Transregio 391

TRR 391 Working Paper #13

June 2026

A robust test for equal predictive accuracy

Matei Demetrescu[a], Christoph Hanck[b], and Yannick Hoga[b]

Suggested citation:
Demestrescu, M., Hanck, C., and Hoga, Y. (2026). "A robust test for equal predictive accuracy". *TRR 391 Working Paper #13*. DOI: 10.17877/DE290R-26668.

Version 1.0, June 2026

[a] Department of Statistics, TU Dortmund University
[b] Faculty of Business Administration and Economics, University of Duisburg-Essen
Corresponding author: mdeme@statistik.tu-dortmund.de

A robust test for equal predictive accuracy*

Matei Demetrescu , Christoph Hanck , and Yannick Hoga
TU Dortmund University
University of Duisburg-Essen

May 21, 2026

Abstract

This paper studies what we call a "robust" Diebold–Mariano type test. The unique feature of our test is that—even in the absence of *any* knowledge of the forecasting method—it is robust to estimation noise in the forecasts, i.e., size is kept irrespective of estimation effects induced by model fitting. We obtain this feature by a test statistic that is based on rolling-window means whose length is a vanishing fraction of the total evaluation sample. This leads to non-standard Gumbel limit laws. Other desirable features of our test are that it is easily robustified against time-varying volatility, and that it naturally uncovers time-varying differences in predictive ability under the alternative. Simulations demonstrate the benefits of our multiply robust implementation vis-à-vis several competitors. An empirical application to forecasts for several variables, horizons, vintages and methods from the Survey of Professional Forecasters illustrates the relevance of the new approach, allowing us to identify forecasters with superior models. Such conclusions are in fact impossible to infer by extant tests, since information on the models and estimation procedures behind the forecasts are typically proprietary and, hence, estimation effects cannot be factored out.

Key words: Equal predictive ability; Estimation error; Rolling windows; General cost-of-error function; Hypothesis testing

JEL classification: C12 (Hypothesis Testing), C22 (Time-Series Models), C52 (Model Evaluation, Validation and Selection), C53 (Forecasting and Prediction Methods)

*The authors would like to thank Kasumi Nakayama for excellent research assistance, as well as the participants at the A&E seminar at the University Carlos III of Madrid and the WTSE conference in Zaragoza for very helpful comments and remarks. Demetrescu and Hanck gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within TRR 391: Spatio-temporal Statistics for the Transition of Energy and Transport (520388526). Hoga also thankfully recognizes support of the DFG through grants 460479886, 531866675 and 568876076. All remaining errors are ours. Contact details: mdeme@statistik.tu-dortmund.de, christoph.hanck@vwl.uni-due.de, yannick.hoga@vwl.uni-due.de.

1 Motivation

Forecasting serves as a cornerstone of decision-making across economics, business strategy, and policy making. The stakeholders relying on predictive outputs to handle uncertainty range from central banks and international institutions to private sector firms and households. Among the major institutions issuing forecasts on a regular basis are the IMF, the OECD, the Federal Reserve, and commercial providers (e.g., the Survey of Professional Forecasters; SPF). These predictions directly influence fiscal policy, investment strategies, and risk management processes.

This pervasive reliance on predictions has generated a rich academic literature on the comparative evaluation of forecasts. In their seminal work, Diebold and Mariano (1995) were the first to provide formal statistical tests for the equal predictive ability (EPA) of forecasts, i.e., tests that assess whether competing forecasts yield statistically equivalent losses. We call tests in this vein "classic" EPA tests. Since then, the field has seen a rapid development, which is nicely reviewed by Diebold (2015).

By now, EPA tests come in two flavors. The first sort follow in the spirit of Diebold and Mariano (1995) and take the observed forecasts as primitives (see, e.g., Giacomini and White, 2006). This is taken to mean that the complete forecasting *method* is evaluated—including not only the forecasting model, but also the estimation methodology, that is, the type of estimator used (e.g., least squares, maximum likelihood, etc.) and the estimation scheme (e.g., rolling, fixed or expanding window). Next to Diebold and Mariano (1995) and Giacomini and White (2006), tests from this strand of the literature include (e.g.) Giacomini and Rossi (2010), Li et al. (2022) and Lan et al. (2024). In contrast, the second type of EPA tests acknowledge the sampling variation induced by parameter estimation, and—by accounting for it—they may be seen as rather evaluating the forecasting *model*. Tests in this tradition were first proposed by West (1996) and now include, among others, Clark and McCracken (2001, 2009), Patton (2020), and Demetrescu et al. (2022). For more discussion on the relative merits of the two approaches, we refer to Giacomini and White (2006, Sec. 2.1). We contribute to both strands of the literature in this paper.

Our first contribution extends classic EPA tests by making them more sensitive to local episodes of different predictive ability. We achieve this by considering means of the loss differences calculated on rolling windows of width . A key step in our construction is that we let vanish relatively to the total size of the evaluation sample. Our proposal thus uses a window size that, although increasing asymptotically, is only a vanishing fraction of , i.e., , as . The means are then aggregated suitably to provide a more detailed analysis of the full sample. This allows us to identify differences in predictive accuracy even when, on average, the loss differences have zero mean—which would render classic Diebold and Mariano (1995) tests (DM tests) powerless.

Our second and foremost contribution is to the literature dealing with parameter estimation. One of the key insights of the papers in the West (1996) tradition is that factoring out estimation effects is highly situation-specific, requiring in particular knowledge of the forecasting model, the length of the estimation sample, and the type of estimator used to fit the model. Such knowledge is, however, unavailable in most contexts, as the predictive methodology is proprietary for almost all providers of forecasts. Therefore, this route is closed in typical empirical applications. West (1996) and others also argue that if the evaluation sample is a vanishing fraction of the estimation window (i.e.,), then effects from model fitting no longer impact the limiting distribution. In this case, such factoring out would not be necessary. Of course, in practice, the evaluation sample is always a non-negligible fraction of the estimation sample (which, unlike , is not controlled by the analyst and may, in fact, not even be known), such that the product of estimation noise and will impact the EPA test statistics' distributions. Hence, traditional tests of EPA may not properly control size under typical circumstances.

Nonetheless, this insight suggests that our proposed "vanishing rolling-window paradigm" leads to limiting distributions that are *not* affected by estimation noise. The key intuition is that, whenever the estimation error matters in the limit—and we follow those who argue that they often do in practice—our is, by construction, only a vanishing fraction of and, hence, also a vanishing fraction of (even when). Therefore, estimation effects become negligible on each of the smaller rolling windows. We indeed prove that, for a suitably constructed test, the effect stays asymptotically negligible upon aggregation to the full-sample level. Thus, ours is the first EPA test that validly allows to compare predictive models even when details on the estimation methodology are unavailable, which is our second and most important contribution.

More concretely, our test builds on, but then suitably rescales, the fluctuation test of Giacomini and Rossi (2010). Their test is based on a moving sum whose window size is a *non-vanishing* fraction of , i.e., . Under such asymptotics, Demetrescu et al. (2022) show that the West (1996) effect is, in general, still present asymptotically for the fluctuation test. To suitably rescale the Giacomini and Rossi (2010) statistic, we use certain sequences following from extreme value theory. More specifically, Section 2 derives closed-form critical values from an extreme value Gumbel limit and shows these to be the same *both* in the presence and in the absence of estimation error.

Therefore, the advantages of the "vanishing " setting are that (i) even short periods of different predictive ability can be picked up and, more importantly, (ii) estimation effects disappear from the limiting distribution for a judicious choice of . Not least, we also show that our test can (iii) be robustified against time-varying volatility and dynamics of the loss differentials in a natural manner; see, e.g., Coroneo and Iacone (2020), Demetrescu et al. (2022) and Harvey et al. (2024) for the empirical relevance of time-varying behavior

of loss differentials.

Besides its connection to Giacomini and Rossi (2010), our test statistic with "vanishing " rolling window sizes is also related to the maximum subsampling statistic of Lan et al. (2024). They suggest to randomly draw (in total) subsamples of length from the complete set of loss differences, similar to our "vanishing " moving windows. By doing so, the average distance between the randomly drawn loss differences is . In turn, this implies that the loss differences are close to independent, such that they can be standardized with a sample variance estimator. In particular, this obviates the need to use heteroskedasticity and autocorrelation consistent (HAC) estimators, which is the main motivation for Lan et al. (2024) to introduce their maximum subsampling statistic. In contrast, our reasons for considering small subsamples in the form of rolling windows are completely different, viz. increased sensitivity to short-lived episodes of different predictive accuracy (since the "vanishing " rolling window sizes naturally lead to an estimator of the local average loss differential) and, more importantly, vanishing estimation effects.

By considering time-varying variances, our work also relates to Harvey et al. (2024), who provide more efficient implementations of the DM test based on GLS-type weighting. However, they do not address estimation noise and they only consider constant alternatives.

Finally, our work nicely connects to Escanciano and Parra (2026), who derive tests for comparing forecasting models that are based on machine learning (ML) algorithms. They argue that a fast rate of convergence of the machine learners is key to validly compare prediction models. Our approach in fact yields suitable *implied* rates of convergence by only considering moving sums over *narrower* windows in our test statistic. Therefore, our tests may also be used to validly compare forecasts issued from ML models.

Section 3 presents a series of Monte Carlo experiments designed to highlight the above crucial challenges in conducting inference regarding EPA. These include size and power, the latter both for the case of constant and time-varying superior predictive ability (SPA). We also disentangle the effects of estimation uncertainty and the (asymptotic) lack thereof. Additional simulations discuss extensions to machine learning-type applications. We finally present results for testing EPA under time-varying volatility. Overall, the results support the broad usefulness of the approach developed here: it has robust size throughout the scenarios, and good power in those cases it is designed for, namely detecting short periods of superior predictability.

Section 4 applies the tests considered in this paper to gross domestic product (GDP), housing starts, inflation and unemployment forecasts for several horizons, vintages and methods from the Survey of Professional Forecasters (Croushore, 1993). This application illustrates the relevance of the new approach, allowing us to identify forecasters with superior models. Such conclusions are impossible to infer by extant tests, since information on the models and estimation procedures behind the forecasts are typically not available

and, hence, estimation effects cannot be controlled for. Our results for example highlight that SPF forecasts typically outperform more statistical methods such as autoregressive or no-change forecasts mainly in (typically relatively short) recessionary periods, while they are comparable in more tranquil times. Such time-varying SPA is often overlooked by average-based tests such as the Diebold and Mariano (1995) test.

Section 5 concludes. All proofs as well as additional simulations and empirical results are relegated to an Online Appendix.

2 Robust tests of EPA

2.1 Setup

We denote the forecast target by and by the predictors used to generate the forecasts for time (which are available at the time of setting up the forecast). As usual, may contain various leading indicators as well as lags thereof; furthermore, the two competing forecasts may in fact rely on different subsets of , as is not uncommon in practice.

To fix ideas, consider the forecasts and issued by two competing prediction models, where and are true parameters. (Of course, the true parameters are only known in rare cases, such as for random walk forecasts, no-change forecasts, or constant forecasts. We relax the assumption of known parameters in Section 2.3 below.) Associated with the ideal forecasts are the ideal forecasts errors . The outcome of the comparison of the ideal forecasts errors hinges on the specific loss $\mathcal{L}$ chosen by the evaluator. For expositional ease, we consider the following class.

Assumption 0 (loss function)**.** *Let* $\mathcal{L}$ $_{\mathcal{L}}$ $_{\mathcal{L}}$ **1** *, where* **1** *is the usual indicator function, some positive integer, and* $_{\mathcal{L}}$ *.*

Assumption 0 corresponds to the family of asymmetric power loss functions due to Elliott et al. (2005). This class of loss functions nests, for $_{\mathcal{L}}$, the absolute error ( ) and the squared error ( ) loss functions, while $_{\mathcal{L}}$ implies asymmetric losses (e.g., the pinball loss for). For later reference, we note that $\mathcal{L}$ is times continuously differentiable with the derivative of order being uniformly Lipschitz on . Such smoothness of the loss function will allow us to control for estimation error.

We then consider the ideal loss differences $\mathcal{L}$ $\mathcal{L}$. Going back to Diebold and Mariano (1995) and West (1996), the null hypothesis of interest in EPA-type testing is

$$\text{for all} \tag{2.1}$$

This is the null hypothesis we consider throughout the paper.

We assume that an evaluation sample of actual outcomes as well as of forecasts , , is available. Then, to formally investigate , most tests in the Diebold and Mariano (1995) tradition rely on the full sample mean $\bar{}$. Under weak regularity conditions, a limiting standard normal null distribution obtains for

$$\mathcal{DM} \quad \bar{} \frac{\bar{}}{\text{—}} \tag{2.2}$$

for some suitable long-run variance estimate of . See Diebold and Mariano (1995) and also the assumptions below.

We identify three econometric challenges that arise in practice when testing . We address each of these challenges in turn in Sections 2.2–2.4. First, the DM family of tests is insensitive to alternatives where $\bar{}$, yet (say) for and for (such that differences in predictive ability cancel each other out in the full sample mean). Similarly, the power of DM-type tests will typically suffer in cases where (2.1) is violated only for a small fraction of time periods .

This motivates Giacomini and Rossi (2010) to introduce their fluctuation test. For no estimation error, this test is based on rolling a moving window

$$\text{–} \tag{2.3}$$

of length through the evaluation sample of length . Giacomini and Rossi (2010) assume that , as , such that the moving window size is a non-vanishing fraction of the total evaluation sample. Yet, the fluctuation test may also struggle to identify short-lived episodes of differences in predictive accuracy. For instance, consider the alternative

$$^{\text{trans}} \quad \text{for} \quad ^{*} \quad \text{for} \quad ^{*} \quad ^{*} \tag{2.4}$$

and else, where * * is a vanishing fraction of the evaluation sample (such that * , as). Such a brief episode would be hard to identify using a fluctuation test with a relevant that averages over nontrivial parts of the evaluation sample. We address this first challenge in Section 2.2.

The second complication in EPA testing arises since, following West (1996), the true parameters are often not known in practice. Thus, only the forecasts

are available for suitable estimators . Pre-sample data, , may of course

be used by the respective forecasters to estimate the relevant parameters (or, in machine learning lingo, to train their models). In this paper's setup, however, these data are available to the forecaster only, and not to the stakeholder or forecast evaluator. Importantly, we do not require forecast issuers to disclose the model or the estimation method used in constructing their predictions. E.g., may be computed the usual way (in rolling or expanding windows), but may also be obtained by expert consultation.

In particular, this implies that the ideal forecast errors are not available, but the evaluation of the forecasts has to build (quite naturally) on the observed forecast errors

Based on these, we define the observed loss differences

$$\mathcal{L} \qquad \mathcal{L} \tag{2.5}$$

The availability of only the presents econometric challenges in testing , which is phrased in terms of the unobservable . Concretely, it is known since West (1996) that estimation error, i.e., noisy forecasts implying noisy loss differentials via

$$\mathcal{L} \qquad \mathcal{L} \qquad\qquad \mathcal{L} \qquad\qquad _{t \quad i}$$

may affect the limiting distribution of $\overline{\quad}^{-} \qquad ^{-}$. We again refer to West (1996) for a generic discussion of the issue and to Demetrescu et al. (2022) for a specific treatment of the Giacomini and Rossi (2010) fluctuation test in West's framework. Once more, we emphasize that the problem here is not the presence of estimation error: West (1996) does discuss in detail the steps required to handle estimation noise. Rather, what is problematic is that the information required to account for estimation noise is quite often not publicly available. Therefore, corrections are typically unfeasible in practice. For example, in our empirical application (Section 4) to the SPF, the specific sample sizes and forecasting strategies of the various forecasters (in fact, even their identities) contributing to the SPF are generally unknown. This is the second obstacle in testing that we address in Section 2.3.

Thirdly, time-varying volatility, and more generally time-varying dynamics, are particularly problematic for the fluctuation test; see Demetrescu et al. (2022) again. (Interestingly, this is not the case with the full-sample $\mathcal{DM}$ test, where the required average long-run variance can be estimated consistently using Newey–West type estimators; see Cavaliere, 2005, Lemmata 4 and 5.) Specifically, the fluctuation test exhibits limiting null distributions depending on nuisance parameters stemming from time-varying volatility or time-varying dynamics, implying lack of asymptotic size control when ignored. Section 2.4 shows that

robustness may nevertheless be achieved for our "vanishing " test.

2.2 The Baseline Case of No Estimation Error

To fix ideas we start with the case of no estimation noise. Our test is based on the maximum-type statistic

$$\mathcal{M}$$

where is a long-run variance estimator characterized below, but at suitable rates as .[1]

2.2.1 Behavior under the null

When for some fixed , $\mathcal{M}$ corresponds exactly to the fluctuation test statistic of Giacomini and Rossi (2010). Given , Giacomini and Rossi (2010, Proposition 1) show under suitable regularity conditions that, as ,

$$\mathcal{M} \tag{2.6}$$

where denotes a standard Brownian motion.

As we show in Theorem 2.1 below, our "vanishing " approach yields an interestingly different asymptotic limit with a different standardization. Specifically, we employ

$$\mathcal{M} \quad \mathcal{M} \tag{2.7}$$

with standardizing sequences

$$\tag{2.8}$$

where . This may seem slightly involved compared to the original $\mathcal{DM}$ or $\mathcal{M}$ tests. This adjustment, however, yields relevant benefits in that it allows to solve all three issues mentioned in Section 2.1 in the "vanishing " framework. Furthermore, we document in our simulations that, for relatively small window widths (see Section 3 for details), the asymptotics of Giacomini and Rossi (2010) are not reliable throughout compared to the Gumbel-based approximation we discuss.

We now introduce our baseline assumptions.

[1] Intuitively, our "vanishing " fluctuation test comes closer to testing $_0$ in (2.1) because it is, in a sense, truly "local" when compared to the test of Giacomini and Rossi (2010).

Assumption 1 (window size). *The moving window size satisfies , as , where is defined in the proof of Theorem 2.1.*

Assumption 2 (moment conditions forecast errors). *It holds that for , some and from Assumption 0.*

Assumption 3 (dependence forecast errors). *The sequence of forecast errors ′ ∈ is -mixing with -mixing coefficients satisfying − for some and from Assumption 2.*

Assumption 4 (weak stationarity loss differences). *The loss differences ∈ are weakly stationary.*

Assumption 5 (long-run variance estimation). *There exists an estimator that satisfies − − , as , where →∞ − is the long-run variance of the loss differences.*

Assumption 1 sets out the "vanishing " framework by requiring the moving window size to diverge asymptotically at a rate slower than the sample size, yet not too slowly. The conditions are standard in the literature on MOSUM-based change point testing; see, e.g., Hušková and Slabỳ (2001, Eq. (1.5)), Eichinger and Kirch (2018, Eq. (2.3)) and Kirch and Klein (2023, Assumption 2) for identical requirements. Note that Assumption 1 also implies that . In the simulations in Section 3, we provide a rule of thumb for the choice of for a feasible implementation of our test. Assumption 2 is a standard moment bound, which is required for most large-sample results for partial sums. Assumptions 3 and 4 impose standard serial dependence and stationarity conditions. Finally, Assumption 5 is a high-level condition. Note the limit exists thanks to Kuelbs and Philipp (1980, Theorem 4). Since , the imposed convergence rate is only slightly stronger than the usual requirement of consistency, i.e., . In particular, Eichinger and Kirch (2018, Theorem 2.3 (b)) and Horváth and Rice (2024, Theorem 3.1.2) give conditions under which estimators satisfy such a rate. We refer to Horváth et al. (2020, Sec. 5.1) for more possibilities, and to Andrews (1991, Theorem 1 (b)), who shows that even estimators with unbounded kernels may achieve our required rate of convergence.

Our first main result is the following.

Theorem 2.1. *Under and Assumptions 0–5,*

$$\underset{\to\infty}{\mathcal{M}} \qquad -$$

Proof. See Online Appendix A. □

Remark 1.

(a) Writing identifies the limiting distribution from Theorem 2.1 as Gumbel . We hence reject when the "scaled" $\mathcal{M}$ statistic exceeds $\mathcal{D}$, i.e., the -quantile of the Gumbel distribution. Note also that the convenient closed-form does not require simulation of critical values.

(b) For known parameters (or forecasts not involving parameter estimation, such as random walks or constant forecasts), Theorem 2.1 enables us to put uniform significance bounds around the time axis, thus allowing to detect episodes of different predictive accuracy: reject the null at level when, for some , the normalized rolling average exceeds $\mathcal{D}$. A graphically intuitive depiction consists of plotting the sequence of rolling averages (which estimate) against time, , and compare them with significance bounds constructed as

$$\quad \frac{\mathcal{D}}{} \qquad \quad \frac{\mathcal{D}}{} \tag{2.9}$$

(c) Thanks to the uniform validity of the significance bounds, Theorem 2.1 can also be used for *monitoring* equal predictive ability. Specifically, for a prospective sample to be observed at times , one simply has to compute successively the rolling means from (2.3) and reject as soon as $\mathcal{D}$ at some time . The only difference between the monitoring application and more usual EPA tests is the computation of . Specifically, one requires data suitable for the estimation of already at the start of the monitoring period. Without training data for , this could be for instance achieved by using only the data from the first window, , to obtain an initial estimator , and then updating as new data become available one by one.

(d) The difference in the vanishing- Gumbel asymptotics from Theorem 2.1 and the Giacomini and Rossi (2010) fixed- asymptotics (2.6) may also be viewed as one of small vs. fixed bandwidths prominent in related literatures such as HAC- vs. HAR-(or "fixed- ") inference as studied by Andrews (1991) and Kiefer and Vogelsang (2005). Of course, the $\mathcal{DM}$-statistic (2.2) is an application of the former approach, where Coroneo and Iacone (2020) study $\mathcal{DM}$ under a fixed-bandwidth, or fixed- approach. Kiefer and Vogelsang (2005) demonstrate that fixed- inference provides better size in a wide variety of situations. Interestingly, the situation is reversed in the present situation in that it is the small- asymptotics that provide better size control (cf. Section 3).

(e) The proof of Theorem 2.1 shows that a key intermediate step in obtaining the Gumbel

limit builds on strong convergence of partial sums of the form

$$\leq \leq \qquad (2.10)$$

And indeed, the mixing and stationarity conditions in Assumptions 3–4 allow us to invoke the weighted strong invariance principle of Kuelbs and Philipp (1980, Theorem 4) for the loss differences . Note that (2.10) may in fact be established under alternative sets of conditions, see, e.g., Eberlein (1986), Ling (2007) and Wu (2007), such that the validity of our approach does not hinge on the specific form of Assumption 3.

