***	**	**	**				

	PGDP (First) - PC/SPF						PGDP (Final) - PC/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	***	***	***	***	***	***	**	***	***
0.1	***	***	***	***	***	**	***	***	***	**	***	**
0.2	***	***	***	**	***	*	***	**	***	**	**	*
0.3	***	**	***	**	***	*	***	**	***	**	**	*
0.4	***	**	***	**	**	*	***	**	***	**	**	*
0.5	***	**	***	**	**		**	**	***	**	**	
0.6	***	**	***	**	**		***	**	***	**	**	
0.7	***	**	***	**	**		***	**	***	**	**	
0.8	***	**	***	**	**		**	**	***	**	**	
0.9	***	**	***	**	**		**	**	***	**	**	
1	***	**	***	**	**		**	**	***	**	**	

Table 14: Test decisions for the time-variation $\mathcal{T}^{\{Q,C,F\}}$-statistics for time-variation in the predictive ability of competing term spread (TMS) and Phillips curve (PC) forecasts against the SPF - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the first and final data release. Evaluation sample runs from 1969Q4 to 2017Q2.

RGDP (First) - TMS/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	***	**	**	**	*		***	***	**	***		
0.1	***	***	***	***	***	***	**	**	*	*			***	**	***	***		
0.2	***	**	***	***	***	**	**	**	**	**	*	*	***	**	***	**		
0.3	**	**	**	**	*	*	***	***	***	***	***	**	***	**	***	**		
0.4	**	*	**	**			***	***	***	***	***	***	***	**	***	**		
0.5	*	*	**	**			***	***	***	***	***	**	***	**	***	**		
0.6	*		**	**			**	**	**	**	**	**	***	**	***	**		
0.7	*		**	**			***	***	***	***	***	**	***	**	***	**		
0.8	*		*	**			***	***	***	***	***	***	***	**	***	**		
0.9	*		**	**			***	***	***	***	***	***	***	**	***	**		
1	*		**	**			***	***	***	***	***	***	***	**	***	**		

RGDP (Final) - TMS/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***												
0.1	**	**	**	***	***	**												
0.2	**	**	**	**	**	**												
0.3	**	*	**	**	*	*												
0.4	*	*	**	**														
0.5	*	*	**	**														
0.6	*	*	*	**														
0.7	*		*	**														
0.8	*	*	*	**														
0.9	*		*	**														
1	*		*	**														

Table 15: continued from Table 14.

PGDP (First) - PC/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	***	**	**	**	*	*	***	**	**	***	**	***
0.1	***	***	***	***	***	***	***	**	***	***	**	**	**		**	**		*
0.2	***	**	***	**	***	**	***	**	***	**	*	**	*		*	*		
0.3	**	*	***	**	**	*	***	*	***	**	*	*	*		*	*		
0.4	**	*	**	**	*		**		***	**			*		*			
0.5	**	*	**	**	*		**		**	*					*			
0.6	**	*	**	**	*		**		***	*					*			
0.7	**	*	**	**	*		**		***	*					*			
0.8	**	*	**	**	*		**		**	*					*			
0.9	**	*	**	**	*		**		**	*					*			
1	**	*	**	**	*		**		***	*					*			

PGDP (Final) - PC/SPF

	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	**			*			**	**	**	***	**	***
0.1	***	***	***	***	***	***	***	*	**	**	**	**	*		**	**		*
0.2	***	**	***	**	***	**	**	*	**	**	**	**	*		*	*		
0.3	**	*	**	**	**	**	***	*	***	**	**	**			*	*		
0.4	**	*	**	**	*		**	*	**	**	*	**			*			
0.5	**		**	**	*		**	*	**	**	*	*			*			
0.6	**		**	*	*		**	*	***	**	*	*			*			
0.7	**		**	**	*		**	*	***	**	*	**			*			
0.8	**		**	**	*		***	*	***	**	**	**			*			
0.9	**		**	**	*		***	*	***	**	*	**			*			
1	**		**	**	*		**	*	***	**	*	**			*			

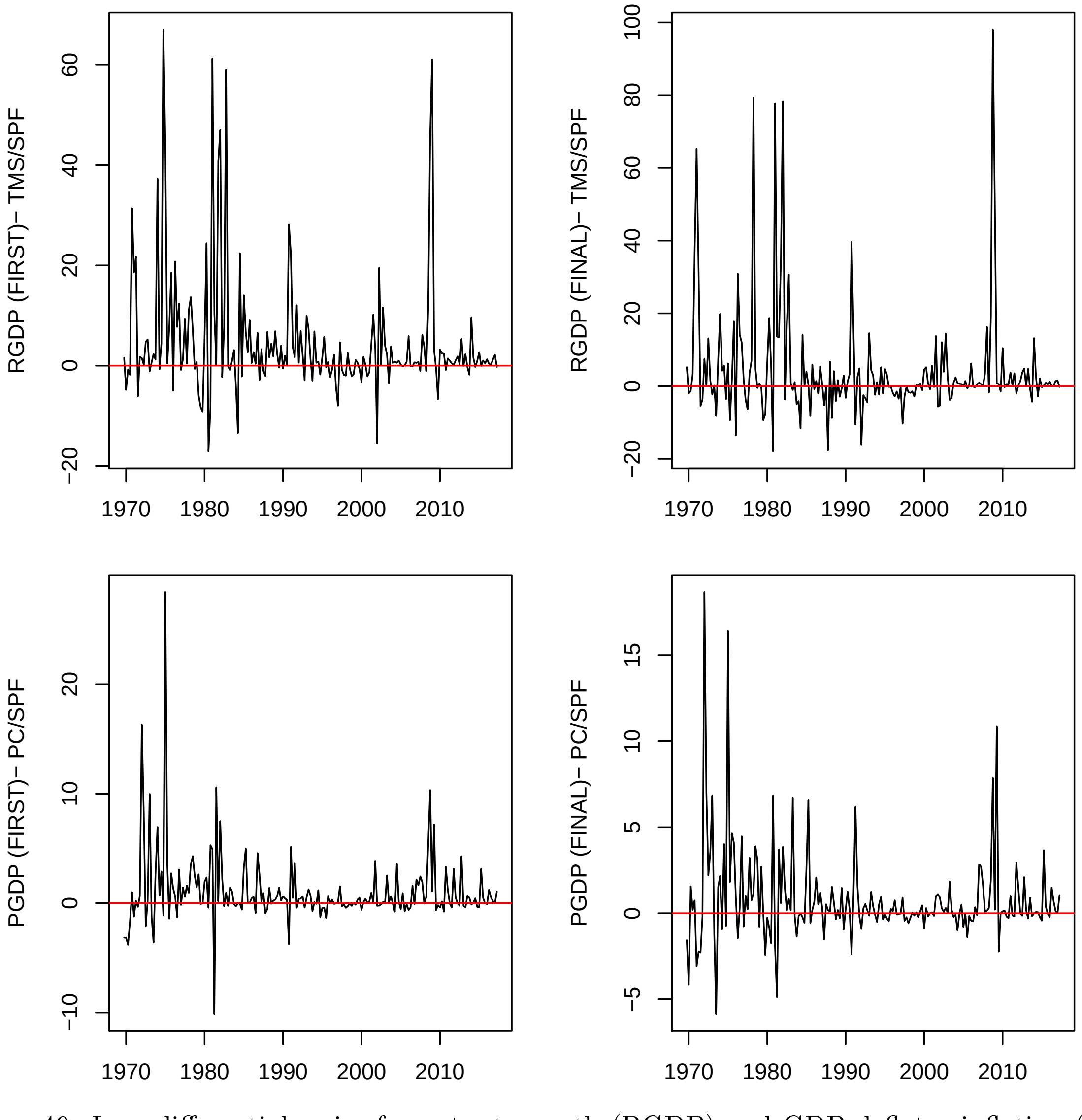


Figure 49: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing TMS forecasts against the SPF. Nowcasts ($h = 0$) are evaluated against the first and the final release.