}

Remark 2. It may seem unusual in the context of predictive ability testing that a Gumbel limit arises for a functional of partial sums of the loss differences. To develop some intuition for this result, note that

$$\mathcal{M} \qquad (2.11)$$

$$\text{i.i.d.} \qquad (2.12)$$

where denotes a standard Wiener process. Here, the first step approximates the maximum over *all* rolling sums by that over only *non-overlapping* rolling sums, the second step exploits a functional central limit theorem, and the third step follows from well-known properties of Brownian motion. Moreover, we have for all (with a similar computation valid for) that, as ,

$$\qquad (2.13)$$

by classical results from extreme value theory for independent, identically distributed (i.i.d.) random variables (see, e.g., Leadbetter et al., 1983, Theorems 1.5.3 & 1.8.3). }

Remark 3. The scalings and from (2.8) are not unique and can be replaced by any other sequences ′ and ′ satisfying ′ and ′ , as . }

2.2.2 Convergence Under Local Alternatives

Since our test statistic and its limiting distribution are non-standard, it is of interest to investigate the local power of our test more formally. The analysis of genuinely time-varying differences in predictive ability is, however, complicated as results for extremes of *heterogeneous* sequences are substantially more difficult to obtain. We therefore gauge power under time-variation in simulations below, and present in the following the analysis of the case where the mean of the loss differentials is constant. This allows us to characterize *the types of local alternatives* against which power is expected to be nontrivial.

To do so, we now consider the specific data generating process with

$$\text{———} = \text{for all} \tag{2.14}$$

where , and simply normalizes the magnitude of the alternative. Since as , the *magnitude* of the alternative is suitably decreasing with the sample size.[2]

We now present our local power result.

Theorem 2.2. *Under and Assumptions 0–5,*

$$\underset{\to\infty}{\mathcal{M}} \qquad - \quad -$$

Proof. See Online Appendix B. □

For , Theorem 2.2 reduces to Theorem 2.1. However, for (and in particular irrespective of the sign of), Theorem 2.2 shows that $\mathcal{M}$ has non-trivial asymptotic local power in $^{-}$-neighborhoods of : its rejection probability is

$$\mathcal{M} \quad \mathcal{D} \qquad \underset{\to\infty}{\qquad} \qquad \mathcal{D}$$

where

$$- \quad - \qquad -$$

such that $\mathcal{D}$ $\mathcal{D}$.

For diverging , the asymptotic probability of rejecting the null converges to one. This is because, as ,

$$\mathcal{D} \qquad - \quad -\mathcal{D}_0 \; -$$

for any significance level .

[2] Since we consider a local alternative where vanishes as , the are strictly speaking a triangular array, such that ; for notational ease, we however suppress the dependence of on .

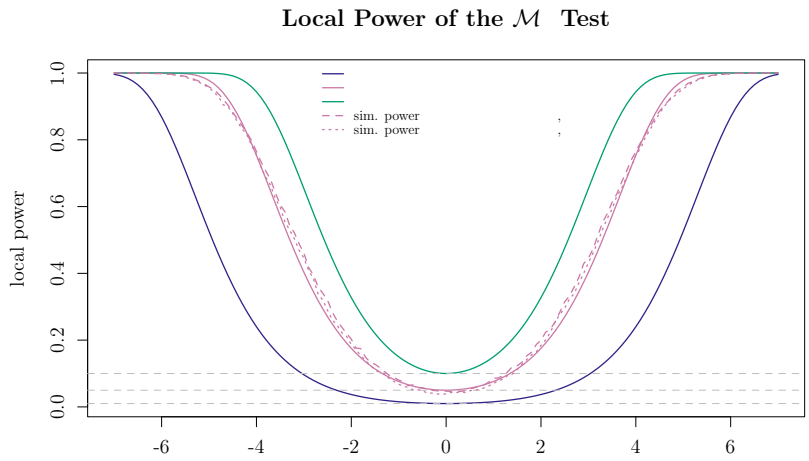

Figure 1: Asymptotic and simulated power curves for a difference of chi square loss differentials

Figure 1 illustrates these effects by plotting the asymptotic power curve of the scaled $\mathcal{M}$ test against for conventional significance levels . For the 5% nominal level, we additionally report simulated power curves for generated as mean-zero differences of variates normalized by two, the standard deviation of the difference, and shifted by as specified in (2.14). We simulate for large and choices of inspired by the rule of thumb to be discussed further in Section 3. We note the test's consistency already for intermediate values of . Moreover, the asymptotic power curve provides an accurate prediction of finite-sample power. A more thorough assessment of the properties of the $\mathcal{M}$ test under relevant forecast scenarios will be provided in simulations in Section 3.

Remark 4.

(a) Theorem 2.2 shows that we obtain local power against alternatives in $\bar{}$-neighbourhoods of . Thus, contingent on the chosen , local power may lie arbitrarily close to the $\bar{}$-local power of the usual $\mathcal{DM}$ or fluctuation test $\mathcal{M}$. Regarding the window width, we note that typical MSE-optimal window choices from nonparametric regression need not be optimal in this respect.

(b) Theorem 2.2 implies a size–power trade-off for the choice of the moving window size . While larger imply local power in ever smaller neighborhoods of the null, such large also tend to lead to poorer size, because the number of normal variates in (2.12) decreases in , such that the extreme value approximation in (2.13) becomes less reliable. An additional reason that size may deteriorate with increasing is that the approximation of the maximum in (2.11) is no longer accurate. To see this, consider the case of an extremely large , in which case the maximum over *all* rolling sums is approximated by the maximum over only the *two* non-overlapping

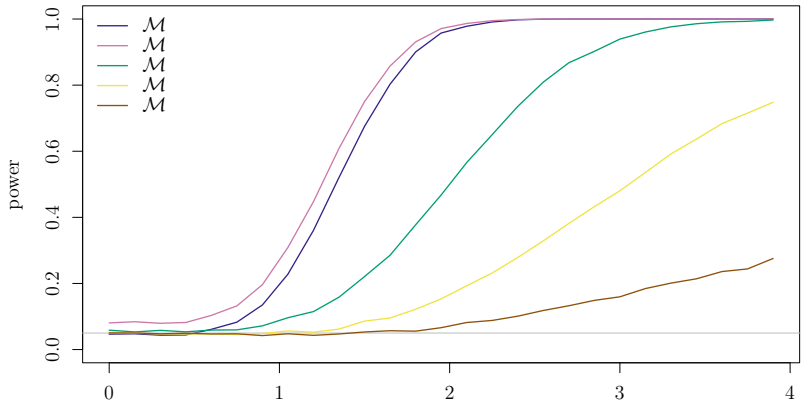

Figure 2: Power under transitory episodes of superior predictive ability (see Remark 5 for discussion of the experiment)

windows from and .

}

Remark 5. For intuition for power under short episodes of SPA, return to the scenario from $^{\text{trans}}$ in (2.4). We generate loss differentials, as in Figure 1, as under and shift by under $^{\text{trans}}$. For replications and , we set and construct $\mathcal{M}$ with the rule of thumb $_{\text{rot}}$ to be discussed further in Section 3. This implies here, and we set * . Figure 2 reveals that the Giacomini and Rossi (2010) $\mathcal{M}$ tests share some of the features of the average-based Diebold and Mariano (1995) for larger choices of in that evidence of SPA averages out over larger windows, leading to low power. When the window for the $\mathcal{M}$ test is chosen according to our rule of thumb ($\mathcal{M}$ $_{\text{rot}}$), power is competitive with our $\mathcal{M}$, albeit at the cost of some size distortions. Section 3 will present simulation evidence in more realistic setups. }

Remark 6. Figure 3 provides an alternative way to see the "local" nature of $\mathcal{M}$. We first use the asymptotic distribution in (2.6) to simulate critical values of the Giacomini and Rossi (2010) statistic $\mathcal{M}$ $\mathcal{M}_{\lfloor \ \rfloor}$ at levels for a fine grid . To accurately simulate such tail events, we opt for a large of 10 million and 100,000 replications when simulating the Brownian motions in (2.6). We then transform these critical values, equivalent to the event of a $\mathcal{M}$ test having -value , to the $\mathcal{M}$ statistic via the scaling factors from (2.8) with (the solid lines in Figure 3). The dashed lines report the Gumbel critical values $\mathcal{D}$. Note that, in the "vanishing " region of small , the implied $\mathcal{M}$

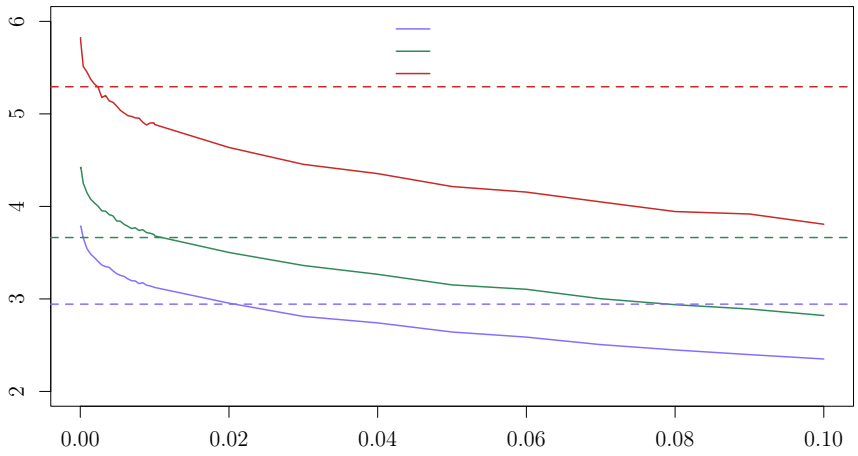

Figure 3: A small- analysis of $\mathcal{M}$ and $\mathcal{M}$. The experiment is discussed in Remark 6.

statistics exceed the $\mathcal{D}$, such that $\mathcal{M}$ would then reject, while $\mathcal{M}$ would (barely) not. Therefore, our joint limit (where jointly) leads to different critical values than the sequential limit (where first, and then). This suggests a crucial conceptual difference to Giacomini and Rossi (2010)-style tests. }

2.3 Dealing with Estimation Error

We next consider the impact of parameter estimation on our test statistic. Typically, evaluators do not have precise knowledge, if any, about which prediction model and/or estimators forecasters use. Hence, it would seem somewhat presumptuous to make specific assumptions on these, except of course when the purpose is to understand their impact (as in West, 1996). For this reason, we place a low-level requirement on the differences between the fitted and the ideal forecasts (see Assumption 6), which leads to a generic robustness statement. We furthermore require a strengthening of Assumption 4:

Assumption 4* (forecast errors under estimation). *The sequence of forecast errors* $'$ $\in$ *is strictly stationary.*

Assumption 6 (estimation impact on forecasts)**.** *For it holds that, as* ,

$$\text{———} = \tag{2.15}$$

Assumption 4* is stronger than the weak stationarity condition in Assumption 4. We impose it here to not only deduce the weak stationarity of the loss differences (as in Assumption 4), but also the weak stationarity of $\mathcal{L}$ for each .

The latter terms need to be sufficiently regular for estimation effects to vanish (see the proof of Theorem 2.3 below), and weak stationarity is a convenient (and likely relaxable) condition. Finally, Assumption 6 puts an upper bound on how estimators are allowed to impact the forecasts—without requiring details on how this impact occurs in detail. Note the flexibility afforded by the appearance of in (2.15). Due to this, even estimators with poor convergence rates may be accommodated by a suitably small choice of . Example 1 below illustrates this.

We have the following sufficient conditions for Assumption 6 to hold.

Proposition 1. *Suppose Assumptions W1–W4 (in Online Appendix C) hold for both forecasting models , estimators , and true parameters ( ). Then, Assumption 6 holds for with , where is defined in Assumption W1.*

Proof. See Online Appendix C. □

Assumptions W1–W4 correspond exactly to Assumptions 1–4 of West (1996), with the only difference that from West's (1996) Assumption 1(b) is required to satisfy (for some) in our Assumption W1 instead of as in West (1996). The point of Proposition 1 is to show that our Assumption 6 is often naturally satisfied—even in situations where estimation noise would otherwise affect the asymptotic limit (i.e., when); cf. West (1996, Theorem 4.1). In particular, by a suitable choice of , Assumption 6 can always be satisfied. Importantly, given sufficiently strong moment conditions in Assumptions W1(b) and W3(a) (i.e., large), Assumption 6 only requires minimal additional restrictions on the growth of beyond the requirement that , as already implied by Assumption 1.

To make Proposition 1 more concrete, consider the following example.

Example 1. *As a typical situation, consider two competing linear predictive regressions,*

′

*where , , are to be estimated (also) on the basis of pre-sample data, . This results in plugging in (possibly updated) estimators and forecasts ′ . Following West (1996), , and, mild regularity conditions assumed, we have from his Lemma A3(b) that − * uniformly in , where * may be arbitrarily small. Then,*

′ − *

where may, e.g., be bounded using moment properties of . More specifically, if is uniformly -bounded for some , then Markov's inequality yields that

. *Hence, choosing* *for* *is compatible with Assumption 6.*

We now denote by

$$\mathcal{M} \qquad \frac{-}{=} \qquad (2.16)$$

the feasible Giacomini and Rossi (2010) statistic using estimated loss differentials (2.5) and by

$$\mathcal{M} \qquad \mathcal{M} \qquad (2.17)$$

its scaled counterpart. We have the following result.

Theorem 2.3. *Under* *and Assumptions 0–3, 4* and 5–6,*

$$\underset{\to\infty}{} \quad \mathcal{M} \qquad -$$

Proof. See Online Appendix D. □

Comparing Theorems 2.1 and 2.3, we find—surprisingly—that $\mathcal{M}$ (based on the observable) has the same asymptotic limit as its oracle counterpart $\mathcal{M}$ (based on the unobserved). In this sense, $\mathcal{M}$ is robust to estimation effects. In particular, Theorem 2.3 implies that we may employ the same decision rule to reject from (2.1) at level when the "scaled and estimation-based" $\mathcal{M}$ statistic exceeds the -quantile of the Gumbel distribution, i.e., $\mathcal{D}$. Remarks 1(b) and (c) continue to hold.

Remark 7. To develop some intuition for why $\mathcal{M}$ is robust to estimation noise, recall from West (1996) that effects from model fitting vanish asymptotically when , i.e., the "full sample" evaluation window is a vanishing fraction of the estimation window . By a suitably slowly diverging , we ensure that our "subsample" evaluation window vanishes relative to for any . In other words, by considering small we create artificially small evaluation windows, such that estimation effects vanish even when aggregating the results on the length- windows to the full sample. }

Remark 8. Now, is computed based on rather than . Formally, we maintain Assumption 5 to derive the above result, such that the impact of estimation noise is implied to be controlled for when estimating the relevant long-run variance based on observed . In light of Assumption 6, this impact on is actually quite likely to be negligible, but a rigorous result in this respect requires us to specify which long-run variance estimator is used. For this reason we rather prefer to maintain Assumption 5. }

Example 1 (continued)**.** *When* *for* *and, hence, Assumption 6 is satisfied, then Theorems 2.1 and 2.3 imply (under the regularity conditions of Theorem 2.3)*

that $\mathcal{M}$ *and* $\mathcal{M}$ *have the same Gumbel limit—even when* . *In contrast, West (1996) and others show that in this case, the standard DM test statistic* $\mathcal{DM}$ *from (2.2) has a (Gaussian) limit different from that of*

$$\mathcal{DM} \quad -^{-} \tag{2.18}$$

in general. Similarly, Demetrescu et al. (2022) obtain distinct asymptotic limits for the fluctuation statistics $\mathcal{M}_{\lfloor \quad \rfloor}$ *and* $\mathcal{M}_{\lfloor \quad \rfloor}$*, leading to size distortions of tests based on the latter when estimation effects are not controlled for. Overall, this example highlights the robustness of our "vanishing " tests.*

Remark 9. To highlight the importance of the robustness guaranteed by Theorem 2.3 consider the following situation in the framework of Example 1. Let be generated as

where , and are mutually independent (and—to streamline the argument—we take them to be i.i.d. standard normal). Consider as competing forecasts

and

where and are fixed and equal. This, however, is not known to the evaluator. Nonetheless, for symmetry reasons, the two forecasts are equivalent in terms of predictive accuracy. (And, in this case, the DM test is the right tool to compare the EPA of the two forecast models/methods.) Consider now the case where each forecaster actually estimates and by regressing on using the same estimation procedure and the same window(s) of data. This too is not known to the evaluator, but, for symmetry reasons, the forecast methods (and also the forecast models, even if they are not comparable without the relevant information) are again equivalent in terms of predictive accuracy. Given that both forecast models and forecast methods are equivalent, the size of any test comparing the EPA of the two predictions should be held. Yet the sampling distribution of the DM test will depend in general on how (or indeed whether) and are estimated, such that size distortions arise when estimation effects are ignored. Moreover, taking these aspects into account (say along the lines of West, 1996) requires knowledge of the respective estimation steps—while a robust procedure such as ours does not. }

2.4 Time-varying volatility and dynamics

One well-known stylized fact of economic series is that many of them exhibit (slowly) varying variances (e.g., Stock and Watson, 2002; Campbell, 2007; Groen et al., 2013; Amado

and Teräsvirta, 2014). This often spills over to series of loss differentials; see Coroneo and Iacone (2020), Demetrescu et al. (2022), and Harvey et al. (2024). So we now examine our proposal's behavior in such environments. First, we argue that the baseline version is not robust to time-varying volatility and dynamics. This can in fact be seen from the proof of Theorem 2.1, where one key step is the strong approximation of partial sums of loss differentials by a standard Brownian motion; see (2.10). This Brownian motion limit is not given under time-varying volatility or time-varying dynamics in general; see the seminal work of Cavaliere (2005). Remark 2 gives an accessible view of the issue: under stationarity, the limiting distribution of $\mathcal{M}$ is asymptotically equivalent to the distribution of the maximum of certain i.i.d. variables. Yet under time-varying volatility (or under time-varying dynamics), the relevant increments are heterogeneous rather than i.i.d., where the maximum of heterogeneous variables has a different behavior in general.

To deal with time-varying variances in the context of EPA tests, Demetrescu et al. (2022) advocate a wild bootstrap procedure. However, their framework does not cover the "vanishing " case. Furthermore, the wild bootstrap also requires knowledge of the relevant estimation details in order to account for any estimation noise. While exploring a wild bootstrap in our setup is appealing, it is at the same time challenging. We leave the discussion of such an approach for further work, especially since, as show below, one can get by the issue of heteroskedasticity without bootstrapping. Coroneo and Iacone (2020) conduct a split-sample analysis to this end, while Harvey et al. (2024) account for time-varying volatility by variance smoothing; neither however consider estimation error.

Time-varying volatility is naturally captured by variance modulation; see Cavaliere (2005) again. In our setup, a model of the form follows under the null, where is a zero-mean, weakly stationary sequence. In fact, we follow the literature and work with a slowly varying ; see Assumption 2** below for details. If then may be directly interpreted as a *local* variance, and, denoting the long-run variance of by , may be interpreted as a *local* long-run variance.

This gives rise to a different limiting null behavior of the normalized partial sums of . One may still expect "nice" behavior where the continuous-time limit process is Gaussian. However, the limit process is no Brownian, but rather a *variance-transformed* Brownian motion; cf. Cavaliere (2005) and references therein. It is such limiting behavior that invalidates the result of Theorem 2.1 under time-varying variance.

Nevertheless, each *local* average is computed based on approximately homoskedastic variables, owing to the assumed slow variation of volatility. This motivates, in turn, local standardization of the rolling averages to deal with time-varying volatility. In fact, local standardization may be conducted more generally whenever the long-run variance (and not just the variance) varies slowly; see Remark 11 below. Summing up, to robustify against time-varying volatility, we may (like Harvey et al., 2024) adjust our statistic such that the

rolling averages are standardized *individually*:

$$\mathcal{M}^{\text{TVVD}} \qquad \bar{\qquad} \qquad \bar{=} \qquad \bar{}$$

where is an estimator of the *local* long-run variance (which, in the above case of time-varying variance, has a multiplicative structure,).

This prompts the following extensions of some of the previous assumptions:

Assumption 1** (window size under time-varying volatility). *The moving window size satisfies Assumption 1 and , as .*

Assumption 2** (time-varying volatility and moments). *The forecast errors are variance-modulated in the sense that , where with a piecewise uniform Lipschitz function, bounded away from zero, and $'_{\in}$ is uniformly - bounded for some .*

Assumption 3** (dependence of standardized forecast errors). *The standardized forecast errors $'_{\in}$ are -mixing with -mixing coefficients satisfying $\bar{}$ for some and from Assumption 2**.*

Assumption 4** (stationarity of standardized forecast errors). *The sequence of standardized forecast errors, $'_{\in}$, is strictly stationary.*

Assumption 5** (local long-run variance estimator). *There exists an estimator that satisfies $\bar{}\ \bar{}$ uniformly in , as .*

Assumption 1** puts a somewhat stricter condition on how fast compared to Assumption 1 stating only . An intuition behind this condition is that some minimal separation between our setup and the case where is required—but time-varying variance makes this somewhat more difficult than for Theorem 2.1, and therefore needs an additional restriction.

Assumption 2** imposes a modulation pattern on the forecast errors. As a convenient consequence, a modulated structure arises immediately for the loss differentials. In particular, it is easily checked that the loss differentials are given as , with and $\mathcal{L}$ $\mathcal{L}$. To deal with the nonstationarity implied by Assumption 2**, we may exploit the fact , which suggests to estimate the local long-run variance *componentwise*. Following Harvey et al. (2024), this results in a two-step testing procedure: first standardize the loss differentials locally, and then run the test on the basis of the locally standardized loss differential series. Concretely, in this highly empirically relevant case, one computes

$$\mathcal{M} \qquad \bar{\qquad} \qquad \bar{=} \qquad — \qquad (2.19)$$

where is a local variance estimator (obtained, e.g., via local smoothing of) and is a long-run variance estimator based on the locally standardized loss differentials, ; see Harvey et al. (2024) for more details.[3] Then, uniform consistence of and consistency of suffices to establish the high-level Assumption 5**.

The following theorem derives the limiting behavior of

$$\mathcal{M} \qquad \mathcal{M}$$

It implies asymptotic robustness of $\mathcal{M}$ under the null.

Theorem 2.4. *Under and Assumptions 0, 1**–5**, and 6,*

$$\underset{\to\infty}{} \quad \mathcal{M} \qquad -$$

Proof. See Online Appendix E. □

Remark 10. Alternatively, in contrast to (2.19), one may consider plugging in a local long-run variance estimator prior to computing the rolling averages, i.e.,

$$\mathcal{M}^* \qquad \underset{-}{} \quad \overline{=} \qquad -$$

Regardless of the structure of , we expect the same asymptotics for $\mathcal{M}^*$ as for $\mathcal{M}$. }

Remark 11. Assumption 2** may actually be extended to cover not only time-varying volatility, but also time-varying dynamics. The easiest way to accomplish that is to require the loss differences to satisfy that, as ,

$$\underset{\leq \leq}{} \quad \underset{-}{} \quad \lfloor \rfloor \quad \overline{\qquad}$$

where $\underset{\geq}{}$ is a standard Brownian motion, and $\underset{-}{}$ is a Lipschitz-continuous function on , and . Note that, in the case of a generic time-varying autocovariance function, the local long-run variance is not proportional to the local variance in general. Either way, the estimator needs to be consistent at a suitable rate, as specified in Assumption 5**. Local long-run variance estimators, as required by Assumption 5**, have been discussed in the literature; see, e.g., Belotti et al. (2023). }

[3] Harvey et al. (2024) also discuss a GLS-type transform (namely scaling by ^{2} rather than) of the loss differentials prior to applying the usual DM test; this however does not remove the heteroskedasticity but merely changes the shape of the modulation, and therefore does not lead to robustness in our framework.