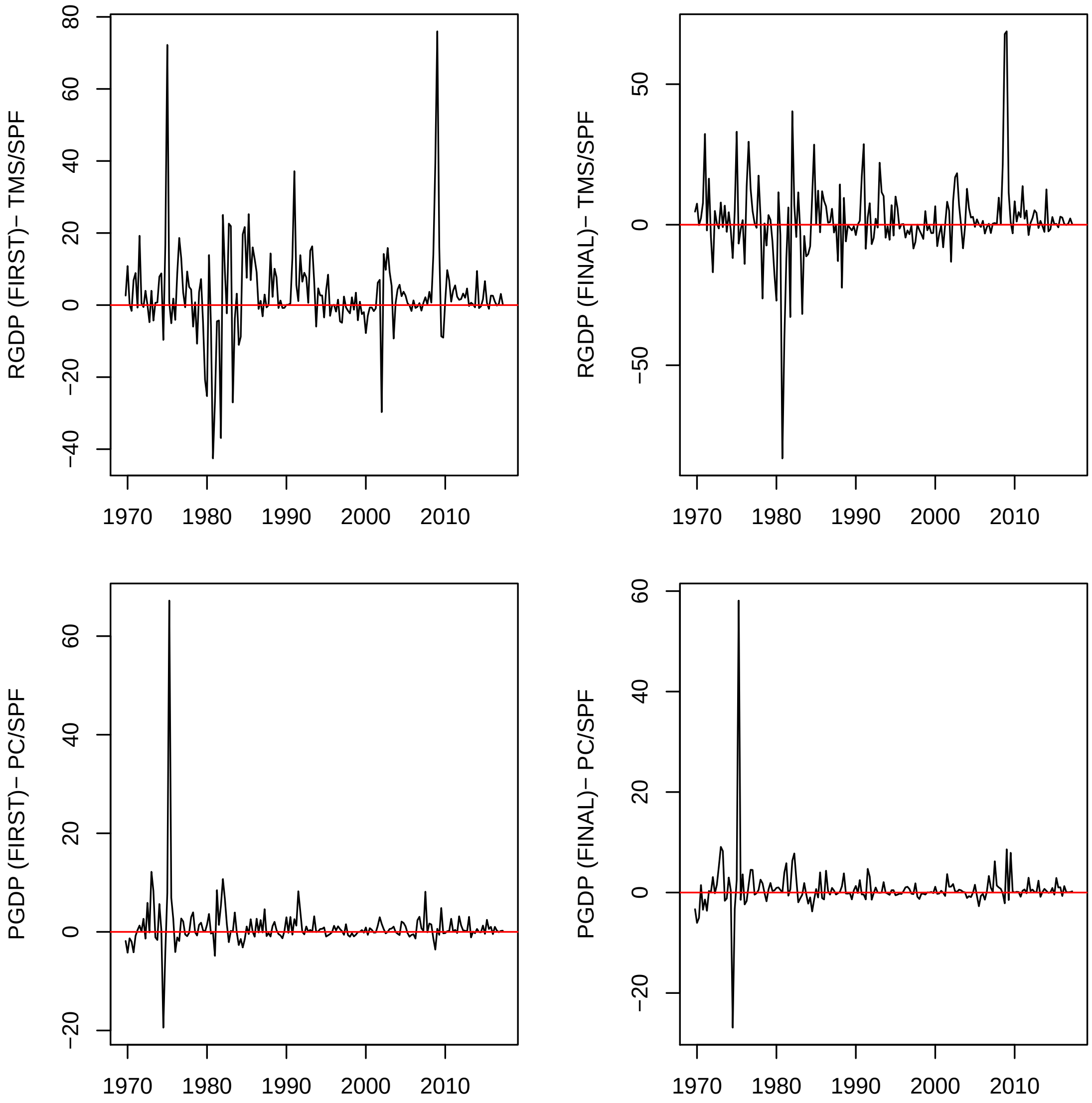


Figure 50: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing TMS forecasts against the SPF. One-quarter ahead forecasts ($h = 1$) are evaluated against the first and the final release.

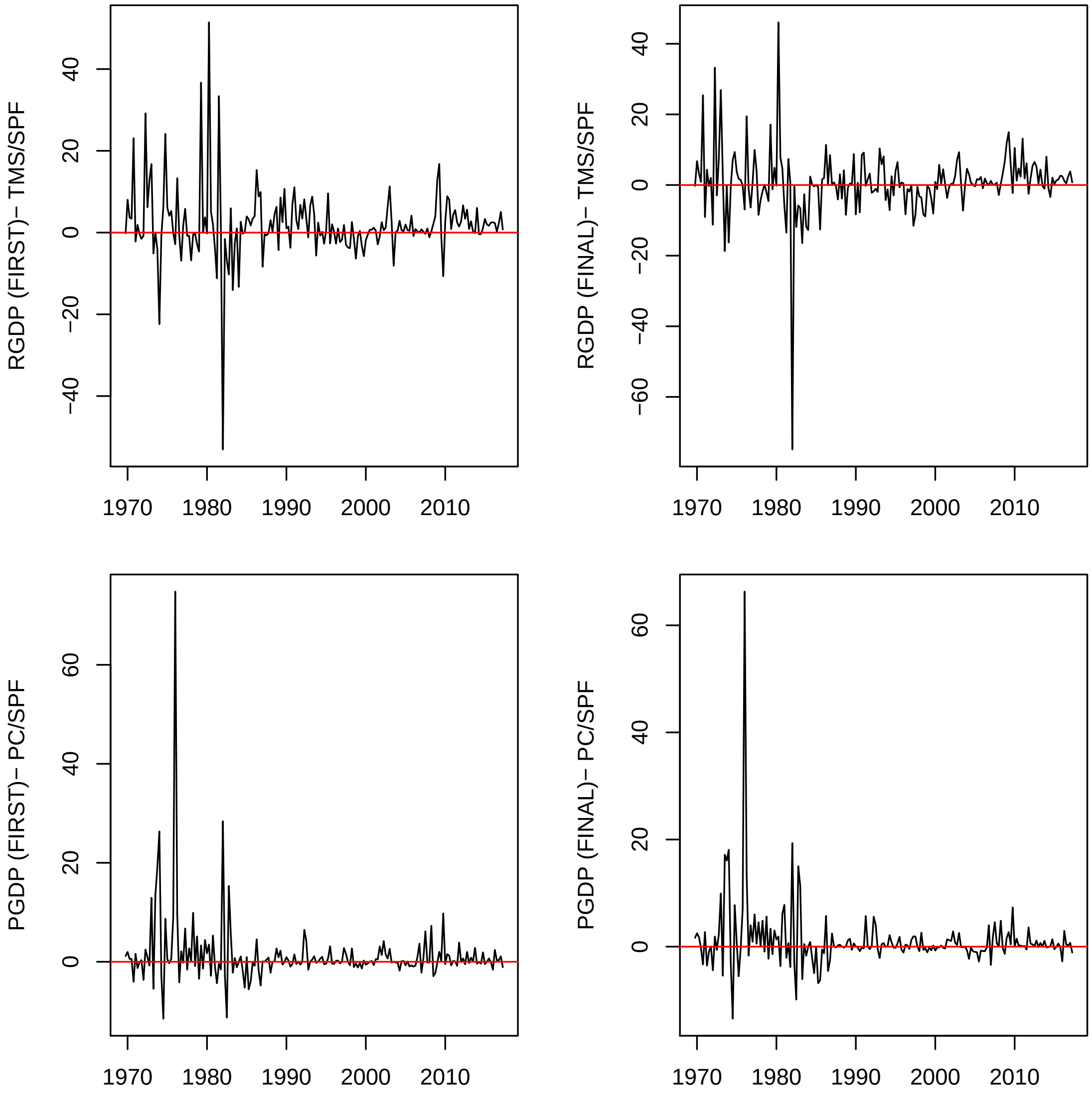


Figure 51: Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing TMS forecasts against the SPF. One-year ahead forecasts ($h = 4$) are evaluated against the first and the final release.