3 Simulations

This section presents the results of a series of Monte Carlo experiments designed to highlight various relevant aspects in conducting inference regarding SPA. These include, of course, size and power, the latter both for the case of constant and time-varying superior predictive ability. We also aim to disentangle the effects of estimation uncertainty (cf. Section 2.3), and the (asymptotic) lack thereof. Additional simulations discuss extensions to machine learning-type applications.

Concretely, Section 3.1 revisits the data-generating process (DGP) studied by Demetrescu et al. (2022) to assess size and power in a scenario in which OLS estimation implies an orthogonality condition à la West (1996), such that estimation effects do not matter asymptotically for the Diebold and Mariano (1995) and Giacomini and Rossi (2010) fluctuation tests. (Recall that our approach is robust to estimation effects under our maintained assumptions.) Section 3.2 presents results for a DGP with endogenous right-hand-side variables as predictors, as studied in Section 5.2 of West (1996). This requires instrumental variable (IV) estimation, such that the lack of orthogonality of predictors and residuals implies asymptotic estimation effects whenever . Section 3.3, inspired by the machine learning-type applications of Escanciano and Parra (2026), discusses results for an extended version of the DGP of Section 3.1. We here estimate high-dimensional forecasting models using LASSO and ridge to gauge the properties of the tests under such schemes.[4] Section 3.4 extends Section 3.1 to also consider time-varying volatility as discussed in Section 2.4. Finally, Section 3.5 provides power results.

We compare the "rolling Gumbel" statistic $\mathcal{M}$ from (2.17) to the Diebold and Mariano (1995) statistic $\mathcal{DM}$ from (2.18), and the Giacomini and Rossi (2010) fluctuation statistic $\mathcal{M}$ from (2.16). In our simulations, we found a rule of thumb of $_{\text{rot}}$ to perform well across the different DGPs.[5] This rule implies, e.g., choices of for . In relative terms, it yields small . Finally, we consider the subsampling-based -statistic by Lan et al. (2024).

All comparisons are based on squared error loss $\mathcal{L}$ and all tests are conducted two-sided at the 5%-level. Throughout, we set the number of replications to . Where applicable, long-run variance estimators in all statistics rely on the defaults of the `lrvar` command in `R`'s `sandwich` package (Zeileis, 2004; Zeileis et al., 2020).

[4] Online Appendix F.2 additionally explores a simple nested DGP as in McCracken (2020).

[5] We use two bandwidths for the Giacomini and Rossi (2010) statistic. First, we use the "small"-bandwidth based on our rule of thumb $_{\text{rot}}$. We then find the critical value of their statistic by simulating the functional from (2.6) on a grid for significance levels , essentially refining their Table I. The test is then conducted by comparing the test statistic to the critical value for the closest to $_{\text{rot}}$. Second, we also consider a "large"- statistic where we choose . Since the "fixed-bandwidth" approach of Giacomini and Rossi (2010) works with a separate critical value for each , their test should, at least asymptotically, also be robust to the specific choice of .

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.034	0.064	0.102	0.034	0.024
100	50	0.030	0.061	0.099	0.032	0.022
200	50	0.023	0.068	0.074	0.035	0.026
300	50	0.021	0.074	0.086	0.032	0.020
50	100	0.068	0.061	0.142	0.048	0.030
100	100	0.048	0.051	0.112	0.036	0.024
200	100	0.048	0.056	0.101	0.036	0.034
300	100	0.040	0.060	0.104	0.040	0.032
50	200	0.084	0.047	0.149	0.027	0.034
100	200	0.073	0.045	0.132	0.035	0.033
200	200	0.058	0.052	0.122	0.039	0.034
300	200	0.057	0.046	0.114	0.034	0.027
50	300	0.096	0.046	0.161	0.033	0.035
100	300	0.074	0.046	0.139	0.042	0.030
200	300	0.062	0.045	0.112	0.029	0.031
300	300	0.072	0.043	0.135	0.036	0.038
50	400	0.096	0.042	0.171	0.034	0.029
100	400	0.083	0.050	0.144	0.041	0.035
200	400	0.079	0.043	0.136	0.035	0.036
300	400	0.066	0.058	0.126	0.045	0.036

Table 1: Empirical size of the rolling Gumbel test $\mathcal{M}$, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at and of the Lan et al. (2024) test, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Section 3.1.

3.1 Demetrescu et al. (2022): the OLS case

We investigate the simple and widely relevant case of regression-based prediction through competing univariate predictors. Concretely, we predict an ARMA(1,1)-process $_{-}$ $_{-}$, , through two competing AR(1)-processes $_{-}$, . Here, $\boldsymbol{u}$ $'$ is generated independently from a multivariate normal distribution with correlation matrix $\boldsymbol{\Upsilon}$ specified further below. The predictions of via the , and hence loss differentials to be used for all the test statistics, are obtained by simple (recursive) ordinary least squares (OLS).[6] This DGP assesses size and power in a scenario where OLS estimation implies an orthogonality condition between predictors and forecast errors such that estimation effects do not matter asymptotically for all the tests considered.

The size experiments use an equicorrelation matrix $\boldsymbol{\Upsilon}$ $\boldsymbol{\Upsilon}$ with identical off-diagonal

[6]Selected simulations for rolling estimation suggest qualitatively similar performance.

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.079	0.329	0.191	0.203	0.078
100	50	0.065	0.294	0.166	0.179	0.062
200	50	0.053	0.220	0.136	0.112	0.038
300	50	0.044	0.176	0.113	0.086	0.035
500	50	0.042	0.148	0.112	0.075	0.032
50	100	0.115	0.384	0.209	0.307	0.139
100	100	0.092	0.352	0.180	0.248	0.118
200	100	0.071	0.276	0.151	0.176	0.084
300	100	0.049	0.249	0.117	0.153	0.069
500	100	0.058	0.185	0.124	0.120	0.054
50	200	0.150	0.448	0.255	0.411	0.181
100	200	0.113	0.404	0.204	0.342	0.154
200	200	0.088	0.373	0.159	0.274	0.125
300	200	0.077	0.315	0.152	0.227	0.116
500	200	0.067	0.258	0.121	0.172	0.081
50	300	0.189	0.452	0.284	0.447	0.196
100	300	0.151	0.456	0.253	0.418	0.185
200	300	0.112	0.410	0.186	0.341	0.155
300	300	0.096	0.374	0.149	0.295	0.129
500	300	0.082	0.311	0.144	0.235	0.098

Table 2: Empirical size of the rolling Gumbel test $\mathcal{M}$, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at and of the Lan et al. (2024) test, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Section 3.2.

elements . This yields a scenario in which and have equal predictive ability for so that the null hypothesis of the tests is true. The corresponding power experiments are presented in Section 3.5. We consider the parameter grid , and .

Table 1 reports selected results for (unless indicated otherwise, omitted parameter combinations led to qualitatively similar results). Rejection rates are relatively insensitive to the estimation sample size for a given . All tests have relatively acceptable size here, although some differences emerge. $\mathcal{DM}$ is most accurate here, with $\mathcal{M}$ being somewhat undersized for smaller . In turn, the small- , or "local" $\mathcal{M}_{\text{rot}}$ Giacomini and Rossi (2010) test performs well for smaller , but has some notable upward size distortion for larger . The less local $\mathcal{M}$ test shares more of the properties of the global $\mathcal{DM}$ test. The test is somewhat undersized, but controls size across and here.

3.2 West (1996): instrumental variable estimation

Here, we consider a virtually identical setup as in West (1996, Section 5.2). Let $'$ be an i.i.d. normal random vector with an identity covariance matrix. The data are then generated as , where , . We consider the two competing predictive models for (simplified to omit West's constants):

(3.1)

(3.2)

where the regressors are correlated with the 's and . In particular, in (3.1) and (3.2). Each equation is estimated by 2SLS with instruments in (3.1) or in (3.2). That is, let _ ¯ _ be the 2SLS estimate of . For , one-step-ahead prediction errors are computed as .

In view of IV estimation of the forecast models' parameters, residuals and predictors are not orthogonal here, such that this DGP investigates the empirical relevance of parameter estimation effects on the size of the tests, in particular for cases where is large relative to . Here, we study the grid and .

Table 2 presents the results. We now observe strong size distortions for $\mathcal{DM}$ of up to 45%, highlighting its lack of robustness to estimation errors. These typically occur, in line with the estimation noise theory of West (1996) and others, when is large relative to . To a lesser, but still substantial extent, we also observe size distortions of the Giacomini and Rossi (2010) fluctuation tests, which rejects the null in up to almost 30 or 45% of the tests conducted for the small- and large- versions. Again, these mainly occur in situations with ratios far from zero.

By comparison, the accuracy of $\mathcal{M}$ is superior, especially in ranges of evaluation sample sizes typically encountered in empirical practice of up to 200. For larger , there are some size distortions as well, although they are still substantially smaller than those observed for the other tests, with for example a reduction of the null rejection rates of close to 50% relative to the Giacomini and Rossi (2010) fluctuation test.[7] The test performs similarly to $\mathcal{M}$ in these size experiments. Overall, the present results illustrate the usefulness of using an inferential approach robust to estimation effects.

[7] To further illustrate the relevance of estimation effects under nonorthogonality, Table F.2 in Online Appendix F provides counterfactual results in which estimated IV forecast functions are replaced with the true predictors $_1$ and $_2$ (recall the true coefficients are equal to one). In this infeasible scenario without estimation noise, all approaches control size, or at least have dramatically smaller size distortions in smaller samples.

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.042	0.150	0.122	0.080	0.048
100	50	0.029	0.134	0.086	0.060	0.042
200	50	0.019	0.102	0.069	0.044	0.042
300	50	0.016	0.093	0.063	0.039	0.037
500	50	0.012	0.098	0.057	0.039	0.046
50	100	0.080	0.176	0.158	0.092	0.081
100	100	0.059	0.150	0.127	0.081	0.082
200	100	0.048	0.120	0.114	0.060	0.066
300	100	0.040	0.102	0.095	0.047	0.066
500	100	0.034	0.110	0.090	0.058	0.067
50	200	0.124	0.220	0.200	0.144	0.103
100	200	0.090	0.216	0.160	0.124	0.110
200	200	0.067	0.182	0.117	0.094	0.094
300	200	0.065	0.153	0.123	0.088	0.086
500	200	0.064	0.143	0.113	0.072	0.083
50	300	0.159	0.276	0.241	0.172	0.127
100	300	0.121	0.232	0.192	0.149	0.121
200	300	0.086	0.214	0.146	0.118	0.118
300	300	0.074	0.203	0.128	0.106	0.120
500	300	0.069	0.164	0.118	0.086	0.100

Table 3: Empirical size of the rolling Gumbel test $\mathcal{M}$, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at and of the Lan et al. (2024) test, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Section 3.3.

3.3 Escanciano and Parra (2026): shrinkage estimators

This section aims to evaluate the performance of the tests when the losses are based on shrinkage estimators of more highly parameterized models. In view of their predictive effectiveness, such techniques are nowadays widely used in forecasting, but have recently also received increased attention in the EPA testing literature (e.g., Escanciano and Parra, 2026).

Similar to the scenario studied in Section 3.2, such shrinkage estimators typically violate the orthogonality condition between their forecast errors and predictors, unlike OLS-based estimators such as those studied in Section 3.1. It is hence of interest to study the extent to which the size of EPA tests is affected by this non-orthogonality.

We employ an extension of the setup from Section 3.1. Concretely, we use the same ARMA(1,1) DGP as in Section 3.1. Yet, for the forecasts, we generate an additional AR(1) regressor $_{-}$ without predictive content. We then recursively fit a LASSO

model for using lags of both and as well as a ridge model for using the same number of lags of and of . The fitted values of these models are finally used as predictions for to generate and .[8] We consider the sample size grid from the previous subsection, , , and report results for .

The results in Table 3 are qualitatively similar to those from the IV DGP from the previous Section 3.2: Again, $\mathcal{M}$ has fairly accurate size throughout. The large- Giacomini and Rossi (2010) $\mathcal{M}$ fluctuation test also performs somewhat satisfactorily, while the more local $\mathcal{M}_{\text{rot}}$ test is a little more size distorted. The $\mathcal{DM}$ test has the most serious size distortions. The $\mathcal{M}$ test also shares, albeit to a substantially smaller extent, some of these distortions for very large ratios ; has comparable size to $\mathcal{M}$.

Compared to the DGP from the previous Subsection 3.2, the size distortions of the competitor tests, while still marked or substantial, are somewhat smaller. This is likely due to the orthogonality condition between forecast errors and predictors being "less" violated by LASSO and ridge than IV, because the latter—as penalized versions of OLS—are still closer to satisfying the orthogonality conditions. Once more and again in line with the estimation noise theory of West (1996), the distortions of the other tests are more pronounced for the non-robust tests when is large relative to .

3.4 Demetrescu et al. (2022) with time-varying volatility

We now gauge the tests' capability to also control size under time-varying volatility. Concretely, we augment the DGP from Section 3.1 by multiplying the equicorrelation matrix $\boldsymbol{\Upsilon}$ with a scale factor after the break date, so that both the variance and the covariances increase by this factor. This models a permanent break in the covariance structure, while still maintaining equal predictive ability of the two predictors and considered in Section 3.1. The break occurs at , with the taken from Section 3.1 now acting as a variance rather than a time-varying predictive ability break.

We build on the approach of Harvey et al. (2024) discussed around (2.19) and estimate a local variance to standardize the loss differentials. Concretely, we estimate a local linear regression for the squared loss differentials using the default bandwidth of the `KernSmooth` (Wand, 2025) package in `R`, where we truncate estimated variances away from zero using a threshold of $^{-}$. The parameter grid is as in Section 3.1; in addition, we consider values . We now report the "robustified" Giacomini and Rossi (2010) statistic $\mathcal{M}$ from (2.19) as well as the corresponding scaled Gumbel statistic $\mathcal{M}$.

Table 4 reports results for a late ( ) upward ( ) variance break. The HAC-type $\mathcal{DM}$ statistic confirms its robustness to heteroskedasticity in this scenario. The

[8] For the LASSO, we select the shrinkage parameter as the value where the deviance of including a further predictor increases by less than $^{-3}$. Ridge predictions are made at a penalty parameter .

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$		$\mathcal{M}_{\text{rot}}$
50	50	0.132	0.044	0.158	0.019	0.016	0.342
100	50	0.115	0.044	0.150	0.028	0.012	0.340
200	50	0.130	0.036	0.160	0.025	0.008	0.335
300	50	0.117	0.039	0.144	0.022	0.015	0.333
50	100	0.104	0.030	0.122	0.031	0.009	0.487
100	100	0.102	0.038	0.127	0.032	0.010	0.494
200	100	0.094	0.036	0.118	0.036	0.012	0.506
300	100	0.100	0.041	0.116	0.039	0.012	0.508
50	200	0.074	0.038	0.090	0.044	0.012	0.616
100	200	0.063	0.034	0.081	0.045	0.006	0.615
200	200	0.076	0.037	0.098	0.046	0.011	0.601
300	200	0.070	0.039	0.088	0.050	0.011	0.601
50	300	0.058	0.044	0.074	0.052	0.008	0.650
100	300	0.059	0.040	0.074	0.062	0.009	0.675
200	300	0.061	0.039	0.079	0.064	0.008	0.678
300	300	0.049	0.032	0.062	0.058	0.006	0.687
50	400	0.051	0.038	0.070	0.068	0.009	0.740
100	400	0.052	0.040	0.083	0.063	0.009	0.752
200	400	0.045	0.040	0.065	0.056	0.009	0.742
300	400	0.038	0.044	0.063	0.066	0.011	0.744

Table 4: Empirical size of the rolling Gumbel test $\mathcal{M}$, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at , of the Lan et al. (2024) test and of the nonrobust Giacomini and Rossi (2010) $\mathcal{M}$ test at $_{\text{rot}}$, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Section 3.4, .

"local" maximum-type statistics $\mathcal{M}_{\text{rot}}$ and $\mathcal{M}$ similarly enjoy good size for sufficiently large , i.e., when the local standardization can be based on sufficiently large sample sizes. The less local $\mathcal{M}$ shares more of the characteristics of the HAC-robust $\mathcal{DM}$ statistic. The Lan et al. (2024) -statistic holds size, but is fairly undersized. The last column starkly illustrates the relevance of robustifying against time-varying volatility: when such volatility is present, the default non-robust Giacomini and Rossi (2010) statistic $\mathcal{M}_{\text{rot}}$ exhibits strong size distortions.

Table 5 provides further results for an early ( ) downward ( ) break. These additional results largely confirm the findings of Table 4. The statistic again holds size, suggesting a certain robustness to time-varying volatility at least for the DGP considered here. (Note that we did not attempt any steps to suitably standardize this statistic—a task that similarly appears manageable, but that we leave for future research.)

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$		$\mathcal{M}_{\text{rot}}$
50	50	0.097	0.064	0.118	0.069	0.013	0.264
100	50	0.092	0.058	0.114	0.070	0.012	0.282
200	50	0.096	0.058	0.115	0.076	0.012	0.286
300	50	0.087	0.059	0.109	0.071	0.008	0.299
50	100	0.055	0.053	0.070	0.075	0.014	0.381
100	100	0.057	0.054	0.070	0.068	0.008	0.392
200	100	0.058	0.052	0.073	0.081	0.008	0.399
300	100	0.058	0.049	0.072	0.073	0.010	0.397
50	200	0.038	0.047	0.050	0.084	0.009	0.461
100	200	0.031	0.050	0.046	0.087	0.009	0.452
200	200	0.036	0.052	0.052	0.087	0.010	0.432
300	200	0.034	0.049	0.047	0.082	0.011	0.433
50	300	0.024	0.047	0.038	0.094	0.007	0.510
100	300	0.028	0.046	0.044	0.091	0.015	0.512
200	300	0.025	0.049	0.042	0.081	0.010	0.495
300	300	0.031	0.049	0.047	0.087	0.011	0.512
50	400	0.035	0.050	0.057	0.089	0.014	0.570
100	400	0.029	0.044	0.053	0.084	0.008	0.576
200	400	0.028	0.053	0.050	0.096	0.013	0.560
300	400	0.022	0.051	0.045	0.090	0.009	0.548

Table 5: Empirical size of the rolling Gumbel test $\mathcal{M}$, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at , of the Lan et al. (2024) test and of the nonrobust Giacomini and Rossi (2010) $\mathcal{M}$ test at $_{\text{rot}}$, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Section 3.4, .

3.5 Demetrescu et al. (2022), power

After having established the properties of the various tests under the null of equal predictive ability in the previous subsections, we now turn to some power results. We do so for the DGP of Section 3.1 where all tests, in view of orthogonality, are in principle valid to make such a power comparison meaningful. In our power experiments, we specify three distinct scenarios.

First, in order to generate time-varying forecasting ability, we specify a simple switch from an equicorrelated (see the size experiments) matrix to a "Toeplitz" matrix $\boldsymbol{\Upsilon}$ at time . More specifically, after the break both and as well as and are correlated (with a correlation coefficient of) while and are uncorrelated. Thus, is independent of and therefore has no predictive power anymore, in contrast to . Hence, the structural break in predictive power emerges from a time-varying correlation matrix $\boldsymbol{\Upsilon}$. The later the break occurs, the less time the tests have to detect the emerging

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.033	0.066	0.105	0.032	0.023	0.028	0.058	0.096	0.029	0.024
100	50	0.026	0.066	0.095	0.033	0.028	0.024	0.062	0.084	0.031	0.023
200	50	0.034	0.084	0.104	0.038	0.027	0.025	0.066	0.089	0.030	0.022
300	50	0.027	0.072	0.082	0.034	0.033	0.018	0.059	0.072	0.027	0.027
50	100	0.094	0.096	0.181	0.042	0.041	0.081	0.054	0.153	0.032	0.026
100	100	0.078	0.101	0.170	0.049	0.052	0.060	0.062	0.128	0.031	0.033
200	100	0.080	0.104	0.163	0.047	0.046	0.054	0.052	0.138	0.030	0.035
300	100	0.089	0.100	0.189	0.053	0.054	0.057	0.057	0.134	0.032	0.034
50	200	0.160	0.140	0.266	0.074	0.061	0.123	0.072	0.200	0.034	0.045
100	200	0.170	0.161	0.278	0.103	0.061	0.109	0.078	0.197	0.039	0.040
200	200	0.179	0.176	0.271	0.106	0.074	0.120	0.077	0.192	0.037	0.042
300	200	0.192	0.177	0.292	0.123	0.074	0.120	0.084	0.209	0.040	0.045
50	300	0.236	0.235	0.328	0.188	0.094	0.164	0.088	0.254	0.038	0.048
100	300	0.274	0.234	0.380	0.210	0.095	0.160	0.087	0.248	0.045	0.051
200	300	0.274	0.248	0.386	0.235	0.106	0.166	0.096	0.253	0.052	0.050
300	300	0.277	0.264	0.390	0.248	0.100	0.166	0.104	0.252	0.055	0.044
50	400	0.337	0.300	0.467	0.294	0.114	0.224	0.106	0.330	0.048	0.056
100	400	0.348	0.332	0.475	0.352	0.125	0.223	0.111	0.316	0.056	0.046
200	400	0.374	0.370	0.498	0.386	0.139	0.238	0.120	0.332	0.070	0.046
300	400	0.383	0.356	0.500	0.392	0.145	0.245	0.128	0.348	0.084	0.059

Table 6: Power (time-varying superiority) for different across combinations of and . We consider the $\mathcal{M}$ test, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at and of the Lan et al. (2024) test, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Sections 3.1 and 3.5.

superior predictive ability of . This is anticipated to negatively affect all tests' power, but less so for tests such as ours or that of Giacomini and Rossi (2010) that are designed to capture time-varying predictive ability.[9]

Conversely, we also consider scenarios in which the superior predictive ability occurs at the beginning of the forecasting sample to then vanish after time .

Second, we consider a scenario with "mixed" time-varying forecasting ability. Here, the Toeplitz structure for $\boldsymbol{\Upsilon}$ is reversed in the beginning of the forecasting sample until , such that has predictive ability for , while does not. For , and have equal predictive ability. For the remainder of the sample, has predictive ability as in the first scenario, while ceases to do so.