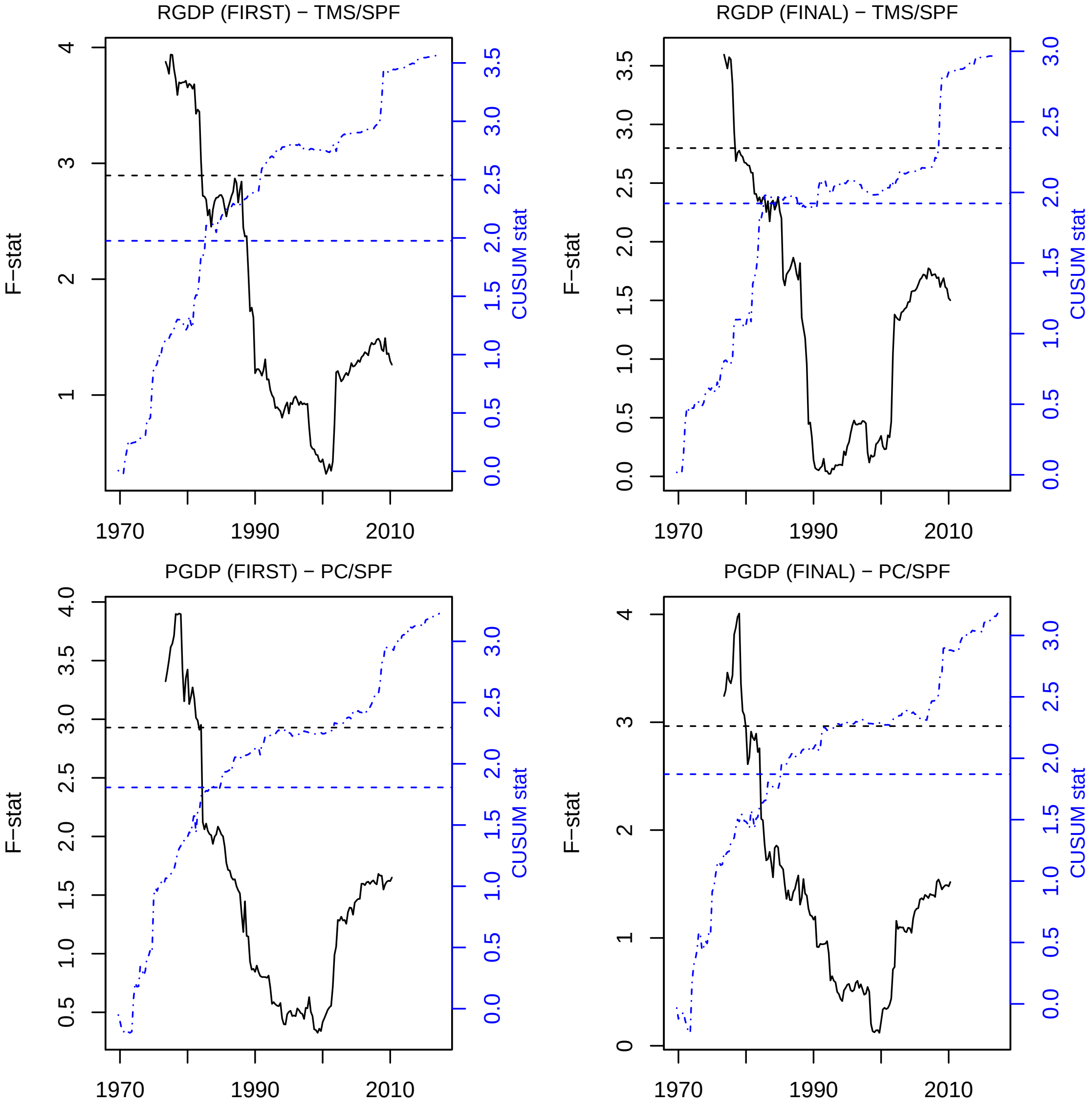


Figure 52: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. Nowcasts ($h = 0$) are evaluated against the first and final release; $b = 0.2$, $\nu = 0.3$.

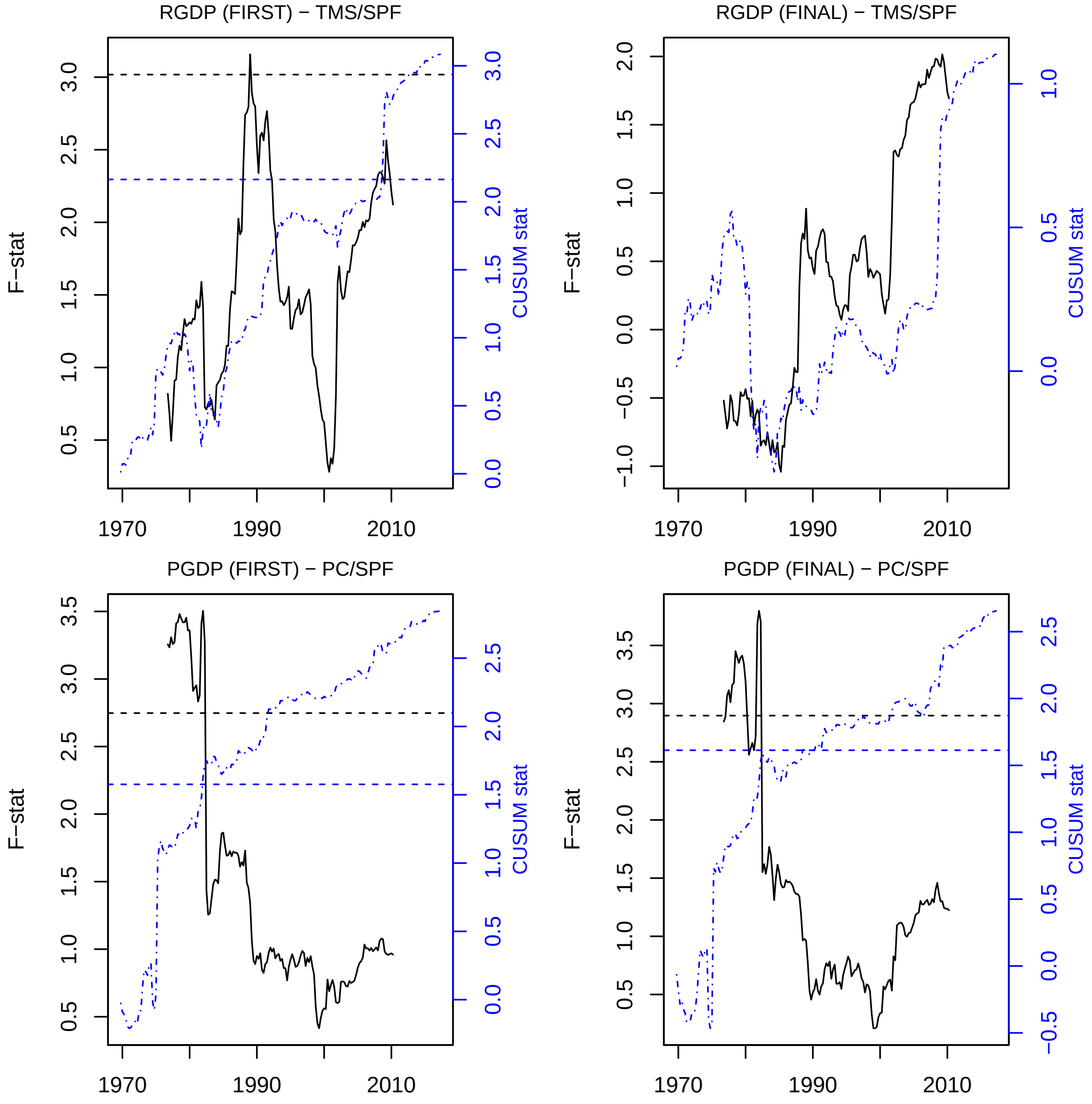


Figure 53: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. One-quarter ahead forecasts ($h = 1$) are evaluated against the first and final release; $b = 0.2$, $\nu = 0.3$.

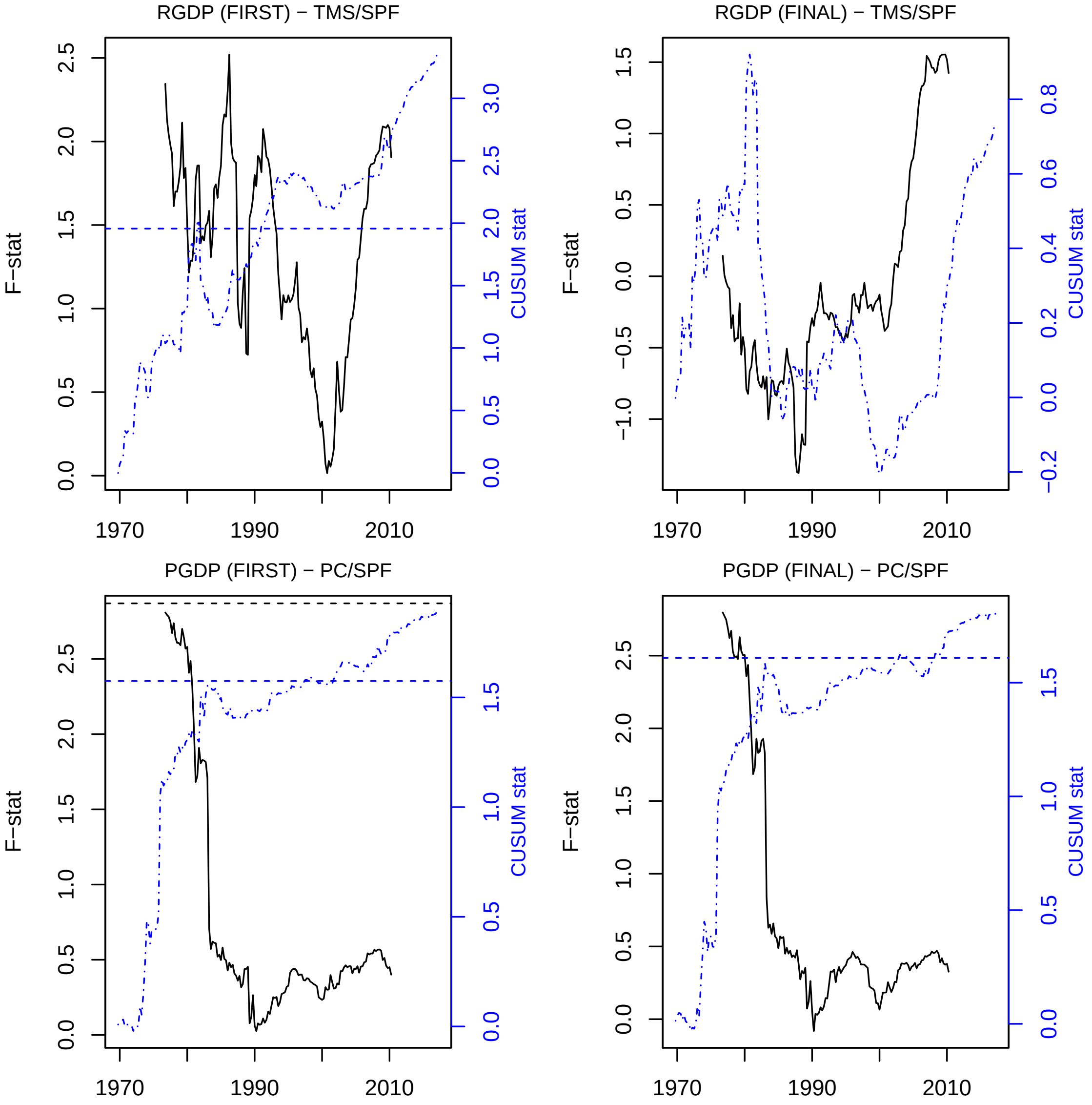


Figure 54: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. One-quarter ahead forecasts ($h = 4$) are evaluated against the first and final release; $b = 0.2$, $\nu = 0.3$.