[9]We refrain from simulating scenarios such as those where each predictor has an identical, but time-varying advantage over the other for half of the evaluation period each. In such cases, the power of the mean-based Diebold and Mariano (1995) test would trivially reduce to size.

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.019	0.065	0.076	0.027	0.029	0.023	0.085	0.080	0.041	0.036
100	50	0.019	0.064	0.080	0.033	0.033	0.026	0.086	0.096	0.034	0.035
200	50	0.024	0.066	0.080	0.033	0.028	0.021	0.083	0.087	0.040	0.030
300	50	0.022	0.063	0.085	0.025	0.028	0.022	0.086	0.089	0.038	0.034
50	100	0.073	0.067	0.154	0.026	0.042	0.085	0.103	0.184	0.051	0.055
100	100	0.062	0.063	0.148	0.032	0.040	0.084	0.115	0.171	0.057	0.062
200	100	0.062	0.070	0.132	0.032	0.042	0.086	0.120	0.180	0.061	0.058
300	100	0.065	0.060	0.143	0.032	0.033	0.084	0.120	0.180	0.070	0.069
50	200	0.138	0.082	0.215	0.054	0.046	0.197	0.199	0.309	0.161	0.094
100	200	0.144	0.071	0.228	0.042	0.044	0.212	0.191	0.322	0.150	0.074
200	200	0.143	0.089	0.227	0.039	0.051	0.211	0.202	0.315	0.158	0.084
300	200	0.119	0.093	0.195	0.041	0.042	0.198	0.204	0.304	0.164	0.089
50	300	0.202	0.102	0.294	0.062	0.055	0.310	0.272	0.408	0.284	0.115
100	300	0.195	0.104	0.282	0.060	0.058	0.303	0.273	0.413	0.276	0.116
200	300	0.208	0.110	0.297	0.067	0.052	0.297	0.307	0.402	0.291	0.127
300	300	0.208	0.104	0.296	0.063	0.065	0.295	0.314	0.406	0.301	0.129
50	400	0.285	0.127	0.371	0.086	0.060	0.416	0.362	0.535	0.422	0.136
100	400	0.278	0.138	0.371	0.090	0.057	0.389	0.376	0.518	0.416	0.151
200	400	0.261	0.126	0.364	0.082	0.066	0.410	0.397	0.532	0.432	0.154
300	400	0.266	0.133	0.372	0.081	0.066	0.405	0.391	0.532	0.457	0.163

Table 7: Power (early time-varying superiority) for different across combinations of and . We consider the $\mathcal{M}$ test, the Diebold and Mariano (1995) $\mathcal{DM}$ test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at and of the Lan et al. (2024) test, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Sections 3.1 and 3.5.

Third, we also consider a time-invariant "Toeplitz" structure for $\boldsymbol{\Upsilon}$ for all . Such a design is expected to favor the Diebold and Mariano (1995) test, as one predictor is consistently superior to the others.

The parameter grid considered again is , , for early to late breaks and .[10]

Table 6 provides power results for the case of "late" time-varying superiority of .[11] We

[10]Note that is the upper bound for to ensure a positive definite $\boldsymbol{\Upsilon}$ in the power scenarios.

[11]Results for the global superior predictive ability case are presented in Table F.1 in Online Appendix F. While has no relevant effect here, as expected, all tests' power increases in . It also, again intuitively, increases in as the superior predictor then has stronger correlation with the target variable. Similarly, Diebold and Mariano (1995) clearly is most powerful in this case as its mean-based design allows to exploit the superiority across all . The power of the nonlocal Giacomini and Rossi (2010) $\mathcal{M}_{0\,4}$ test comes second, again because it shares many of the properties of the global $\mathcal{DM}$ test, followed by . The

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.025	0.059	0.083	0.028	0.031	0.022	0.051	0.073	0.023	0.024
100	50	0.021	0.051	0.077	0.026	0.031	0.020	0.048	0.072	0.026	0.023
200	50	0.018	0.061	0.064	0.029	0.024	0.016	0.052	0.063	0.028	0.022
300	50	0.024	0.058	0.082	0.025	0.021	0.022	0.053	0.070	0.030	0.022
50	100	0.054	0.044	0.121	0.022	0.028	0.066	0.044	0.155	0.032	0.030
100	100	0.055	0.049	0.126	0.025	0.032	0.066	0.042	0.152	0.029	0.032
200	100	0.051	0.052	0.110	0.029	0.030	0.062	0.036	0.148	0.024	0.027
300	100	0.047	0.040	0.114	0.025	0.028	0.067	0.038	0.143	0.028	0.030
50	200	0.099	0.040	0.186	0.026	0.036	0.153	0.036	0.259	0.043	0.030
100	200	0.093	0.045	0.170	0.031	0.034	0.146	0.042	0.239	0.037	0.033
200	200	0.101	0.038	0.169	0.029	0.036	0.155	0.033	0.254	0.043	0.034
300	200	0.100	0.040	0.167	0.024	0.037	0.156	0.036	0.254	0.036	0.033
50	300	0.156	0.050	0.240	0.038	0.032	0.257	0.023	0.360	0.054	0.029
100	300	0.162	0.034	0.240	0.031	0.037	0.245	0.032	0.339	0.057	0.028
200	300	0.159	0.042	0.242	0.032	0.030	0.243	0.034	0.355	0.053	0.025
300	300	0.156	0.044	0.226	0.036	0.032	0.264	0.031	0.377	0.063	0.033
50	400	0.220	0.048	0.321	0.038	0.035	0.355	0.031	0.477	0.076	0.034
100	400	0.206	0.040	0.305	0.039	0.038	0.338	0.031	0.468	0.085	0.029
200	400	0.225	0.036	0.317	0.038	0.033	0.342	0.030	0.460	0.082	0.039
300	400	0.210	0.034	0.302	0.028	0.038	0.341	0.029	0.472	0.074	0.033

Table 8: Power (mixed time-varying superiority) for different across combinations of and . We consider the $\mathcal{M}$ test, the Diebold and Mariano (1995) $\mathcal{DM}$ test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $_{\text{rot}}$ as well as at and of the Lan et al. (2024) test, at nominal level for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Sections 3.1 and 3.5.

report results for the case of a relatively late break .[12] Again, the effect of is negligible for the sample sizes considered here.

Interestingly and in line with the underlying idea of the different tests, we however now observe that the time-varying tests ($\mathcal{M}$ and small- Giacomini and Rossi (2010)) are, for smaller , competitive with the Diebold and Mariano (1995) and its cousin $\mathcal{M}$ test. They in fact outperform the latter for larger , in particular when the break occurs late. While $\mathcal{M}_{\text{rot}}$ rejects more often than $\mathcal{M}$, it has to be recalled that the former has relevant upward size distortions (cf. Table 1) such that some or all of these additional rejections

$\mathcal{M}_{\text{rot}}$ statistic has small advantage over $\mathcal{M}$. Some or most of this superiority is also driven by the differences in size discussed previously.

[12]For a relatively early break, the DGP has longer periods of superior predictive ability and thus gets closer to the case considered under global superiority again, such that results are then qualitatively more similar to those of Table F.1 in Online Appendix F.

	DAR	DARM	IAR	NC	SPF
DAR	124		0		4.0659
DARM		NA			
IAR			124		4.0659
NC	1.0003		1.0003	124	4.8329
SPF	1.8540		1.8540	2.2940	125

Table 9: Loss differences for CPI at horizon 1 and vintage 1. Above diagonal: for the pair. Below diagonal: for the pair. Main diagonal: Evaluation sample size for this method.

are spurious. We also note that has low power in these experiments, maybe because its resampling procedure, under late breaks in predictive ability, mostly mixes observations from periods of equal predictive ability, leading to small test statistics.

Tables 7 and 8 provide results for the mirror-image case in which ceases to have superior predictive ability relatively early in the forecasting sample and for the mixed time-varying superiority case. The results are qualitatively similar to those of Table 6 in that $\mathcal{M}$, $\mathcal{DM}$ and struggle to identify these short episodes of SPA. Again, on a size-adjusted basis, $\mathcal{M}_{\text{rot}}$ and $\mathcal{M}$ emerge as equally attractive options here.

Obviously, all tests have lower power in the time-varying predictive ability cases than in the case of global superiority, as the effective period of superiority is now smaller, which makes it harder for all tests to detect it.

Overall, the results of Section 3 suggest that there is a potential benefit of using the rolling Gumbel tests $\mathcal{M}$ and $\mathcal{M}$ when the analyst suspects estimation effects, time-varying volatility and/or that the superiority of one forecasting approach over another is not pervasive, but only applies over some potentially small fraction of the evaluation period. The higher power of the small- Giacomini and Rossi (2010) test over $\mathcal{M}$ is likely driven by the differences in size discussed previously, and only $\mathcal{M}$ provides good size control over all scenarios of potential estimation effects.

4 Illustration: the Survey of Professional Forecasters

This section applies the inferential techniques discussed above to the Survey of Professional Forecasters (Croushore, 1993) published by the Federal Reserve Bank of Philadelphia. As is well-known, starting in 1968, the SPF is the oldest quarterly survey of macroeconomic forecasts in the United States. More specifically, we consider forecasts for real GDP, inflation, unemployment and housing starts.[13]

Along with the SPF forecasts, we obtain their no-change (NC), indirect autoregressive

[13]Forecasts and realizations for these (and other) variables are provided at https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/error-statistics.

	DAR	DARM	IAR	NC	SPF
DAR	217	0.0033	0	0.0018	0.0115
DARM	0.0016	214			0.0084
IAR		0.0016	217	0.0018	0.0115
NC	0.0016	0.0015	0.0016	217	0.0097
SPF	0.0030	0.0019	0.0030	0.0023	227

Table 10: Loss differences for housing at horizon 1 and vintage 1. See notes to Table 9.

	DAR	DARM	IAR	NC	SPF
DAR	225		0		19.6631
DARM		NA			
IAR			225		19.6631
NC	11.8700		11.8700	225	31.3613
SPF	17.1414		17.1414	28.5805	226

Table 11: Loss differences for real GDP at horizon 1 and vintage 1. See notes to Table 9.

(IAR), direct autoregressive (DAR) as well as direct autoregressive with monthly information (DARM) forecasts. Each of these methods is used to produce 1- to 5-step ahead forecasts, which can each be used to predict five different realizations, or vintages of the data. That is, we may consider $\bar{}$ tests of EPA per variable, horizon and vintage, giving rise to hypotheses to be tested.[14] We consider the five $\mathcal{M}$, $\mathcal{DM}$, $\mathcal{M}_{\text{rot}}$, $\mathcal{M}$ and statistics introduced previously, as well as the time-varying volatility robust versions $\mathcal{M}$, $\mathcal{M}_{\text{rot}}$ and $\mathcal{M}$. We use data as available through to February 2026.[15]

This leads to an evaluation sample size of in the range of 124–227 (cf. Tables 9–12).[16] Since the SPF forecasts are proprietary, we have no information on the estimation sample size nor on the various methods used for constructing the forecasts. This highlights the attractiveness of an inferential approach that is robust to the lack of knowledge about the method and the sample size used to construct a specific forecast. We work with squared error loss so that Assumption 0 is satisfied.

Figure 4 presents loss differences (2.5) as well as their rolling mean (2.3) for comparing the different methods for forecasting housing at horizon and vintage 1. (Figures G.9–G.11 in Online Appendix G provide results for other pairs of horizons and vintages.) We additionally report level- ( ) significance bounds discussed in Remark 1(b).[17] We for example observe superior predictive ability (a negative loss differential of the

[14]The effective number is slightly lower, as the indirect and direct autoregressive forecasts coincide at horizon 1, so that for these tests. Hence, tests of EPA are moot here. Also, the DARM forecasts are not available for GDP and inflation.

[15]Missing values for either observations or forecasts are dropped. Table G.1 in Online Appendix G provides an overview of the proportion of missing values, showing that most series are complete, with the proportion of dropped data points never exceeding 3.2%.

[16]Qualitatively very similar results for other horizons and vintages are available upon request.

[17]Some pairs are not available, cf. footnote 14.

	DAR	DARM	IAR	NC	SPF
DAR	227	1.9989	0	1.5563	2.0828
DARM	2.2833	227	−1.9989	−0.4427	0.0838
IAR		2.2833	227	1.5563	2.0828
NC	1.5191	0.3218	1.5191	227	0.5265
SPF	2.2814	0.0851	2.2814	0.4196	227

Table 12: Loss differences for unemployment at horizon 1 and vintage 1. See notes to Table 9.

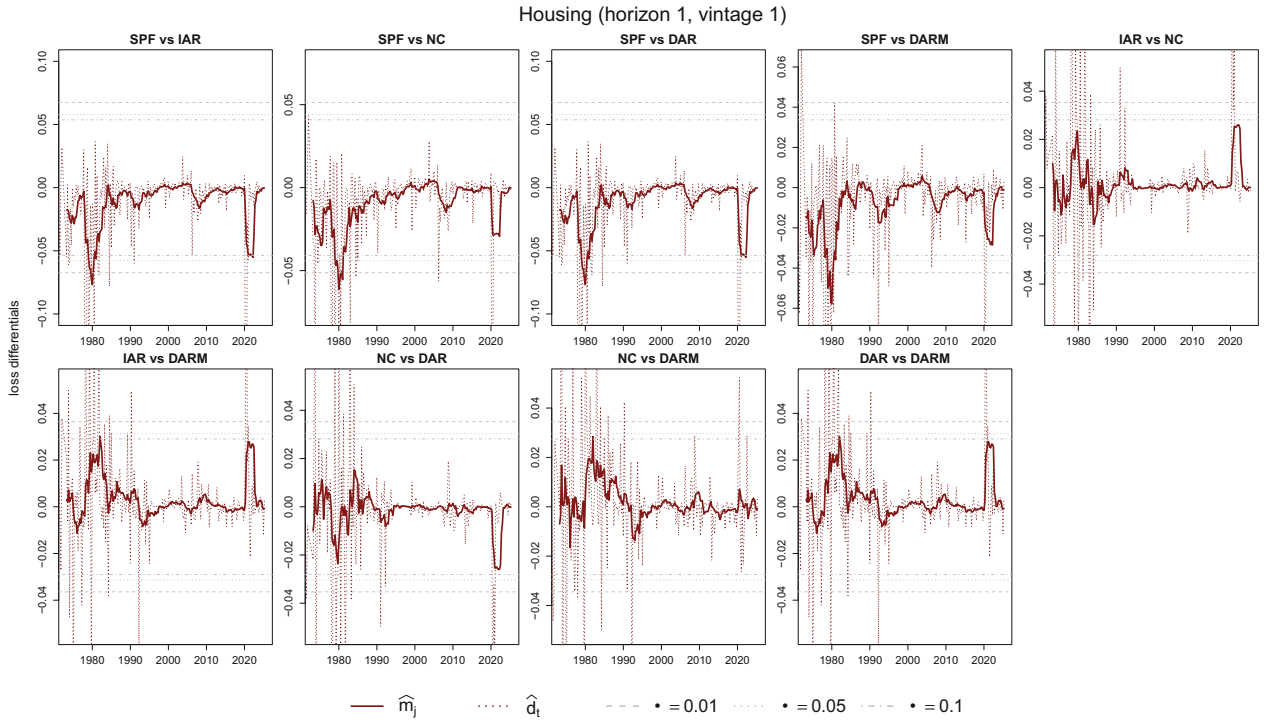

Figure 4: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

first-mentioned method in the plot implies that it outperforms its second-mentioned competitor) of the SPF vs. the other forecasts. Figure 4 also shows that the loss differentials $\widehat{d}_t$ for all comparisons are substantially more volatile until about the mid-1980s (i.e., the onset of the "Great Moderation"), before stabilizing afterwards. This suggests that tests accounting for time-varying volatility (cf. Section 2.4) may be attractive for inference on housing starts.

Figure 4 also highlights the precise source of this superiority. For example, the superior forecasts of the SPF vs. other methods largely stem from recessionary periods due to the oil price crises in the 1980s, as well as the COVID pandemic of the early 2020s. The latter, interestingly, is however not sufficient to achieve significance vis-à-vis all competitors. For the remainder of the evaluation period, the methods perform broadly similarly. Hence, global tests such as $\widehat{\mathcal{DM}}$ might struggle to identify these relatively brief episodes.

Remark 12. Online Appendix G provides additional results for the other variables

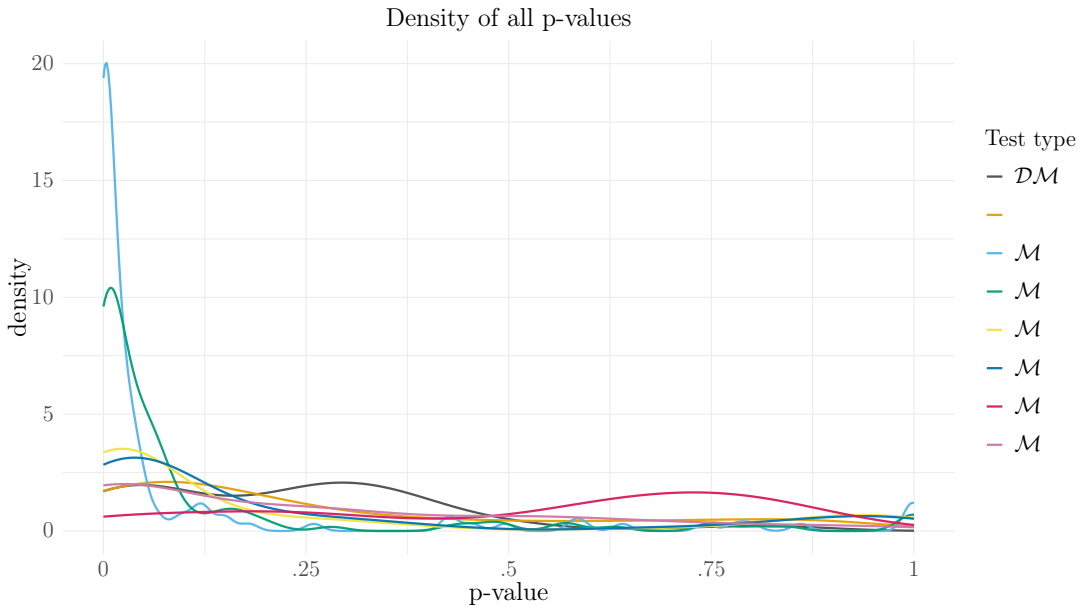

Figure 5: The overall distribution of -values of the EPA tests

considered. E.g., we observe similar patterns in that the SPF reacts more quickly to the COVID pandemic when forecasting unemployment (Figures G.1–G.4) or real GDP (Figure G.5–G.8). In other periods, relatively naive models such as NC forecasts are more similar to those of the SPF, also underscoring the rather unique nature of the 2020 crisis. We however add that this visual impression is partly driven by the scale of the significance bounds that are always included in the plots. Essentially, the long-run variance estimates inflate for series exhibiting such large jumps relative to the remaining variation (compare for example the lower diagonal elements of Tables 9 and 10 with those of Tables 11 and 12), leading to large bounds (2.9). From the above-diagonal elements of Tables 9–12, we see that, on average, the SPF typically performs better than the alternative methods.

For unemployment, augmenting the DAR forecast with monthly information (DARM) renders its forecasts almost competitive with the SPF. This may be expected, since it suggests that professional forecasters incorporate such real-time information in their predictions. For inflation (Figures G.12–G.15), evidence of SPA seems to be largely driven by the financial crisis starting in 2008, next to a smaller impact due to the recent inflationary period following the Russian invasion of Ukraine. }

Figure 5 reports the overall distribution of the -values of the eight above-mentioned tests of EPA.[18] To make such a large number of -values comparable, we provide kernel-density estimated distributions.[19] Recall that the previous discussion implies that rejections

[18]For the Giacomini and Rossi (2010) test, we fit splines to the non-standard distributions simulated in footnote 5 which we then use to compute -values $\mathcal{M}$.

[19]The slight dips of the estimates close to zero are an artifact of the boundary bias of the default kernel-density estimator.

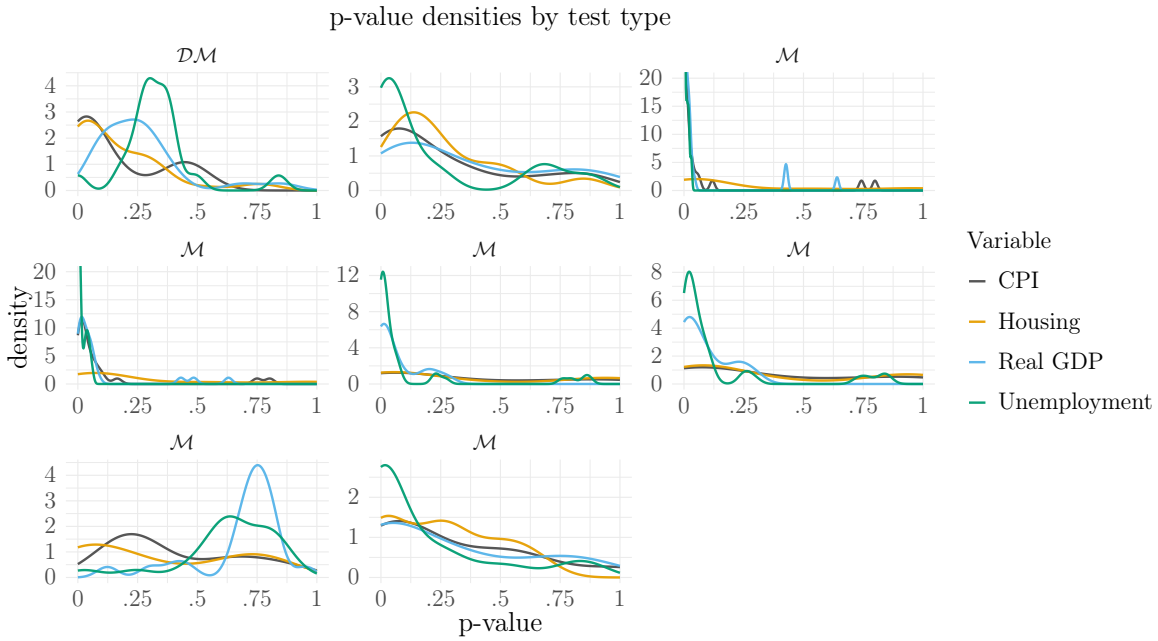

Figure 6: The distribution of -values of the EPA tests by test type

can only be safely taken as evidence against the null for our Gumbel-based tests.