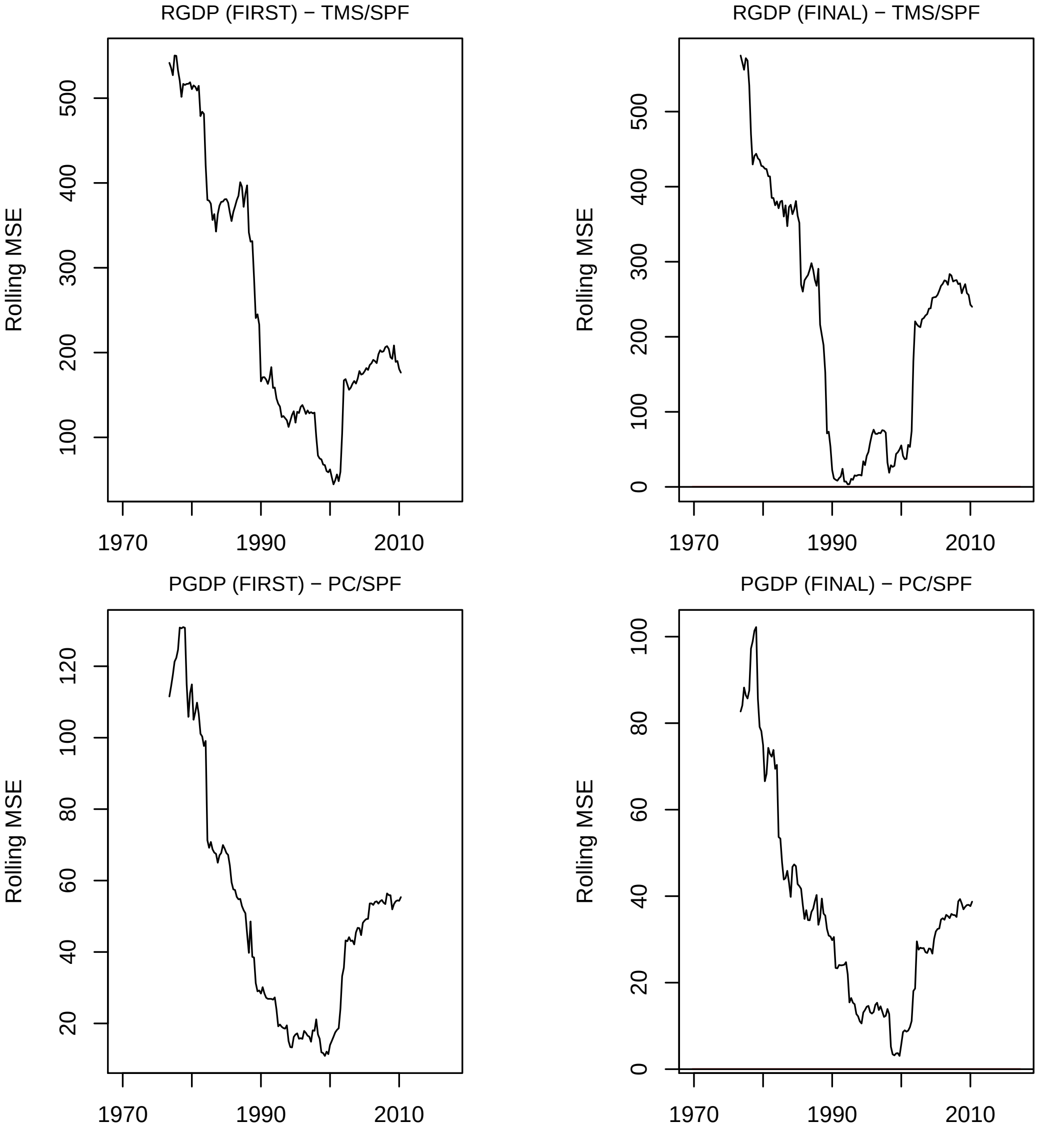


Figure 55: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). Nowcasts ($h = 0$) are evaluated against the first and final release.

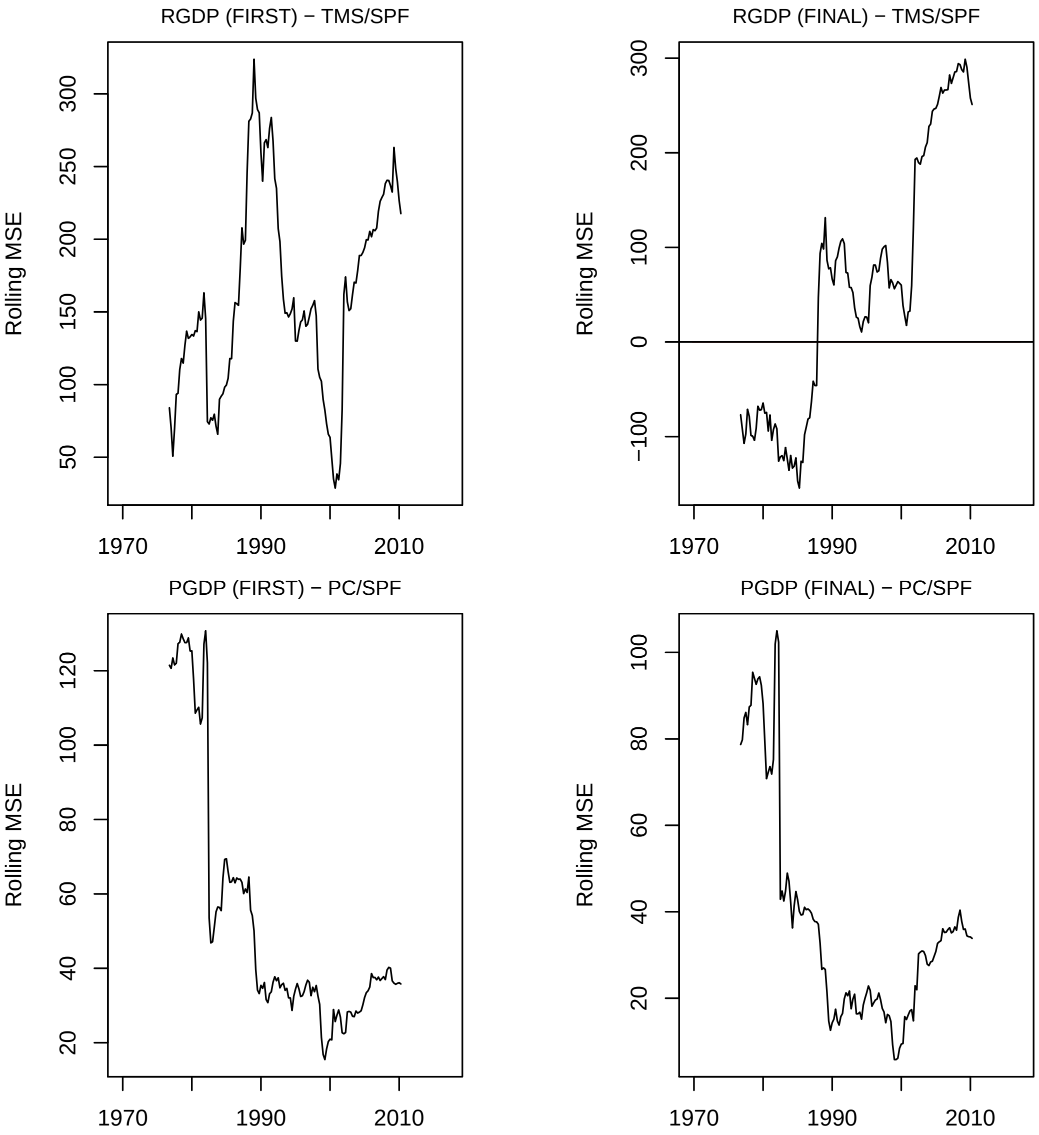


Figure 56: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). One-quarter ahead forecasts ($h = 1$) are evaluated against the first and final release.

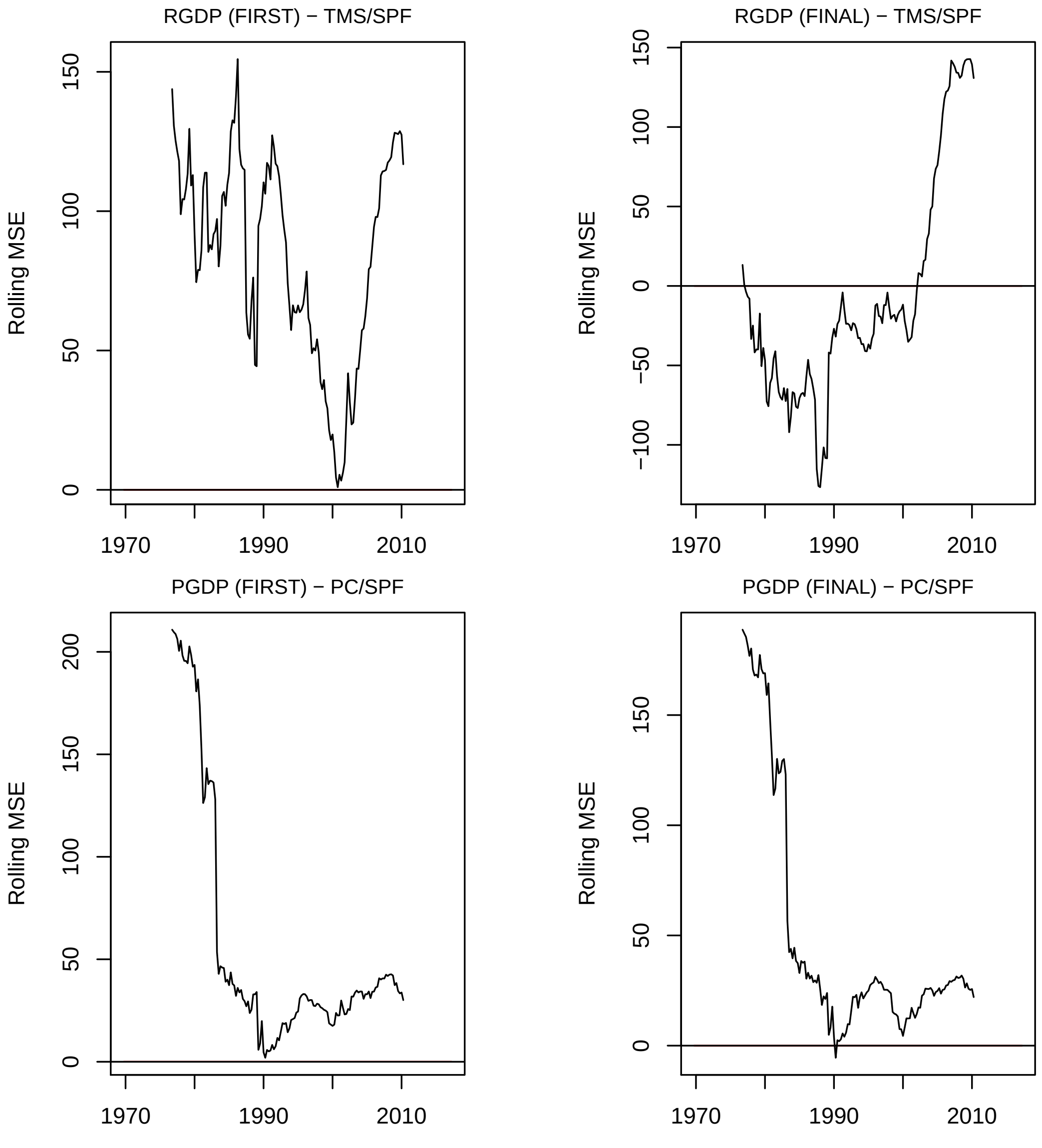


Figure 57: The plots show the rolling mean squared error difference (unscaled, $\nu = 0.3$). One-year ahead forecasts ($h = 4$) are evaluated against the first and final release.

J Additional empirical results - unemployment and housing starts

This section provides some selected results for other SPF variables, viz. unemployment and housing starts for the case of SPF vs. no-change forecasts, to document that our key findings are robust to the choice of variables.[29]

In particular, Figure 58 plots the loss differentials an in Figure 1. Table 16 presents descriptive evidence similar to Table 1. The full-sample results from Table 17 are in line with those of the left panels of Table 2. The findings in Table 18 match those of the upper panels of Tables 3 and 4. Finally, Figure 59 provides evidence regarding time-varying predictability similar to that of the left panels of Figure 2.

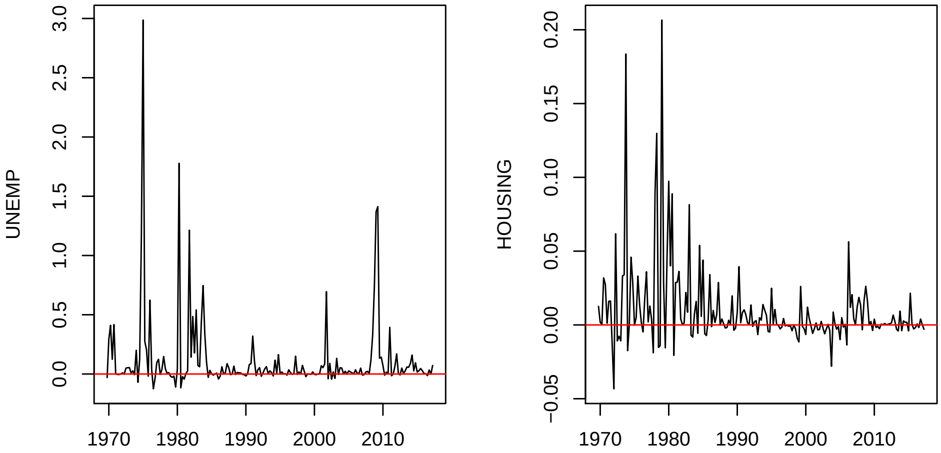


Figure 58: Loss differential series (no-change versus SPF) for the unemployment rate (UNEMP) and housing starts (HOUSING). Nowcasts are evaluated against the first release for mean squared error loss.

Table 16: Summary statistics for the unemployment rate (UNEMP) and housing starts (HOUSING) using the first data release. RelLoss denotes the relative root mean squared error loss of the competing no-change forecasts against the SPF. SD(·) labels the standard deviation of the loss differentials in the subsample I (1969-1984), II (1985-2006) or III (2007-2017). AC(1) denotes the empirical first-order autocorrelation coefficient of the loss differential series.

Statistic		RelLoss(NC/SPF)	SD(I)	SD(II)	SD(III)	AC(1)
Sample		1969-2017	1969-1984	1985-2006	2007-2017	1969-2017
UNEMP	$h = 0$	2.38	0.50	0.09	0.31	0.33
	$h = 1$	1.76	1.15	0.17	0.95	0.58
	$h = 4$	1.43	2.21	0.48	2.16	0.67
HOUSING	$h = 0$	1.40	0.05	0.01	0.01	0.07
	$h = 1$	1.32	0.08	0.02	0.02	0.26
	$h = 4$	1.20	0.25	0.05	0.07	0.61

[29] Further results on the final release, rolling mean squared errors etc., as well as details on the imputation performed on unemployment and housing starts, are available upon request.

Table 17: Test decisions for the full-sample $\mathcal{T}^{DM}$-statistic for equal predictive ability of competing no-change forecasts against the SPF - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the first data release. Evaluation sample runs from 1969Q4 to 2017Q2.

	UNEMP						HOUSING					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	***	***	***	***	***	***	***	***	**
0.1	***	***	***	***	***	***	***	***	***	***	**	**
0.2	***	***	***	***	***	***	***	**	**	**	*	*
0.3	***	**	***	**	***	***	**	*	**	*	*	
0.4	***	**	***	**	***	**	**	*	**	*		
0.5	***	**	***	**	***	**	**	*	*			
0.6	***	**	***	**	***	***	**		*			
0.7	***	**	***	**	***	***	**		*			
0.8	***	**	***	**	***	***	**		*			
0.9	***	**	***	**	***	***	*		*			
1	***	**	***	**	***	***	*		*			