The figure demonstrates that it is mainly the "local" statistics $\mathcal{M}$, $\mathcal{M}_{\text{rot}}$, $\mathcal{M}$ and $\mathcal{M}_{\text{rot}}$ that manage to identify superior predictive ability. Of these, the -values of the volatility-robust versions $\mathcal{M}$ and $\mathcal{M}_{\text{rot}}$ have less probability mass in regions of conventional significance; we return to this issue when discussing Figure 7 below. By contrast, the more "global" statistics $\mathcal{M}$, $\mathcal{M}$, $\mathcal{DM}$ and mostly produce -values outside the significant range. This suggests that the superior predictive ability of one approach versus another is typically not pervasive, but temporary.

Figure 6 differentiates according to test types. It again confirms the higher rejection rates revealed previously, and additionally demonstrates that most of the significant -values are found for the consumer price index (CPI), unemployment and real GDP.[20] The distribution is more uniform for housing, and hence compatible with type-I errors in the case of rejections for these variables.

Figure 7 splits -values according to variables. One interesting pattern emerges regarding the distribution of the volatility-robust local versions $\mathcal{M}$ and $\mathcal{M}_{\text{rot}}$ vs. the non-robust $\mathcal{M}$ and $\mathcal{M}_{\text{rot}}$. Concretely, the distribution is more concentrated in small -values for the latter group (with the exception of housing), which may indicate time-varying volatility leading to some type-I errors when not accounted for (cf. Section 3.4).[21]

[20]The $\mathcal{DM}$ statistic is a slight exception in this regard.

[21]However, Figure G.16 in Online Appendix G also reveals some evidence of "mixed signals", in that -values are not *systematically* smaller for the $\mathcal{M}$-tests. We rather occasionally also observe small -values for the $\mathcal{M}$ tests and large ones for the $\mathcal{M}$-tests. In practice, if the analyst were not to choose the robust versions throughout, we would recommend pre-testing for time-varying volatility prior to choosing a specific

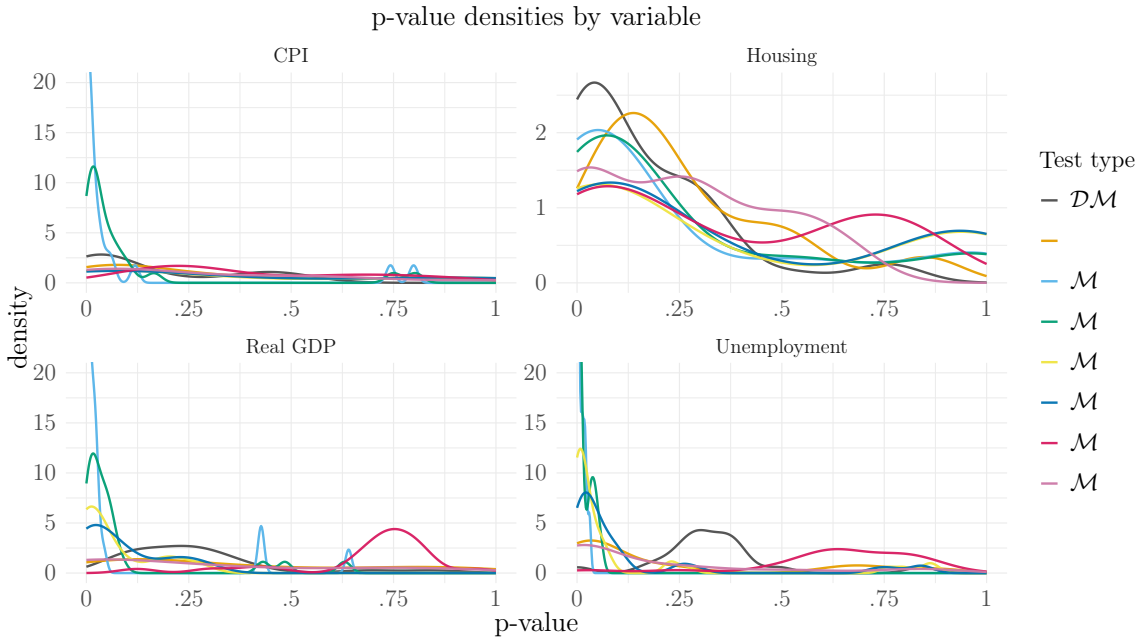

Figure 7: The distribution of -values of the EPA tests by variable

Figure 8 demonstrates that, plausibly, the clearest statistical evidence of SPA arises from comparing SPF forecasts with naive no-change predictions. Finally, Figure 9 compares results across different horizons. It shows that most of the evidence for SPA is present at short horizons (for those tests that do find evidence for SPA).[22]

test for EPA when aiming to capture time-varying forecasting ability.

[22]Figure G.17 in Online Appendix G, comparing results for the first and last vintage, reveals that the vintages have a relatively minor impact on the overall results.

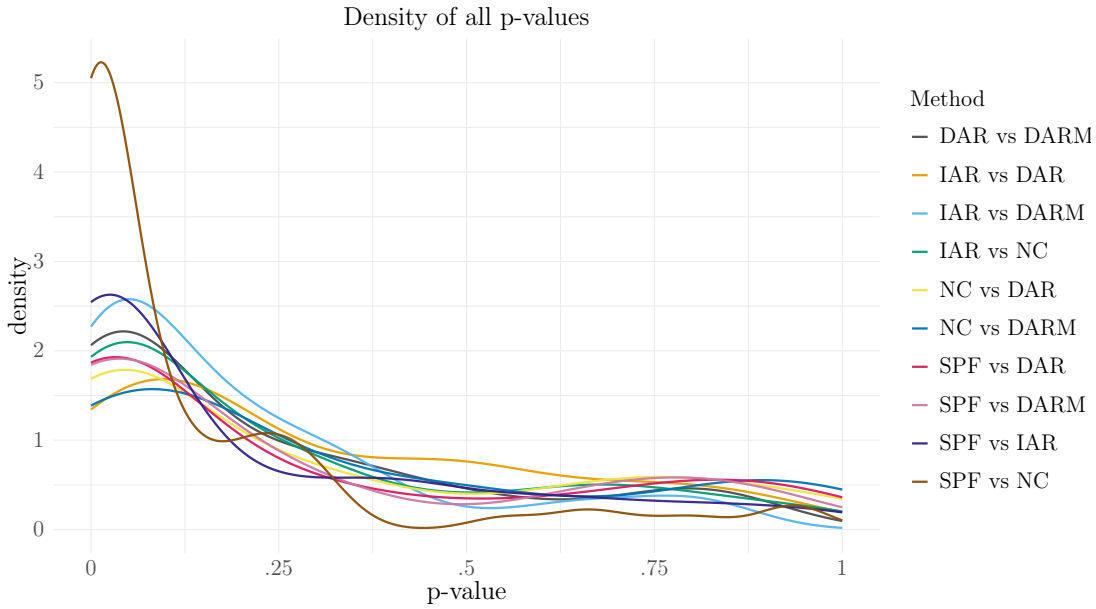

Figure 8: The distribution of -values of the EPA tests by methods

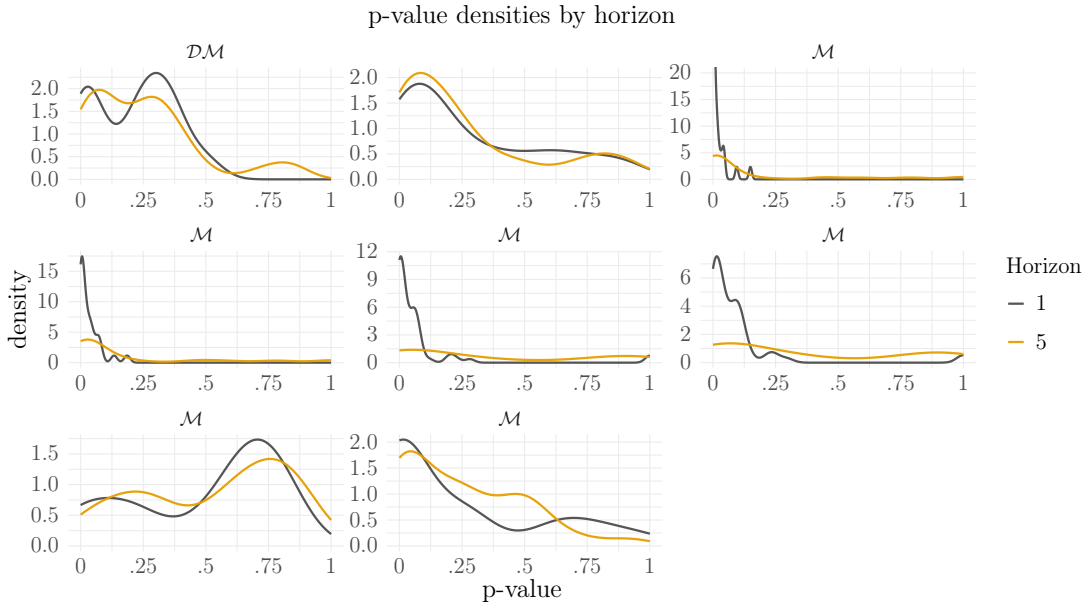

Figure 9: The distribution of -values of the EPA tests by horizon

5 Conclusion

This paper proposes a maximum rolling-window test to assess EPA, where the window size is only a vanishing fraction of the evaluation sample. Therefore, our work is most closely related to the fluctuation test of Giacomini and Rossi (2010), where the difference lies in the length of the moving windows. We show that our test tends to a Gumbel limit asymptotically under weak conditions. Both tests share the property that they are capable of detecting short episodes of superior predictive ability. Our test is also similar to the maximum subsampling test of Lan et al. (2024), where the main distinction is in the choice of the windows—*stochastic* in their case vs. *deterministic* and rolling in our case. The unique feature of our test is that it holds size even in the absence of *any* knowledge on how the forecasts were generated. (Additional results also robustify the test against time-varying volatility.) Of course, this situation is prevalent in empirical work that seeks to identify the most able forecaster among several professional competitors. We emphasize that our test may also be useful when knowledge on forecast generation is available, because correcting for estimation effects *à la* West (1996) is cumbersome and, hence, frequently avoided in practice.

Indeed, rather than accounting explicitly for estimation error, researchers often invoke West (1996) and Giacomini and White (2006)-style asymptotics—showing that estimation effects vanish when —even when empirically is large. E.g., Elliott and Timmermann (2008, Sec. 8) use and , and Li and Patton (2018, Sec. 7) consider and in a "vanishing " framework. In such cases, our simulations suggest that large-sample asymptotics ignoring estimation effects may not be particularly reliable—in contrast to our proposal. Our application to a SPF dataset finds that it is mainly tests that are engineered towards detecting time-varying predictive ability, like ours—rather than "global" tests of the Diebold and Mariano (1995) type—that find evidence against the null hypothesis of EPA.

Finally, we note that our test may be used "as is" when constructing model confidence sets (Hansen et al., 2011) for several competing forecasts, and it could be used to examine hypotheses of equal conditional predictive accuracy like in Giacomini and White (2006) by simply leveraging the loss differentials with suitable test functions. Furthermore, even if we do not pursue the topic here, it seems likely that the conditional superior predictive ability test of Li et al. (2022) may too be robustified along our lines. All in all, we strongly believe that our easy-to-apply test, whose critical values are even available in closed form, will find widespread use in the forecasting community.

References

Amado, C. and T. Teräsvirta (2014). Modelling changes in the unconditional variance of long stock return series. *Journal of Empirical Finance 25*(1), 15–35.

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

Belotti, F., A. Casini, L. Catania, S. Grassi, and P. Perron (2023). Simultaneous bandwidths determination for DK-HAC estimators and long-run variance estimation in nonparametric settings. *Econometric Reviews 42*(3), 281–306.

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. (2005). Unit root tests under time-varying variances. *Econometric Reviews 23*(3), 259–292.

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

Clark, T. E. and M. W. McCracken (2009). Tests of equal predictive ability with real-time data. *Journal of Business & Economic Statistics 27*(2), 441–454.

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*.

Demetrescu, M., C. Hanck, and R. Kruse-Becher (2022). Robust inference under time-varying volatility: A real-time evaluation of professional forecasters. *Journal of Applied Econometrics 37*(5), 1010–1030.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics 33*(1), 1–9.

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

Eberlein, E. (1986). On strong invariance principles under dependence assumptions. *The Annals of Probability 14*(1), 260–270.

Eichinger, B. and C. Kirch (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli 24*(1), 526–564.

Elliott, G., I. Komunjer, and A. Timmermann (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies 72*(4), 1107–1125.

Elliott, G. and A. Timmermann (2008). Economic forecasting. *Journal of Economic Literature 46*(1), 3–56.

Escanciano, J. C. and R. Parra (2026+). Extending the scope of inference about predictive ability to machine learning methods. Forthcoming in the *Journal of Business & Economic Statistics*, 1–30.

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.

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.

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 Y. Zu (2024). Tests for equal forecast accuracy under heteroskedasticity. *Journal of Applied Econometrics 39*(5), 850–869.

Horváth, L., C. Miller, and G. Rice (2020). A new class of change point test statistics of Rényi type. *Journal of Business & Economic Statistics 38*(3), 570–579.

Horváth, L. and G. Rice (2024). *Change Point Analysis for Time Series*. Cham, Switzerland: Springer.

Hušková, M. and A. Slabỳ (2001). Permutation tests for multiple changes. *Kybernetika 37*(5), 605–622.

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

Kirch, C. and P. Klein (2023). Moving sum data segmentation for stochastic processes based on invariance. *Statistica Sinica 33*(2), 873–892.

Kuelbs, J. and W. Philipp (1980). Almost sure invariance principles for partial sums of mixing -valued random variables. *The Annals of Probability 8*(6), 1003–1036.

Lan, W., B. Lei, L. Feng, and C.-L. Tsai (2024). Maximum-subsampling test of Equal Predictive Ability. *Journal of Business & Economic Statistics 42*(4), 1344–1355.

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

Li, J., Z. Liao, and R. Quaedvlieg (2022). Conditional superior predictive ability. *The Review of Economic Studies 89*(2), 843–875.

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

Ling, S. (2007). Testing for change points in time series models and limiting theorems for NED sequences. *The Annals of Statistics 35*(3), 1213–1237.

McCracken, M. W. (2020). Diverging tests of equal predictive ability. *Econometrica 88*(4), 1753–1754.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics 38*, 796–809.

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

Wand, M. (2025). *KernSmooth: Functions for Kernel Smoothing Supporting Wand and Jones (1995)*. R package version 2.23-26.

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

Wu, W. B. (2007). Strong invariance principles for dependent random variables. *The Annals of Statistics 35*(6), 2294–2320.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software 11*(10), 1–17.

Zeileis, A., S. Köll, and N. Graham (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software 95*(1), 1–36.

Online Appendix to "A robust test for equal predictive accuracy"

Matei Demetrescu , Christoph Hanck , and Yannick Hoga
TU Dortmund University
University of Duisburg-Essen

May 21, 2026

Abstract

Throughout, references to equations starting with an "(A.xx)", "(B.xx)", etc. refer to this appendix, whereas references without this prefix refer to the main paper. This supplement contains the proof of all theoretical results. Specifically, we prove Theorem 2.1, Theorem 2.2, Proposition 1, Theorem 2.3, and Theorem 2.4 in Sections A, B, C, D, and E, respectively. Section F provides additional Monte Carlo simulations and Section G presents further results for the empirical application.

A Proof of Theorem 2.1

The proof of Theorem 2.1 requires the following lemma.

Lemma 1. *Suppose that* ( ) *is a standard Brownian motion, and* 0, *as* . *Then, for all* ,

$$\lim \mathrm{P} \quad \sup \quad \frac{(+) \quad ( )}{\sqrt{}} \quad + \quad = \exp \quad 2$$

i.e., the centered and scaled supremum converges in distribution to a Gumbel law, where and are as in (2.8) *with* = 1 1.

Proof. By the scaling property of Brownian motion, it holds that

$$\begin{aligned} \sup \quad \frac{(+) \quad ( )}{\sqrt{}} &= \sup \quad (\quad + 1) \quad (\quad) \\ &= \sup \quad (+ 1) \quad ( ) \\ &=: \sup \quad ( ) \end{aligned}$$

To obtain the desired limiting distribution of this supremum, we apply Corollary A1 of Bickel and Rosenblatt (1973). For this, we have to verify the assumptions of Theorem A1 of said authors.

Before we do so, we collect some helpful results. First, since ( ) has zero mean, it follows that

$$( ) := \mathrm{E} \quad ( ) \quad = 0 \tag{A.1}$$

Second, it holds for the covariance function that

$$\begin{aligned} ( ) &:= \mathrm{E} \quad ( ) \quad (+) \\ &= \mathrm{E} \qquad (+ 1) \qquad ( ) \qquad (+ \quad + 1) \qquad (+) \\ &= \quad 1 \end{aligned} \tag{A.2}$$

as is easy to verify by distinguishing the four cases 1, 1 0, 0 1 and 1.

We can now verify the assumptions of Theorem A1 of Bickel and Rosenblatt (1973).

(i) It holds that ( ) := ( ) $\overline{2 \log \quad}$ = 0, such that uniform boundedness is immediate.

(ii) It is also immediate that ( ) = 0 0 =: ( ) uniformly in , as .

(iii) We have that

$$: \ ( ) \qquad 0 \qquad = \qquad =: \ ( ) \tag{A.3}$$

where ( ) denotes the Lebesgue measure on .

(iv) ( ) = 0 from item (ii) is evidently uniformly continuous on .

(v) ( ) = 1 + , (0 2], as , is obviously satisfied by (A.2) (for = 1 and = 1).

(vi) ( ) d = 1 3 by (A.2).

Since all conditions of Theorem A1 are satisfied, we may apply Corollary A1 from Bickel and Rosenblatt (1973) to deduce that

$$\lim \ \mathrm{P} \quad \sup \quad ( ) \qquad + \qquad = \exp \quad (\quad + \quad) \tag{A.4}$$

where

$$= \sqrt{2\log} \qquad = \sqrt{2\log} \quad \frac{-\log\log \quad +\log \sqrt{} \quad 2}{\sqrt{2\log}}$$

This coincides with given in the statement of the lemma, because $= 1$ and $= 1$ (from item (v)), and $= = 1$ by the comments below Corollary A2 in Bickel and Rosenblatt (1973). Moreover, by (A.3),

$$= \quad \mathrm{d}\,(\ ) = 1 \quad \text{and} \quad = \quad \mathrm{d}\,(\ ) = 1$$

in the notation of Bickel and Rosenblatt (1973). Hence, the conclusion follows from (A.4). □

Proof of Theorem 2.1. It is easy to see that

$$\max \sqrt{} \qquad = \quad \sup \quad \sqrt{} \tag{A.5}$$

Due to Assumptions 2–4, we may apply Theorem 4 of Kuelbs and Philipp (1980) to conclude that, on a new probability space,

$$(\ ) \overset{\text{a.s.}}{=} \tag{A.6}$$

for some $(0\ 1\ 2)$, from Assumption 5, and a standard Brownian motion $(\ )$.

Hence, on this new probability space,

$$\sqrt{} \qquad = \sqrt{} \qquad \sqrt{}$$

$$= \frac{1}{\sqrt{}} \ \frac{1}{} \qquad (\quad +\ ) \quad \frac{1}{\sqrt{}} \ \frac{1}{} \qquad (\quad)$$

$$+ \frac{1}{\sqrt{}} \quad (\quad +\ ) \quad (\quad)$$

$$= \frac{1}{\sqrt{}} \ (\quad +\ ) + (\quad) \quad (1) + \frac{1}{\sqrt{}} \quad (\quad +\ ) \quad (\quad) \tag{A.7}$$

Note for (A.7) that, by (A.6),

$$\frac{1}{} \qquad (\quad) \quad \sup \quad \frac{1}{} \qquad (\ )$$

$$= \sup \frac{1}{} \quad ( )$$

$$\sup \frac{1}{} \quad ( )$$

$$= (\quad) \quad (1)$$

where the (1)-term is uniform in [0 1] by (A.6). Similarly, we get that

$$\frac{1}{} \quad (\quad + \quad) = (\quad + \quad) \quad (1)$$

The first term on the right-hand of (A.7) side vanishes asymptotically, because

$$\sup \frac{1}{=} (\quad + \quad) + (\quad) \quad = \tag{A.8}$$

by Assumption 1.

For the second term on the right-hand side of (A.7), observe that by the scaling property of Brownian motion,

$$\frac{1}{=} \quad (\quad + \quad) \quad (\quad) = \frac{}{} \quad (\quad + \quad) \quad ( )$$

and therefore, by Lemma 1,

$$\lim \mathrm{P} \quad \sup \frac{(\quad + \quad) \quad (\quad)}{} +$$

$$= \lim \mathrm{P} \quad \sup \frac{(\quad + \quad) \quad ( )}{} + \quad = \exp \quad 2 \tag{A.9}$$

(Note we slightly overload notation here by denoting the probability measure on the new probability space also by P .)

Together, (A.7)–(A.9) imply that

$$\lim \mathrm{P} \quad \sup \frac{}{=} \quad + \quad = \exp \quad 2 \tag{A.10}$$

Going back to the original probability space, we obtain that, as ,

$$\max \frac{}{=}$$

$$= \sup \frac{}{=}$$

$$= \frac{}{} \sup \frac{}{=} \frac{}{}$$

$$= 1 + (1) \sup \frac{}{=} \frac{}{}$$

$$= 1 + (1) \sup \frac{}{=} 1 + (1) \frac{}{}$$

Gumbel log(2) 1

where we used (A.5) in the first step, Assumption 5 in the third step, and (A.10) in the final step. This finishes the proof. □

B Proof of Theorem 2.2

The proof of Theorem 2.2 is similar to that of Theorem 2.1. The following is then a simple adaptation of Lemma 1.

Lemma 2. *Suppose that* ( ) *is a standard Brownian motion, and* 0, *as* . *Then, for all* ,

$$\lim \mathrm{P} \sup \frac{(+) ( )}{=} + \frac{}{} + = \exp +$$

where *and* *are as in* (2.8) *with* = 1 1.

Proof. The proof is very similar to that of Lemma 1. To highlight the similarities, we often overload notation by redefining quantities already introduced in the proof of Lemma 1.

By the scaling property of Brownian motion, it holds that

$$\sup \frac{(+) ( )}{=} + \frac{}{} = \sup (+ 1) ( ) + \frac{}{}$$

$$= \sup (+ 1) ( ) + \frac{}{}$$

$$=: \sup ( )$$

To obtain the desired limiting distribution of this supremum, we apply Corollary A1 of Bickel and Rosenblatt (1973). For this, we have to verify the assumptions of Theorem A1 of said authors.

Again, the following two results are helpful. First, it easily follows that

$$( ) := \mathrm{E} \quad ( ) \quad = — \tag{B.1}$$

Second, it holds for the covariance function that, as in (A.2),

$$\begin{aligned} ( ) &:= \mathrm{E} \quad ( ) \quad \mathrm{E}[\ ( )] \quad (\ + \ ) \quad [\ (\ + \ )] \\ &= \mathrm{E} \quad (\ + 1) \quad ( ) \quad (\ + \ + 1) \quad (\ + \ ) \\ &= \ 1 \end{aligned} \tag{B.2}$$

We can now check the assumptions of Theorem A1 in Bickel and Rosenblatt (1973).