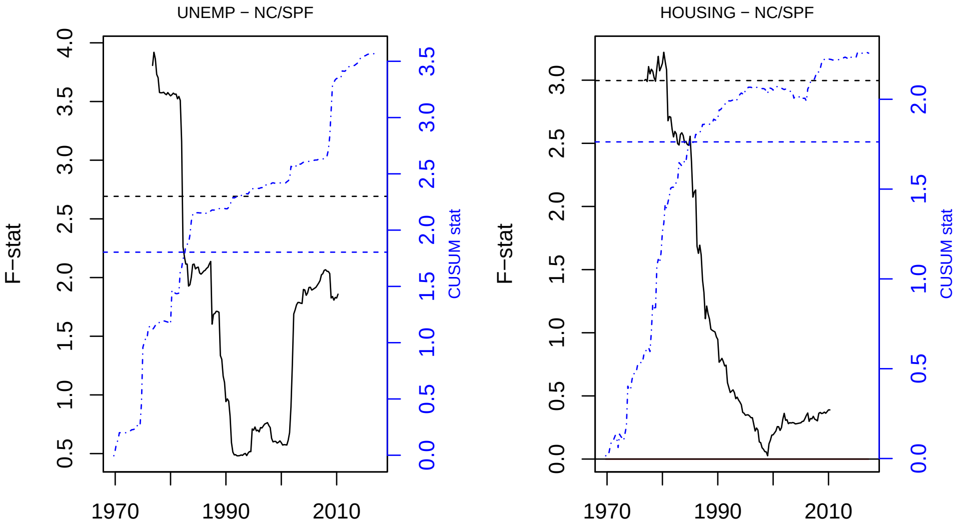


Figure 59: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. Nowcasts are evaluated against the first release; $b = 0.2$, $\nu = 0.3$.

Table 18: Test decisions for the time-variation $\mathcal{T}^{\{Q,C,F\}}$-statistics for time-variation in the predictive ability of competing no-change forecasts against the SPF - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the first data release. Evaluation sample runs from 1969Q4 to 2017Q2.

UNEMP

	$h=0$						$h=1$						$h=4$					
b	$\mathcal{T}^{Q}_{\mathrm{bs}}$	$\mathcal{T}^{Q}_{\mathrm{asy}}$	$\mathcal{T}^{C}_{\mathrm{bs}}$	$\mathcal{T}^{C}_{\mathrm{asy}}$	$\mathcal{T}^{F}_{\mathrm{bs}}$	$\mathcal{T}^{F}_{\mathrm{asy}}$	$\mathcal{T}^{Q}_{\mathrm{bs}}$	$\mathcal{T}^{Q}_{\mathrm{asy}}$	$\mathcal{T}^{C}_{\mathrm{bs}}$	$\mathcal{T}^{C}_{\mathrm{asy}}$	$\mathcal{T}^{F}_{\mathrm{bs}}$	$\mathcal{T}^{F}_{\mathrm{asy}}$	$\mathcal{T}^{Q}_{\mathrm{bs}}$	$\mathcal{T}^{Q}_{\mathrm{asy}}$	$\mathcal{T}^{C}_{\mathrm{bs}}$	$\mathcal{T}^{C}_{\mathrm{asy}}$	$\mathcal{T}^{F}_{\mathrm{bs}}$	$\mathcal{T}^{F}_{\mathrm{asy}}$
0	***	***	***	***	***	***	***	***	***	***	***	**	***	***	***	***	***	***
0.1	***	***	***	***	***	***	***	***	***	***	**	***	***	***	***	***	***	***
0.2	***	**	***	***	***	**	***	**	***	***	**	**	***	***	***	***	**	**
0.3	***	**	***	**	**	*	**	**	***	**	**	*	**	**	***	**	*	
0.4	**	**	**	**	**	*	**	**	**	**	**		**	**	**	**	*	
0.5	**	**	**	**	*		**	**	**	**	*		**	**	**	**		
0.6	**	**	**	**	*		**	**	**	**	*		**	**	**	**		
0.7	**	**	**	**	*		**	**	**	**	*	*	**	**	***	**	*	
0.8	**	**	***	**	*		**	**	***	**	**		**	**	***	**	*	*
0.9	**	*	**	**	*		**	**	**	**	**		**	**	***	**	*	*
1	**	*	**	**	*		**	**	***	**	**		**	**	**	**	*	*

HOUSING

	$h=0$						$h=1$						$h=4$					
b	$\mathcal{T}^{Q}_{\mathrm{bs}}$	$\mathcal{T}^{Q}_{\mathrm{asy}}$	$\mathcal{T}^{C}_{\mathrm{bs}}$	$\mathcal{T}^{C}_{\mathrm{asy}}$	$\mathcal{T}^{F}_{\mathrm{bs}}$	$\mathcal{T}^{F}_{\mathrm{asy}}$	$\mathcal{T}^{Q}_{\mathrm{bs}}$	$\mathcal{T}^{Q}_{\mathrm{asy}}$	$\mathcal{T}^{C}_{\mathrm{bs}}$	$\mathcal{T}^{C}_{\mathrm{asy}}$	$\mathcal{T}^{F}_{\mathrm{bs}}$	$\mathcal{T}^{F}_{\mathrm{asy}}$	$\mathcal{T}^{Q}_{\mathrm{bs}}$	$\mathcal{T}^{Q}_{\mathrm{asy}}$	$\mathcal{T}^{C}_{\mathrm{bs}}$	$\mathcal{T}^{C}_{\mathrm{asy}}$	$\mathcal{T}^{F}_{\mathrm{bs}}$	$\mathcal{T}^{F}_{\mathrm{asy}}$
0	***	***	***	***	***	***	***	***	***	***	***	***	**	**	***	***	**	***
0.1	***	**	***	***	***	***	***	**	***	***	***	***	**		**	**	**	**
0.2	**		**	**	*	*	**		**	**	*	*			*	*		
0.3	*		**	*					**	*								
0.4			*	*					*									
0.5			*															
0.6			*															
0.7			*															
0.8			*															
0.9																		
1			*															

K Additional empirical results - subsample analysis

This section provides an additional analysis on the subsamples studied by Coroneo and Iacone (2020). Their SPF data sample from the post-"Great Moderation" period runs from 1987Q1 to 2016Q4 with a total of 120 observations. The authors form three equally sized subsamples of ten years of quarterly data with 40 observations each. The resulting subsamples show different levels of volatility each. For instance, the third subsample starting in 2007 has higher volatility than the others due to its relation to the "Great Financial Crisis". The foci are now on no-change forecasts and an evaluation against the first release to facilitate a comparison with their results.

First, we run the one-sided $\mathcal{T}^{DM}$-statistic on each of the three subsamples. Table 19 reports the results. Overall, the findings are very similar to Coroneo and Iacone (2020) (their Tables 1 and 2). This is as expected since the volatility varies much more across the individual subsamples rather than within. In a second step, we apply our tests for time-variation on their sample. Thereby, we are able to identify periods of instability (without imposing sample split points ourselves) and compare our findings to their results. In particular, Figure 60 shows the results of the signed time-varying components of the fluctuation and the CUSUM test statistic together with one-sided wild bootstrap critical values at the five percent level. Vertical red dashed lines show the sample split choices by Coroneo and Iacone (2020). Considering the fluctuation test, we find in nearly all cases a rejection in favor of time-varying advantages of the SPF. The case of one-year ahead inflation forecasts is an exception. The CUSUM statistic turns out to be significant in all cases. More importantly, the fluctuation test shows remarkable and significant time-variation within all given subsamples of ten years, especially for output growth in middle subsample from 1997Q1 to 2006Q4 at all horizons. For inflation, we find all subsamples to be affected in this respect. The time-varying nature within subsamples provides some evidence that ad hoc choices might be problematic, as already discussed.

As a next step, we compare the subsample evidence reported in Coroneo and Iacone (2020) (their Tables 1 and 2) to the one obtained via the wild bootstrap for the fluctuation test. Overall, our results show interesting and remarkable differences to those reported in Coroneo and Iacone (2020). For instance, in the case of one-year ahead output growth forecasts, the evidence reported in Coroneo and Iacone (2020) is relatively weak and suggests only for the first subsample that the SPF outperforms the no-change benchmark. A rather opposite result is suggested by the fluctuation test which indicates that the SPF outperforms the benchmark just prior to the "Great Financial Crisis", but not beforehand. Besides these apparent differences, there are some cases in which the test decisions agree (see, e.g., inflation nowcasts), but in general they do not match well. The CUSUM test results generally indicate the importance of time-variation since the 2000s. They further emphasize the time-variation of relative predictive ability within the subsamples. Overall, we find evidence suggesting that the differences in the results are due to different testing environments (ad hoc sample splits versus fluctuation and related tests allowing for endogenous break points).

Table 19: Test decisions for the $\mathcal{T}^{DM}$-statistic for equal predictive ability of competing no-change forecasts against the SPF - either based on wild bootstrap ('bs') or asymptotic critical values ('asy'). Nowcasts ($h = 0$), one-quarter ($h = 1$) and one-year ahead forecasts ($h = 4$) are evaluated against the first data release. Evaluation periods are the subsamples from 1987Q1 to 1996Q4, from 1997Q1 to 2006Q4 and from 2007Q1 to 2016Q4.

	RGDP - 1987Q1 to 1996Q4						PGDP - 1987Q1 to 1996Q4					
	$h = 0$		$h = 1$		$h = 4$		$h = 0$		$h = 1$		$h = 4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	***	**	**	**	**	***	***	*	*
0.1	***	***	***	**	*	*	**	*	***	**	*	
0.2	***	***	***	**	**	**	**	*	**	**		
0.3	***	***	**	**	**	**	*		**	**		
0.4	***	***	**	**	**	**	*		**	**		
0.5	***	***	**	**	**	**			**	**		
0.6	***	***	**	**	**	**			**	**		
0.7	***	***	**	**	**	**			**	**		
0.8	***	***	**	**	**	**			**	**		
0.9	***	***	**	**	**	**			**	**		
1	***	***	**	**	**	**			**	**		

	RGDP - 1997Q1 to 2006Q4						PGDP - 1997Q1 to 2006Q4					
	$h = 0$		$h = 1$		$h = 4$		$h = 0$		$h = 1$		$h = 4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	**	***	***	**	**	***	**	**	**		
0.1	**	**	***	***	**	**	**	**	*	*		
0.2	**	**	***	**	**	**	**	**	*	*		
0.3	**	**	***	***	**	**	**	*	*			
0.4	**	**	***	***	**	*	**	*				
0.5	**	*	***	***	*	*	**	*				
0.6	*	*	***	***	*	*	**	*				
0.7	*	*	***	***	**	*	**	*				
0.8	*	*	***	***	**	**	**	*				
0.9	*	*	***	***	**	**	**	*				
1	*	*	***	***	**	**	**	*				

	RGDP - 2007Q1 to 2016Q4						PGDP - 2007Q1 to 2016Q4					
	$h = 0$		$h = 1$		$h = 4$		$h = 0$		$h = 1$		$h = 4$	
b	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$	$\mathcal{T}^{DM}_{\text{bs}}$	$\mathcal{T}^{DM}_{\text{asy}}$
0	***	***	***	**			***	***	***	***	**	**
0.1	***	**	***	*			***	***	***	***	**	**
0.2	**	**	**	*			**	**	***	**	***	**
0.3	**	**	**				**	**	***	***	**	**
0.4	**	**	**				**	**	***	***	**	*
0.5	**	**	**				**	*	***	**	**	*
0.6	**	*	**				**	*	***	**	**	*
0.7	**	*	*				*	*	***	**	**	*
0.8	*	*	*				**	*	***	**	**	*
0.9	*	*	*				*	*	***	**	**	*
1	**	*	*				*	*	***	**	**	*

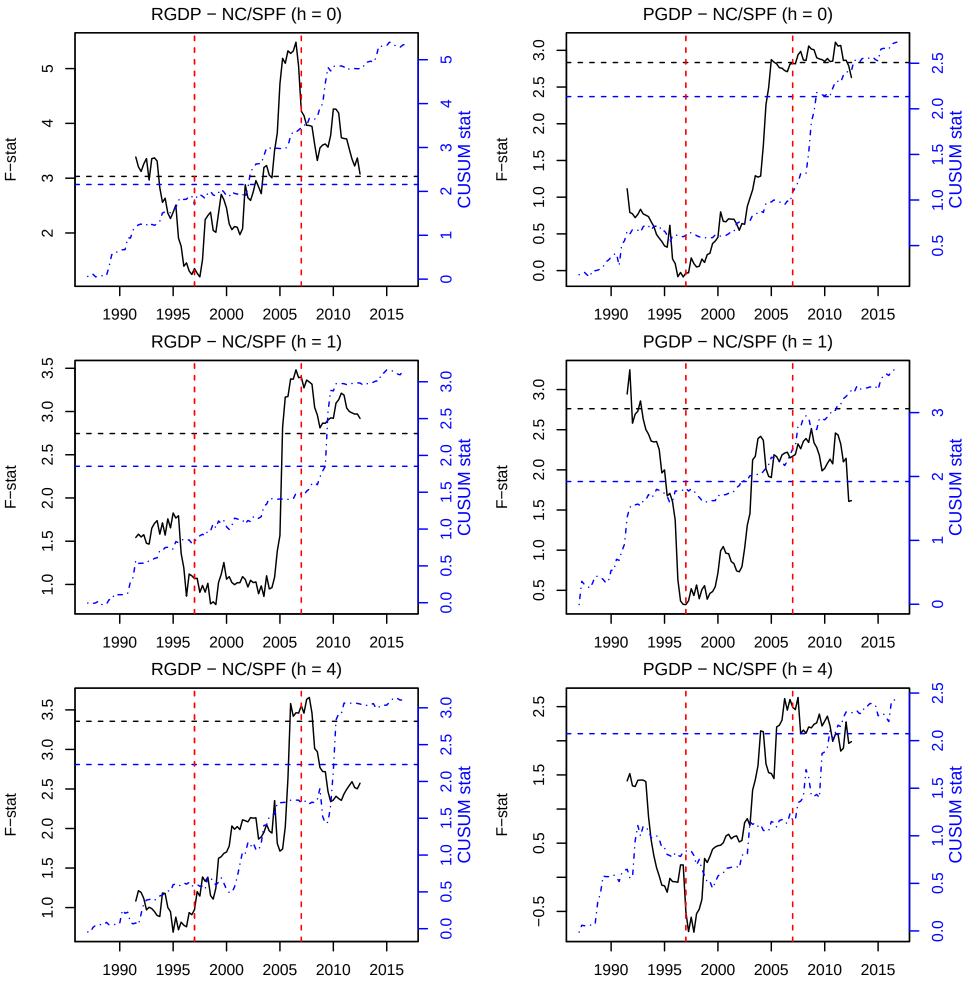


Figure 60: The plots show the time-varying components of the signed fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see equations 4 and 5 and Remark 3. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. Dashed red vertical lines indicate sample split points as in Coroneo and Iacone (2020). Now- and forecasts are evaluated against the first release; $b = 0.2$, $\nu = 0.3$. Evaluation sample runs from 19871Q1 to 2016Q4 as in Coroneo and Iacone (2020).