(i) It holds that $( ) := \ ( ) \ \overline{2 \log \ } =$ by (B.1), such that uniform boundedness is immediate.

(ii) It is also immediate that $( ) = \ =: \ ( )$ uniformly in , as .

(iii) We have that

$$: \ ( ) \quad 0 \quad = \quad =: \ ( ) \tag{B.3}$$

where $( )$ denotes the Lebesgue measure on .

(iv) $( ) =$ from item (ii) is evidently uniformly continuous on .

(v) $( ) = 1 \quad + \quad , \quad (0 \ 2]$, as , is obviously satisfied by (B.2) (for $= 1$ and $= 1$).

(vi) $( ) \mathrm{d} \ = 1 \ 3$ by (B.2).

Hence, we may apply Corollary A1 from Bickel and Rosenblatt (1973) to deduce that

$$\lim \ \mathrm{P} \quad \sup \quad ( ) \quad + \quad = \exp \quad (\quad + \quad) \tag{B.4}$$

where

$$= \ \overline{2 \log \ } \quad = \ \overline{2 \log \ } \quad \frac{- \log \log \quad + \log \ — \quad 2}{\overline{2 \log \ }}$$

This coincides with given in the statement of the lemma, because $= 1$ and $= 1$ (from item (v)), and $= \ = 1$ by the comments below Corollary A2 in Bickel and

Rosenblatt (1973). Moreover, by (B.3),

$$= \quad \mathrm{d}\ (\ ) = \quad \text{and} \quad = \quad \mathrm{d}\ (\ ) =$$

in the notation of Bickel and Rosenblatt (1973). Hence, the conclusion follows from (B.4). □

Proof of Theorem 2.2. Similarly as in the proof of Theorem 2.1, we get from Assumptions 2–4 that we may apply Theorem 4 of Kuelbs and Philipp (1980). In this case, on a new probability space,

$$(\quad) \quad (\ ) \overset{\text{a.s.}}{=} \tag{B.5}$$

for some $(0\ 1\ 2)$, and a standard Brownian motion $(\ )$. Hence, similarly as before,

$$\begin{aligned}
\frac{}{=} \quad &= \frac{}{=} \qquad \frac{}{=} \qquad + \frac{}{=} \\
&= \frac{1}{\frac{}{=}}\ \frac{1}{-} \qquad (\quad + \ ) \\
&\qquad \frac{1}{\frac{}{=}}\ \frac{1}{-} \qquad (\quad) \\
&\qquad + \frac{1}{\frac{}{=}} \quad (\quad + \ ) \quad (\quad) + \frac{}{=} \\
&= \frac{1}{\frac{}{=}} (\quad + \ ) \quad (\quad) \quad (1) \\
&\qquad + \frac{1}{\frac{}{=}} \quad (\quad + \ ) \quad (\quad) + \quad {}^{-} + (1)
\end{aligned} \tag{B.6}$$

where the (1)-term in the final line is uniform on $[0\ 1\ \ ]$.

The first term on the right-hand side of (B.6) vanishes asymptotically uniformly on $[0\ 1\ \ ]$ by (A.8).

For the second and third term on the right-hand side of (B.6), observe that by the scaling property of Brownian motion,

$$\frac{1}{\frac{}{=}} \quad (\quad + \ ) \quad (\quad) + \quad {}^{-} = \frac{}{-} \quad (\ + \quad) \quad (\ ) + \frac{}{}$$

and therefore, by Lemma 2,

$$\lim \mathrm{P} \quad \sup \quad \frac{(\quad + \quad) \qquad (\quad)}{-} + \qquad - \qquad\qquad + $$

$$= \lim \mathrm{P} \quad \sup \quad \frac{(\; + \qquad) \qquad (\;)}{\overline{\qquad}} + \frac{}{} \qquad\qquad + $$

$$= \exp \qquad + \tag{B.7}$$

Together, (B.6)–(B.7) imply that

$$\lim \mathrm{P} \quad \sup \quad \frac{}{=} \qquad\qquad\qquad + \qquad = \exp \qquad\qquad + $$

Combining this with Assumption 5 and then with (A.5) finishes the proof, similarly as in the proof of Theorem 2.1. □

C Proof of Proposition 1

Before providing the proof of Proposition 1, we state the main assumptions used to prove it. For this, let denote the Euclidean norm. In Assumptions W1–W4 we suppress the dependence of models, estimators and parameters on the index that denotes the forecaster in the main paper (1 2).

Assumption W1. *In some open neighborhood around , and with probability one:*

(a) ( ) *is measurable and twice continuously differentiable with respect to* .

(b) There is a constant such that for all , sup ( ) *for a measurable for which* E[ ] *for some* 1.

Assumption W2. *The estimate satisfies* = ( ) ( ), *where* ( ) *is* ( ) *and* ( ) *is* (1), *with*

(a) ( ) $\overset{a.s.}{}$, *a matrix of rank* ;

(b) ( ) = ( ) *for a* (1) *orthogonality condition* ( );

(c) E ( ) = 0.

Assumption W3. *Let*

$$= \; (\;) \qquad = \; (\qquad) \qquad = \frac{}{}(\;) \qquad = \mathrm{E}[\quad]$$

Then:

(a) sup E ( ) *for* 1 *from Assumption W1 (b).*

(b) ( ) E[ ] *is strong mixing, with mixing coefficients of size* 3 (1).

(c) ( ) E[ ] *is covariance stationary.*

(d) Let = E E[ ] E[ ] , = ( ). *Then* *is positive definite.*

Assumption W4. , *as* , *and* lim = , 0 ; = lim = 0.

Proof of Proposition 1. By Assumption W1(a), we may invoke a second-order mean value expansion around the true value , yielding that

$$(\quad) \quad (\quad) = — (\quad) + \frac{1}{2} \text{——} (\quad) \tag{C.1}$$

where is a mean value on the line connecting and .

By Assumption W3(a), E , such that

$$\max \quad = \tag{C.2}$$

by a union bound and Markov's inequality. Using similar arguments, we obtain from Assumption W1(b) that

$$\max \quad = \tag{C.3}$$

Combining (C.2)–(C.3) with (C.1), we deduce that

$$\begin{aligned} \max \quad (\quad) \quad (\quad) \quad \max \quad \max \quad + \frac{1}{2} \max \quad \max \\ = \quad + \\ = \end{aligned}$$

where the penultimate step additionally uses that max = for any 0 (West, 1996, Lemma A3(b)), and the final line follows for sufficiently small (by Assumption W4).

Therefore, for = ( ), Assumption 6 is always satisfied, because

$$\text{——} = \quad = \quad \frac{1}{\text{——}} =$$

as desired. Note that the first equality here follows from Assumption W4, and the second one from 1 1 (2), as specified in the proposition. □

D Proof of Theorem 2.3

Proof of Theorem 2.3. Since Theorem 2.2 implies that

$$\max \quad ^{-} \quad = \quad (\quad)$$

where the order of magnitude is exact, and because of the exact form of the limiting null distribution in Theorem 2.1, we need to show that

$$\max \quad ^{-}(\quad) =$$

Therefore, we show in the following that, for $= 1 \ 2$,

$$\max \quad \mathcal{L}(\quad) \quad \mathcal{L}(\quad) = \quad ^{-}$$

Consider the case $= 1$ first, where the loss function is Lipschitz-continuous. Then, we have $= 1$ that

$$\mathcal{L}(\quad) \quad \mathcal{L}(\quad) = \mathcal{L} \quad (\quad) \quad (\quad) \quad \mathcal{L}(\quad) \quad (\quad) \quad (\quad)$$

implying that

$$\max \quad \mathcal{L}(\quad) \quad \mathcal{L}(\quad) \quad \max \quad (\quad) \quad (\quad)$$

From this the desired result follows with Assumption 6.

This reasoning can be modified for 2 as follows. Given the continuity of the 1st order derivative of $\mathcal{L}$, we may apply a Taylor series expansion with the rest term in differential form. This gives that

$$\mathcal{L}(\quad) = \mathcal{L}(\quad) + \quad \frac{1}{!}\mathcal{L} \quad (\quad)(\quad) + \frac{1}{(\quad 1)!}\mathcal{L} \quad (\quad)(\quad)$$

where lies between and (and we make the usual convention that $= 0$ for 2). Thus,

$$\mathcal{L}(\quad) \quad \mathcal{L}(\quad) = \quad \frac{1}{!}\mathcal{L} \quad (\quad) \quad (\quad) \quad (\quad)$$
$$+\frac{1}{(\quad 1)!} \ \mathcal{L} \quad (\quad) \quad \mathcal{L} \quad (\quad) \quad (\quad) \quad (\quad)$$

Since $\mathcal{L}$ is Lipschitz-continuous, we have

$$\mathcal{L}\quad(\quad)\quad\mathcal{L}\quad(\quad)\qquad\qquad=\quad(\quad)\quad(\quad)$$

such that

$$\mathcal{L}(\quad)\quad\mathcal{L}(\quad)=\quad\frac{1}{!}\mathcal{L}\quad(\quad)\quad(\quad)\quad(\quad)\quad+$$

where $(\quad)\quad(\quad)$. Therefore,

$$\mathcal{L}(\quad)\qquad\mathcal{L}(\quad)\qquad\frac{1}{!}\qquad\mathcal{L}\quad(\quad)\quad(\quad)\quad(\quad)\quad+\quad\max$$

Note that for any ,

$$\mathcal{L}\quad(\quad)\quad(\quad)\quad(\quad)\qquad\max\quad(\quad)\quad(\quad)\quad\max\qquad\mathcal{L}\quad(\quad)$$

where $\mathcal{L}\quad(\quad)$ thanks to Assumption 0. Then,

$$\max\qquad\mathcal{L}\quad(\quad)\qquad\max\qquad\mathrm{E}\qquad+\max\qquad\qquad\mathrm{E}$$

The first term on the right-hand side is $(\quad)$, because we impose existing moments in Assumption 2. The second term is $(\quad^{-}\quad)=\quad(\quad)$ by an application of Theorem 4 from Kuelbs and Philipp (1980) as in Theorem 2.1. Therefore, for $=1\ 2\qquad 1$,

$$\max\qquad\mathcal{L}\quad(\quad)\quad(\quad)\quad(\quad)\quad=\qquad\max\quad(\quad)\quad(\quad)$$

and, summing up,

$$\max\qquad\mathcal{L}(\quad)\qquad\mathcal{L}(\quad)=\qquad\max\quad(\quad)\quad(\quad)$$

Hence, the desired result follows from Assumption 6. □

E Proof of Theorem 2.4

Proof of Theorem 2.4. Examine first

$$\begin{aligned} &= \mathcal{L}(\quad) \quad \mathcal{L}(\quad) \\ &= \quad (\quad)(\mathcal{L}(\quad) \quad \mathcal{L}(\quad)) \\ &:= \quad (\quad) \end{aligned}$$

where we exploited the homogeneity of the loss function. Since ( ) is bounded away from zero and Lipschitz and is a positive integer, ( ) = ( ) is itself bounded away from zero (and bounded for that matter) and piecewise Lipschitz; furthermore is weakly stationary for any given the -boundedness of and .

Then, following the steps of the proof of Theorem 2.1, and noting that ( ) is bounded and bounded away from zero,

$$\max \quad \frac{1}{=} \quad \frac{}{(\quad)} = \max \quad \frac{1}{=}$$

is easily shown to have the same limiting behavior as stated in Theorem 2.1. This also implies

$$\max \quad \frac{1}{=} \quad \frac{}{(\quad)} = \quad (\quad)$$

where the magnitude order is exact. Like in the proof of Theorem 2.3, we therefore only need to show that

$$\max \quad \frac{}{=} \qquad \max \quad \frac{1}{=} \quad \frac{}{(\quad)} = \quad (\quad)$$

To establish this, note that, following the steps of the proof of Theorem 2.3, we may establish that

$$\max \quad \frac{1}{=} \quad \frac{}{(\quad)} = \max \quad \frac{1}{=} \quad \frac{}{(\quad)} + \quad (\quad) \tag{E.1}$$

as follows. The Taylor expansion is essentially the same, such that

$$\mathcal{L}(\quad) \quad \mathcal{L}(\quad) = \quad \frac{1}{!}\mathcal{L} \quad (\quad) \quad (\quad) \quad (\quad) \quad +$$

where ( ) ( ) , and the asymptotic equivalence in (E.1) follows if

$$\max \quad \mathcal{L} \quad (\quad) = \quad ( )$$

But ( ) is uniformly bounded such that

$$\max \quad \mathcal{L} \quad (\quad) \quad \max \quad \mathrm{E}$$

and the same argument used for strictly stationary in the proof of Theorem 2.3 applies here for as required.

Write now

$$\frac{}{=} \quad = \frac{(\quad)}{} \frac{1}{=} \quad \frac{(\quad)}{(\quad)} \frac{}{(\quad)}$$

Given that ( ) is uniformly bounded away from 0, and is uniformly consistent at a rate higher than by Assumption 5*, it holds that

$$\frac{(\quad)}{} = 1 +$$

uniformly in , and, due to the Lipschitz continuity of ( ) (and of ( ) being uniformly bounded away from 0), it also holds that

$$\frac{(\quad)}{(\quad)} = 1 + \quad \frac{}{} = 1 + \quad -$$

again uniformly in . Examine now

$$\frac{1}{=} \quad \frac{(\quad)}{(\quad)} \frac{}{(\quad)} = \frac{1}{=} \quad \frac{}{(\quad)} + \frac{1}{=} \quad \frac{(\quad)}{(\quad)} \quad 1 \quad \frac{}{(\quad)}$$

for which we have that

$$\frac{1}{=} \quad \frac{(\quad)}{(\quad)} \quad 1 \quad \frac{}{(\quad)} \quad \frac{1}{=} \quad \frac{}{(\quad)} \quad \max \quad \frac{(\quad)}{(\quad)} \quad 1$$

where it is tedious yet straightforward to show that $\frac{}{} = ( )$ uniformly in . Thus,

$$\max \quad \frac{1}{=} \quad \frac{(\quad)}{(\quad)} \frac{}{(\quad)} = \quad \max \quad \frac{1}{=} \quad \frac{}{(\quad)} + \quad \frac{}{}$$

Summing up,

$$\max \quad \frac{}{=} \quad = \quad \max \quad \frac{1}{=} \quad \frac{}{(\quad)} + \quad + \quad \frac{}{}$$

$$= \quad \max \quad \frac{1}{=} \quad \frac{}{(\quad)} +$$

as required. □

F Additional Simulation Results

F.1 Additional simulations to subsections in Section 3

This section provides some additional simulation results to some of the subsections discussed in Section 3.

		$= 0\,6$					$= 0\,7$				
		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.062	0.332	0.191	0.161	0.096	0.092	0.494	0.240	0.227	0.145
100	50	0.070	0.405	0.202	0.186	0.135	0.118	0.681	0.304	0.319	0.245
200	50	0.072	0.482	0.216	0.200	0.153	0.131	0.760	0.326	0.351	0.308
300	50	0.065	0.494	0.210	0.200	0.179	0.118	0.801	0.332	0.378	0.334
50	100	0.198	0.631	0.331	0.340	0.265	0.272	0.867	0.456	0.534	0.447
100	100	0.202	0.742	0.357	0.395	0.357	0.297	0.953	0.483	0.635	0.617
200	100	0.203	0.808	0.377	0.426	0.427	0.306	0.972	0.521	0.692	0.720
300	100	0.203	0.845	0.375	0.469	0.464	0.310	0.982	0.518	0.717	0.754
50	200	0.338	0.938	0.492	0.708	0.632	0.453	0.995	0.632	0.920	0.889
100	200	0.353	0.971	0.509	0.788	0.722	0.486	1.000	0.670	0.961	0.952
200	200	0.348	0.987	0.515	0.811	0.775	0.509	1.000	0.684	0.966	0.974
300	200	0.356	0.990	0.520	0.826	0.795	0.487	1.000	0.667	0.973	0.988
50	300	0.459	0.996	0.605	0.923	0.872	0.612	1.000	0.758	0.994	0.988
100	300	0.456	0.999	0.610	0.950	0.917	0.626	1.000	0.782	0.999	0.997
200	300	0.450	0.999	0.611	0.973	0.943	0.600	1.000	0.763	0.997	0.998
300	300	0.453	1.000	0.617	0.964	0.943	0.596	1.000	0.768	0.999	0.999
50	400	0.537	1.000	0.690	0.981	0.954	0.703	1.000	0.840	1.000	1.000
100	400	0.531	1.000	0.703	0.988	0.976	0.715	1.000	0.859	1.000	1.000
200	400	0.533	1.000	0.705	0.992	0.985	0.719	1.000	0.866	1.000	1.000
300	400	0.555	1.000	0.724	0.996	0.987	0.727	1.000	0.870	1.000	1.000

Table F.1: Power (global superiority) for different across combinations of and . We consider the $\mathcal{M}$ test, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $=$ $_{\text{rot}}$ as well as at $= 0\,4$ and of the Lan et al. (2024) test, at nominal level 5% for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Sections 3.1 and 3.5.

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.029	0.066	0.084	0.035	0.023
100	50	0.026	0.069	0.099	0.043	0.021
200	50	0.024	0.062	0.081	0.039	0.021
300	50	0.029	0.063	0.084	0.040	0.020
500	50	0.030	0.068	0.094	0.040	0.020
50	100	0.035	0.058	0.080	0.040	0.028
100	100	0.032	0.066	0.087	0.047	0.024
200	100	0.034	0.057	0.086	0.045	0.020
300	100	0.037	0.057	0.071	0.037	0.021
500	100	0.037	0.052	0.097	0.034	0.020
50	200	0.029	0.052	0.079	0.038	0.026
100	200	0.038	0.063	0.085	0.042	0.029
200	200	0.032	0.057	0.076	0.039	0.020
300	200	0.035	0.054	0.077	0.046	0.024
500	200	0.043	0.055	0.083	0.040	0.020
50	300	0.038	0.051	0.076	0.045	0.030
100	300	0.044	0.049	0.086	0.032	0.025
200	300	0.038	0.049	0.076	0.042	0.026
300	300	0.034	0.054	0.075	0.042	0.026
500	300	0.036	0.055	0.073	0.041	0.031

Table F.2: Empirical size of the rolling Gumbel test $\mathcal{M}$, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $=$ $_{\text{rot}}$ as well as at $= 0\ 4$ and of the Lan et al. (2024) test, at nominal level 5% for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Section 3.2. Here, estimated IV forecast functions are replaced with the true predictors and .

F.2 McCracken (2020): nested estimation

McCracken (2020) exhibits a simple example of a nested DGP. These are, among other things, of interest as, after having accounted for estimation effects, the larger model cannot perform worse than the nested model, inducing potential degeneracies in common test statistics. As we do not know the forecasting approaches used typical EPA comparisons, we also cannot rule out that the comparisons involve nested approaches, such that it is worth investigating how the tests considered here perform under such DGPs.

		$\mathcal{M}$	$\mathcal{DM}$	$\mathcal{M}_{\text{rot}}$	$\mathcal{M}$	
50	50	0.002	0.095	0.016	0.054	0.081
100	50	0.003	0.086	0.011	0.049	0.085
200	50	0.002	0.076	0.016	0.043	0.075
300	50	0.002	0.072	0.010	0.040	0.068
500	50	0.002	0.062	0.013	0.034	0.066
50	100	0.005	0.114	0.016	0.068	0.103
100	100	0.003	0.084	0.017	0.054	0.085
200	100	0.006	0.090	0.016	0.054	0.080
300	100	0.004	0.058	0.013	0.046	0.073
500	100	0.002	0.063	0.016	0.042	0.077
50	200	0.008	0.176	0.030	0.112	0.120
100	200	0.009	0.114	0.022	0.072	0.089
200	200	0.008	0.087	0.023	0.052	0.076
300	200	0.006	0.071	0.020	0.057	0.070
500	200	0.008	0.064	0.024	0.046	0.068
50	300	0.013	0.224	0.033	0.140	0.154
100	300	0.011	0.136	0.030	0.092	0.093
200	300	0.009	0.101	0.026	0.066	0.092
300	300	0.008	0.076	0.031	0.058	0.068
500	300	0.010	0.078	0.038	0.064	0.065
50	500	0.024	0.300	0.060	0.202	0.174
100	500	0.018	0.204	0.053	0.129	0.129
200	500	0.013	0.132	0.046	0.090	0.087
300	500	0.014	0.089	0.044	0.072	0.073
500	500	0.014	0.072	0.035	0.056	0.070

Table F.3: Empirical size of the rolling Gumbel test $\mathcal{M}$, the Diebold and Mariano (1995) test, the Giacomini and Rossi (2010) $\mathcal{M}$ tests at $=$ $_{\text{rot}}$ as well as at $= 0\,4$ and of the Lan et al. (2024) test, at nominal level 5% for various combinations of the forecast estimation window size and evaluation sample size . The DGP is explained in Section F.2.

In his DGP, he shows that the Diebold and Mariano (1995) $\mathcal{DM}$ statistic (he considers a sample variance rather than long-run variance estimator) diverges under the null, i.e., its

asymptotic size will be one. We here assess how our proposal behaves under such a nested DGP. These simulations are of an exploratory nature, as the above theoretical framework does not cover such nested models. It is nevertheless of interest to investigate how the tests studied here behave under such a scenario in which $\mathcal{DM}$ is known to fail.

Concretely, he compares the accuracy of a no-change point forecast for with that of a location model estimated using a fixed window . That is, = 0 while = , where = = for all horizons = 0 1.

Under quadratic loss, the differential then is

$$= (\quad 0) \quad (\quad)$$

He shows that, for i.i.d. N( ), the null of EPA E[ ] = 0 holds for all .

We here consider a grid = = 50 100 200 300 500 to illustrate the diverging nature of the $\mathcal{DM}$ statistic. Table F.3 reports results, confirming the finding of McCracken (2020) that the $\mathcal{DM}$ statistic suffers from increasing rejection rates under the null when increases. The large- Giacomini and Rossi (2010) $\mathcal{M}$ (which, again, is more akin to the standard $\mathcal{DM}$ statistic) and Lan et al. (2024) statistics similarly suffer from overrejections, in particular when is large relative to . In contrast, $\mathcal{M}$ controls size throughout, albeit at the cost of underrejections in this nested DGP.

G Additional Empirical Results

Table G.1 provides an overview of missing values. We report the proportion of missing values relative to the time series length starting with the first available observation per variable.

variable	method	1	2	3	4	5
CPI	DAR	0.000	0.000	0.000	0.000	0.000
	DARM	–	–	–	–	–
	IAR	0.000	0.000	0.000	0.000	0.000
	NC	0.000	0.000	0.000	0.000	0.000
	SPF	0.000	0.000	0.000	0.000	0.000
	realization	0.000	0.000	0.000	0.000	0.000
Housing	DAR	0.018	0.018	0.018	0.018	0.018
	DARM	0.032	0.032	0.032	0.032	0.032
	IAR	0.018	0.018	0.018	0.018	0.018
	NC	0.018	0.018	0.018	0.018	0.018
	SPF	0.000	0.000	0.000	0.000	0.022
	realization	0.000	0.000	0.000	0.000	0.000
Real GDP	DAR	0.004	0.004	0.004	0.004	0.004
	DARM	–	–	–	–	–
	IAR	0.004	0.004	0.004	0.004	0.004
	NC	0.004	0.004	0.004	0.004	0.004
	SPF	0.000	0.000	0.000	0.000	0.022
	realization	0.004	0.000	0.000	0.000	0.000
Unemployment	DAR	0.000	0.000	0.000	0.000	0.000
	DARM	0.000	0.000	0.000	0.000	0.000
	IAR	0.000	0.000	0.000	0.000	0.000
	NC	0.000	0.000	0.000	0.000	0.000
	SPF	0.000	0.000	0.000	0.000	0.022
	realization	0.000	0.000	0.000	0.000	0.000

Table G.1: Fraction of missing values by method and vintage and horizon. For forecasting methods, the columns cover the forecast horizons. For realizations, they describe the data vintage.

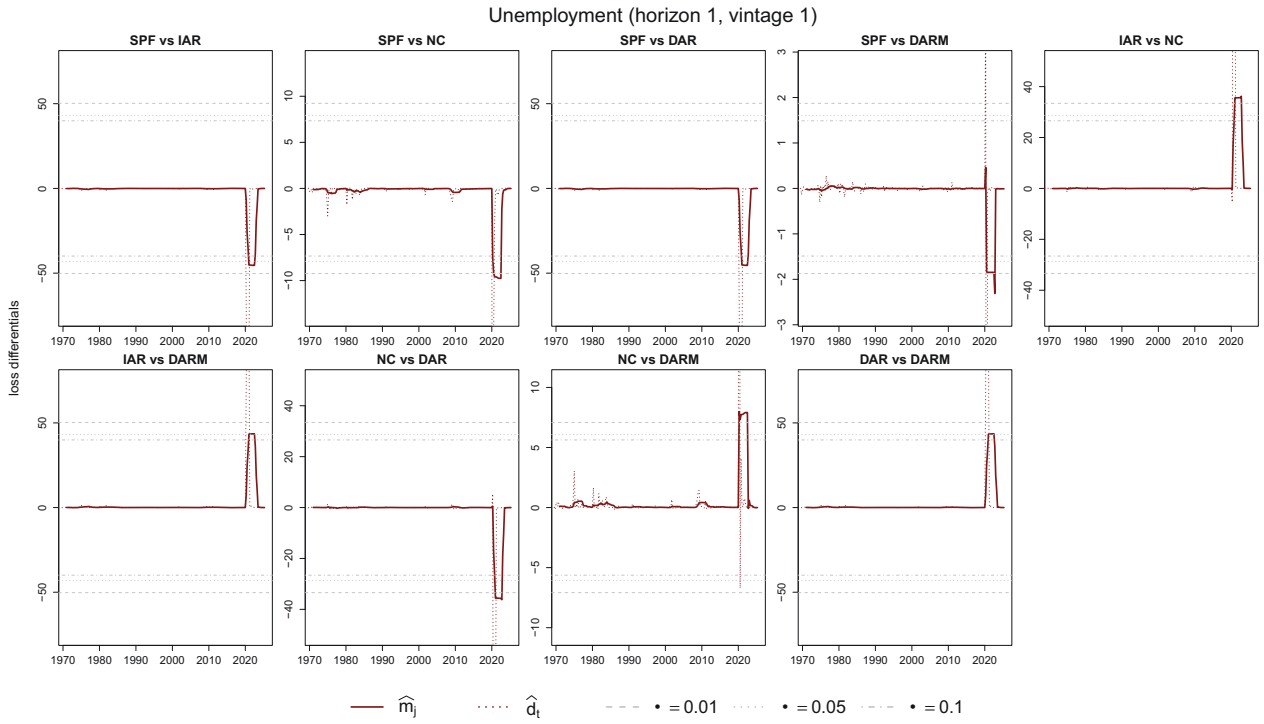

Figure G.1: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

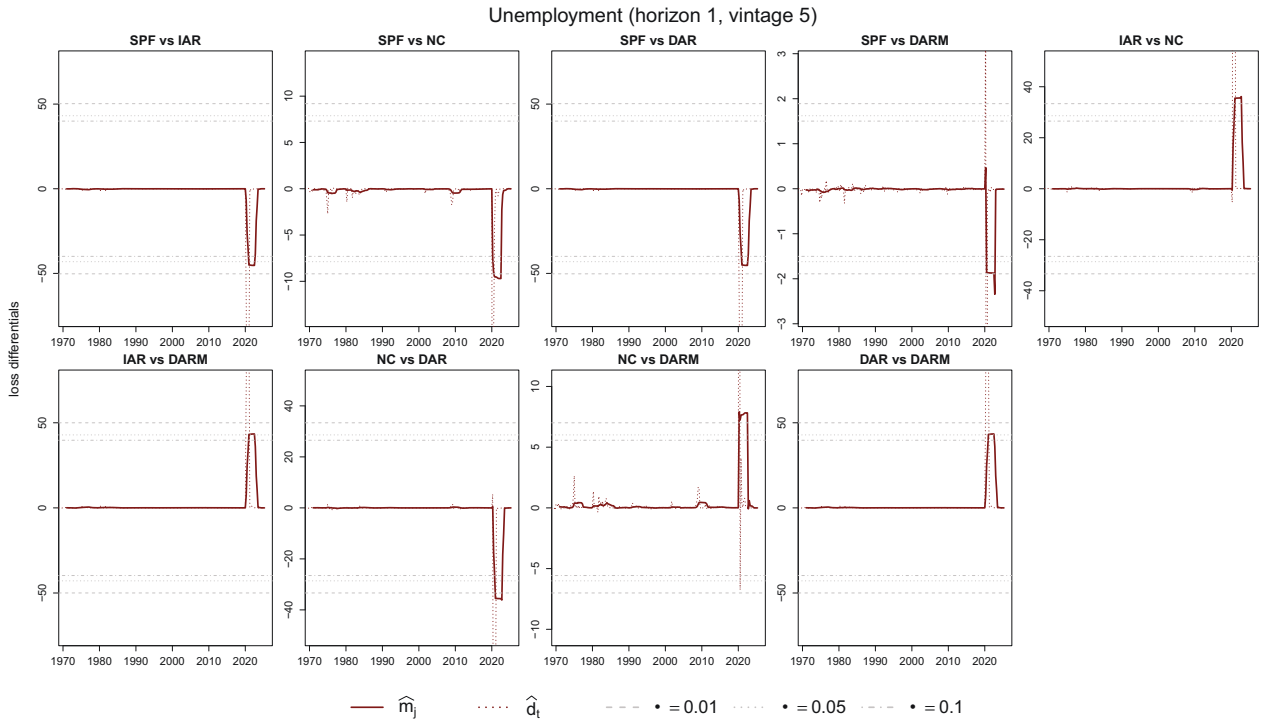

Figure G.2: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

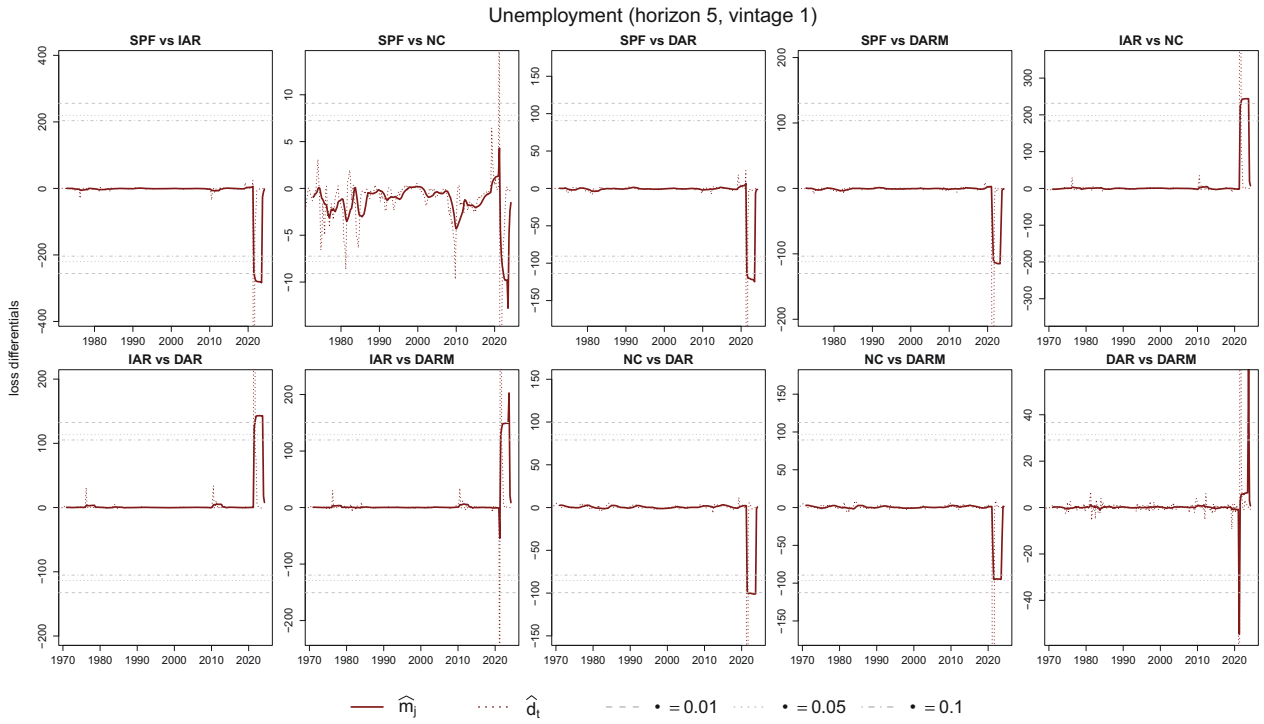

Figure G.3: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

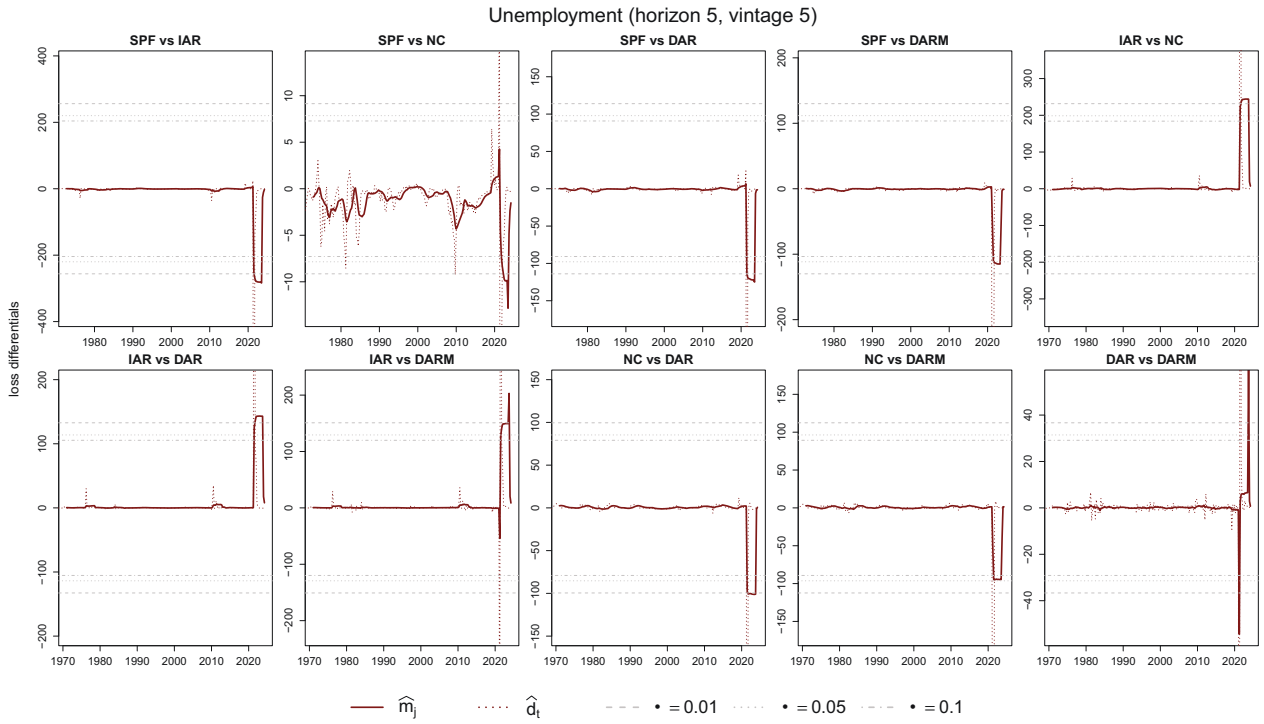

Figure G.4: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

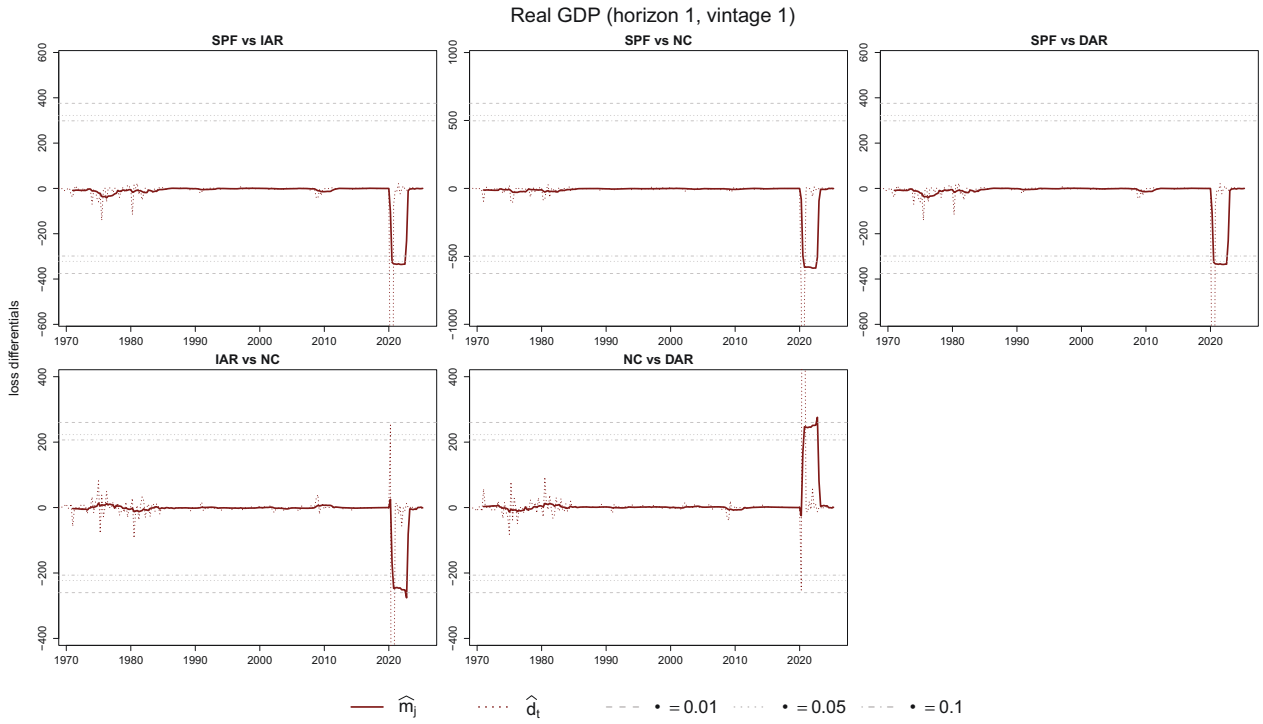

Figure G.5: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

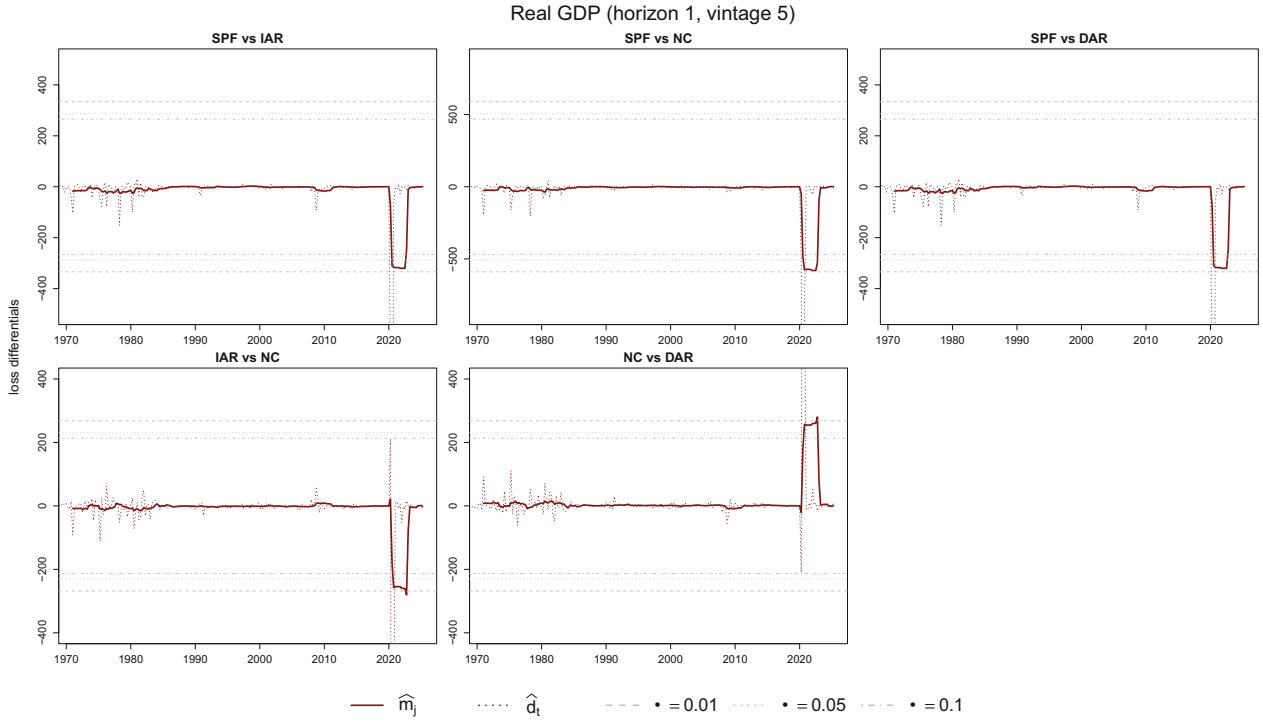

Figure G.6: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

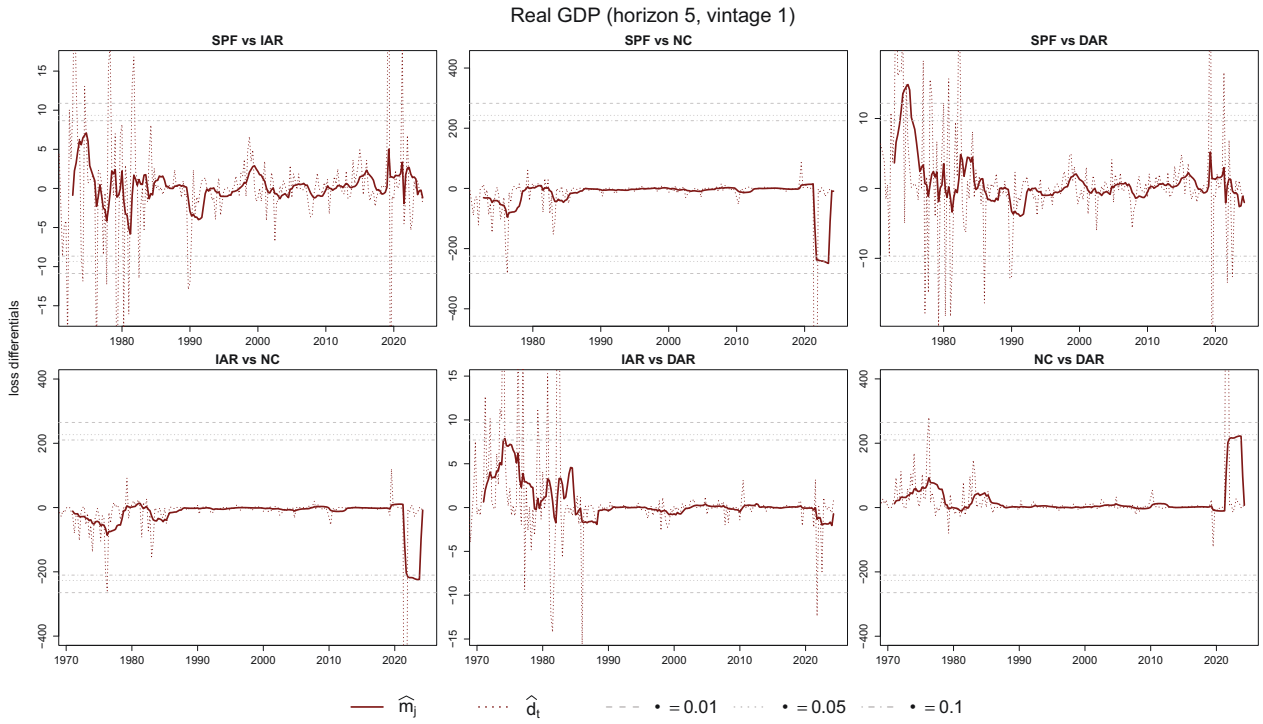

Figure G.7: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

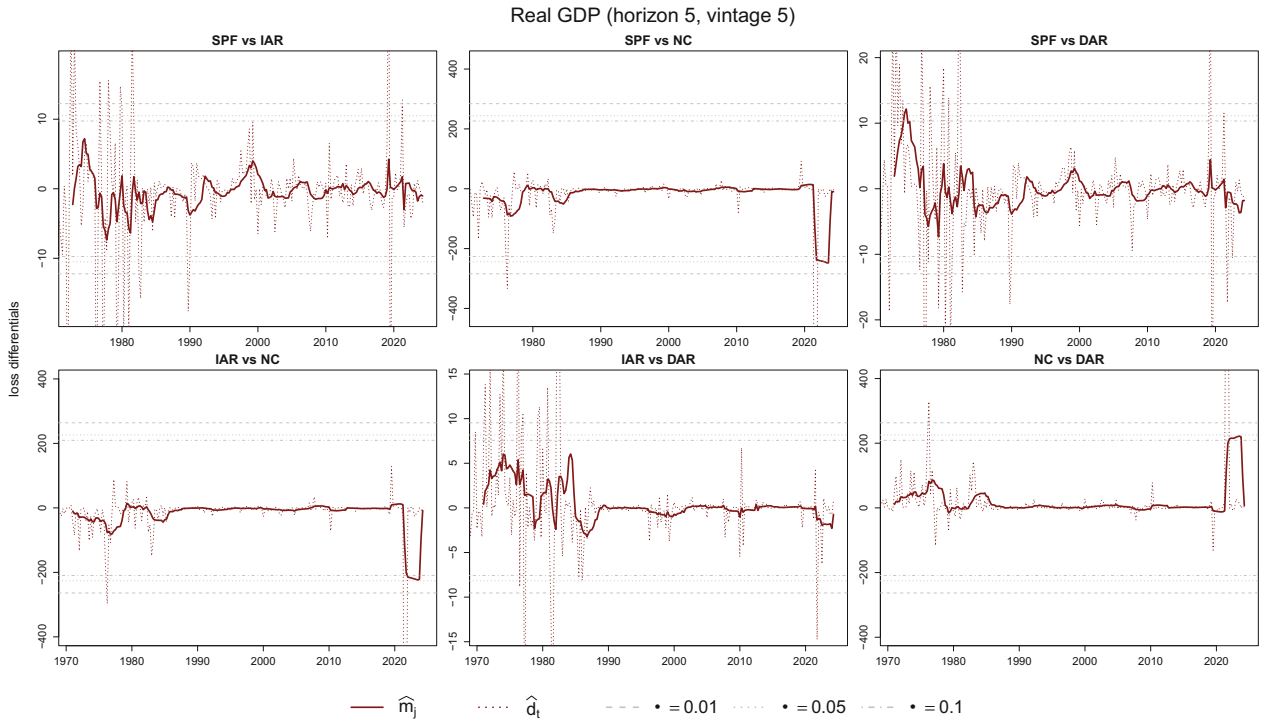

Figure G.8: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

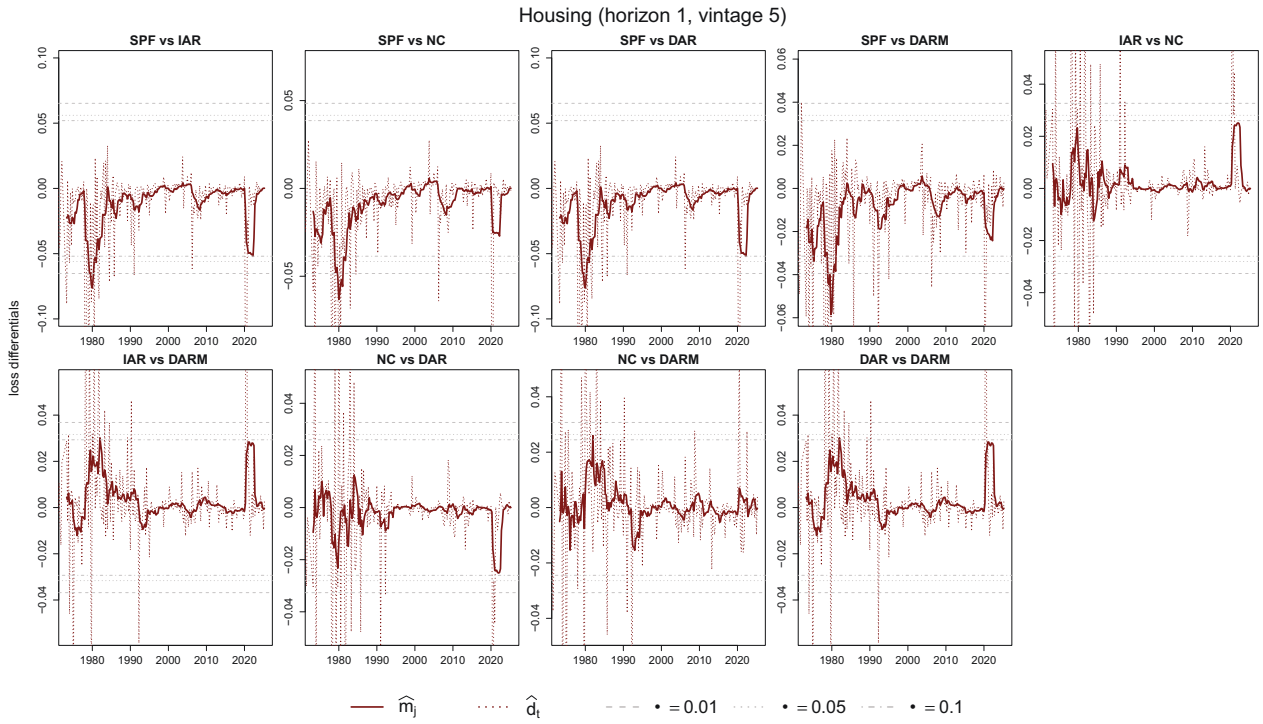

Figure G.9: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

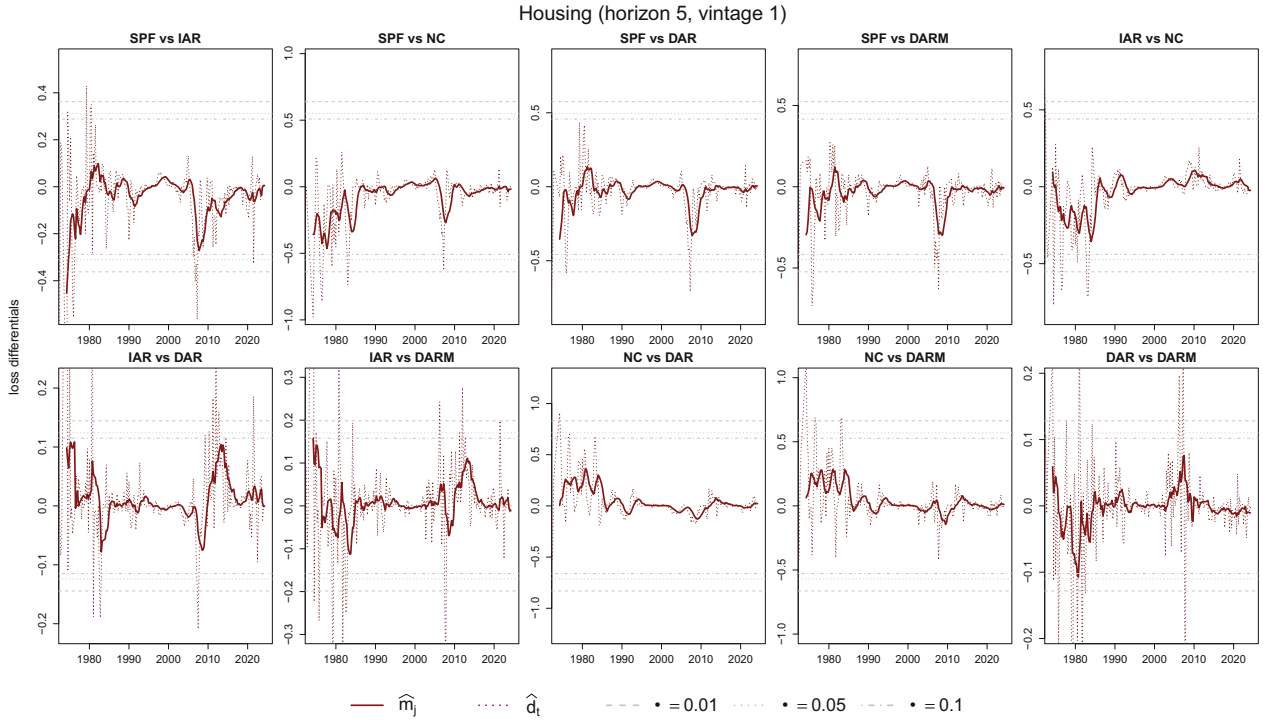

Figure G.10: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

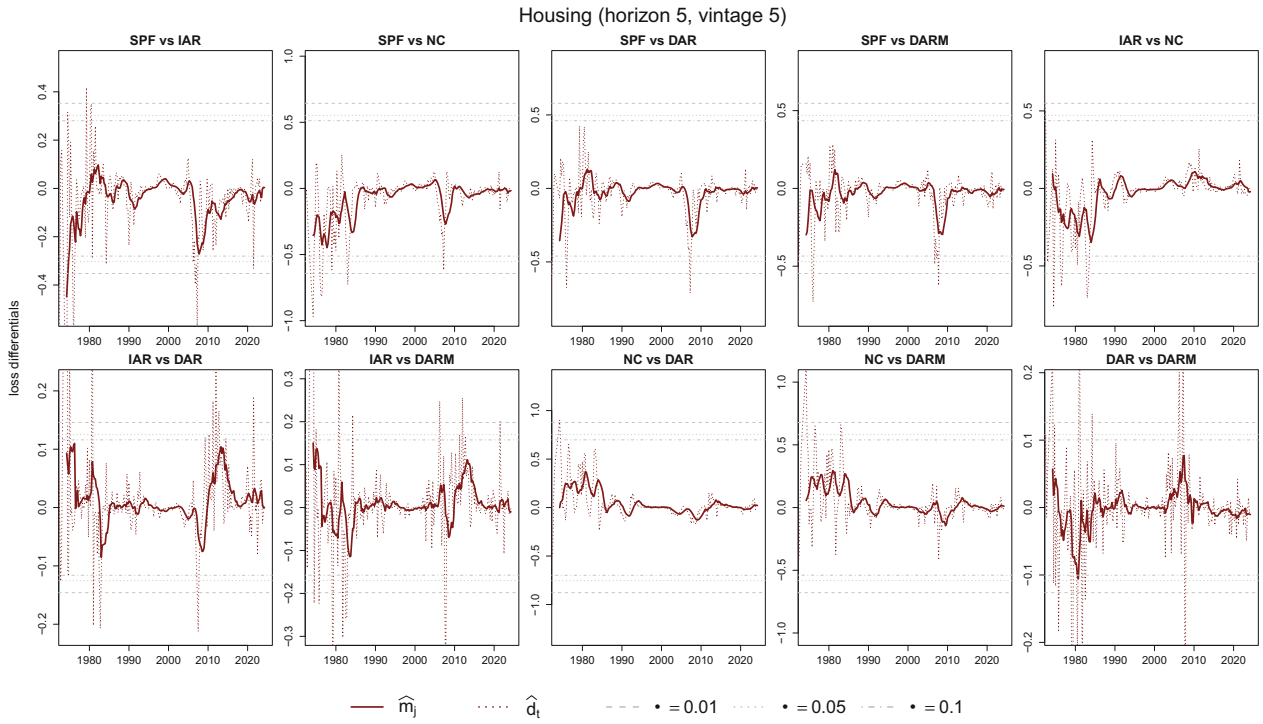

Figure G.11: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

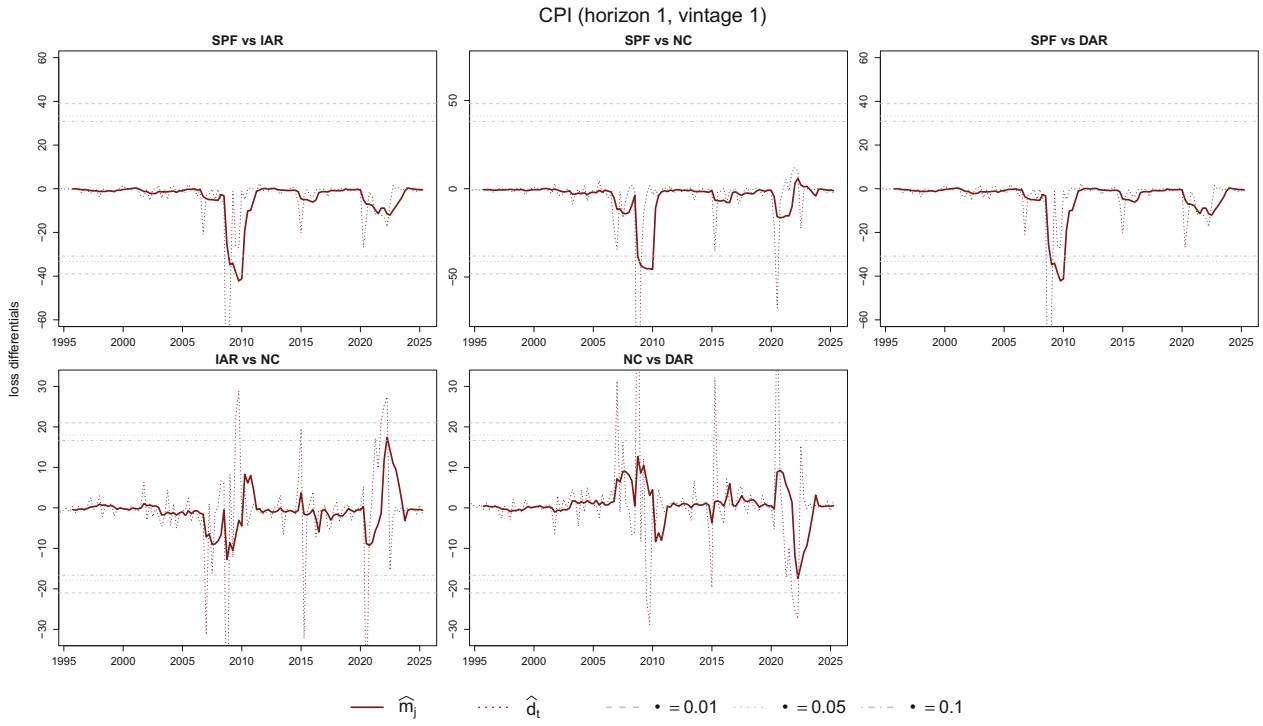

Figure G.12: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

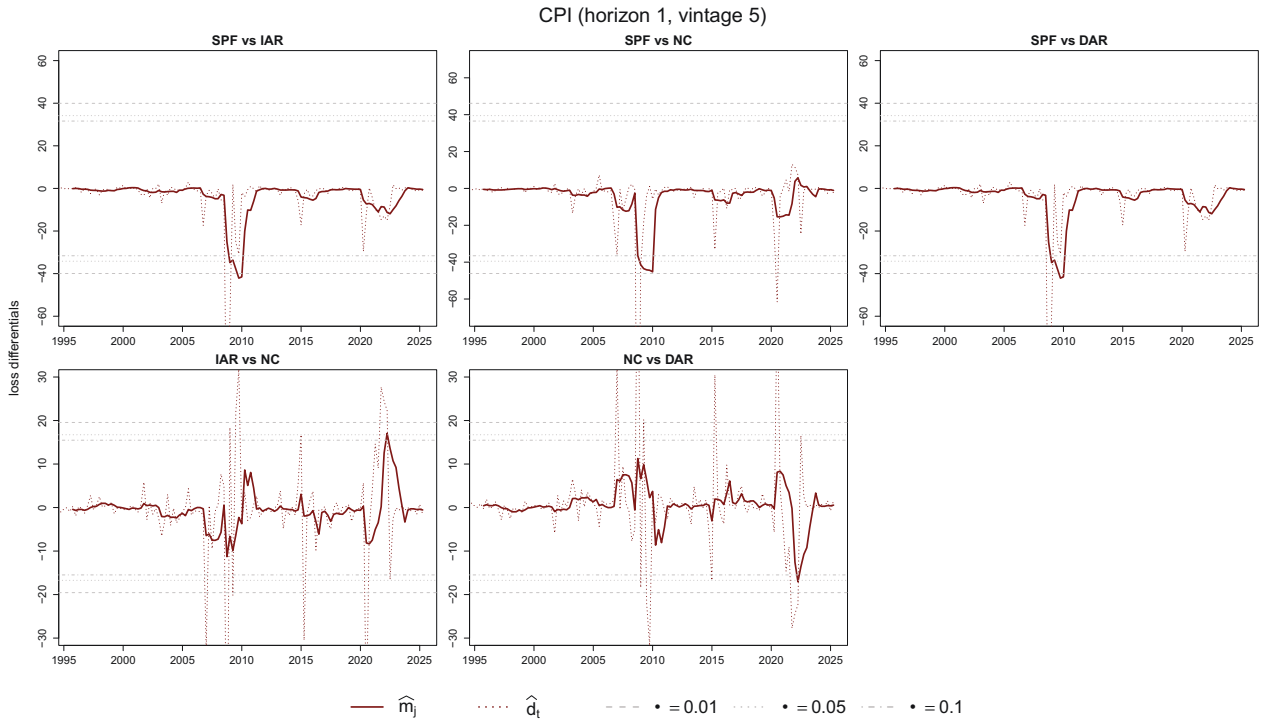

Figure G.13: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

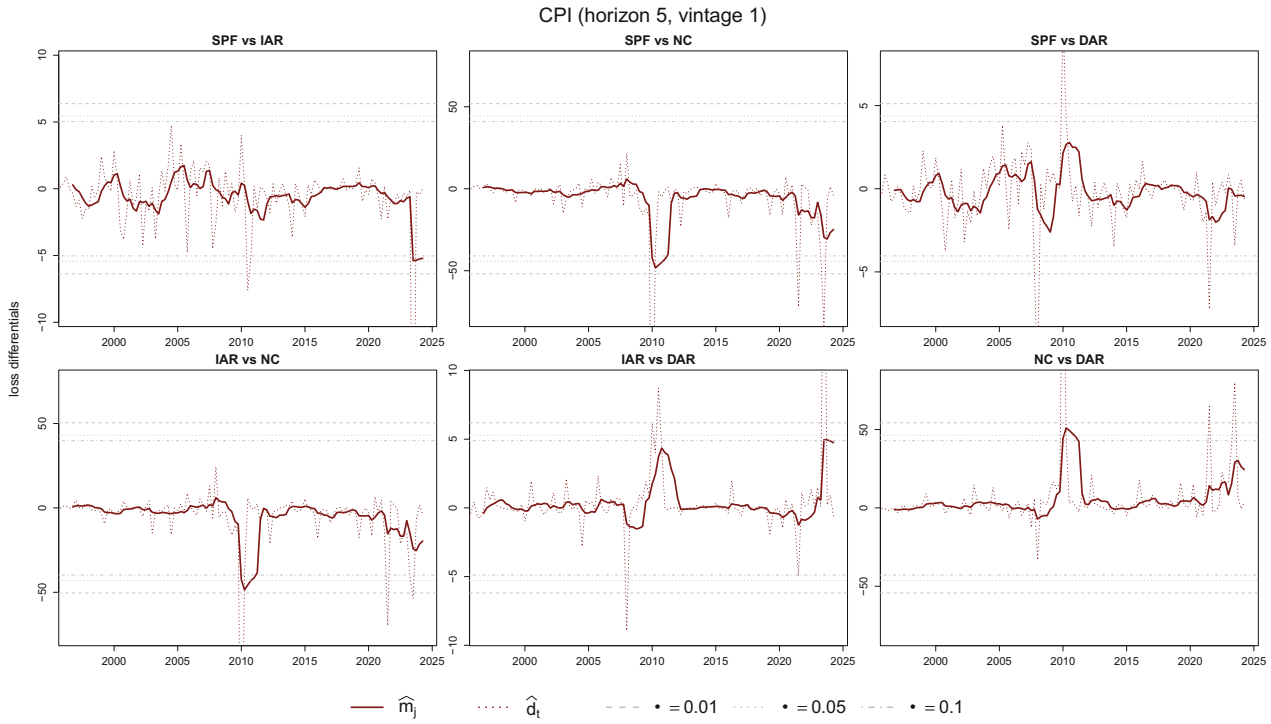

Figure G.14: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

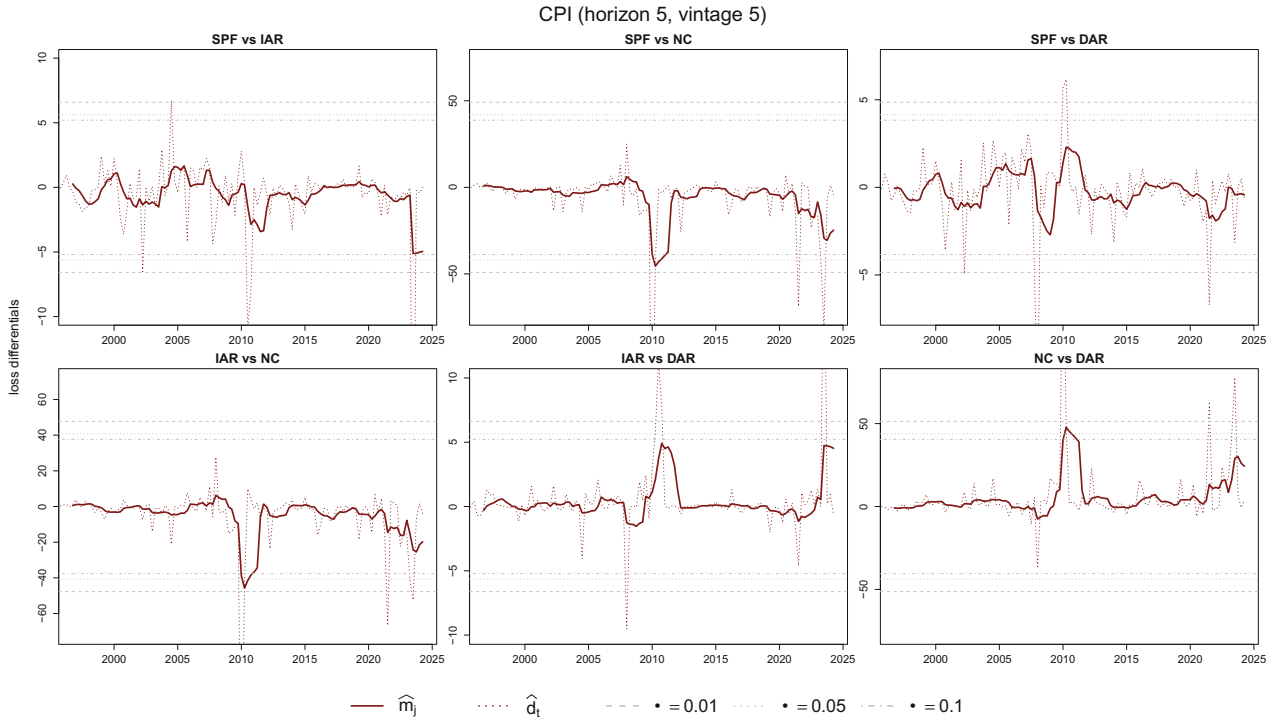

Figure G.15: $\widehat{m}_j$, $\widehat{d}_t$ and significance bounds (cf. Remark 1b)

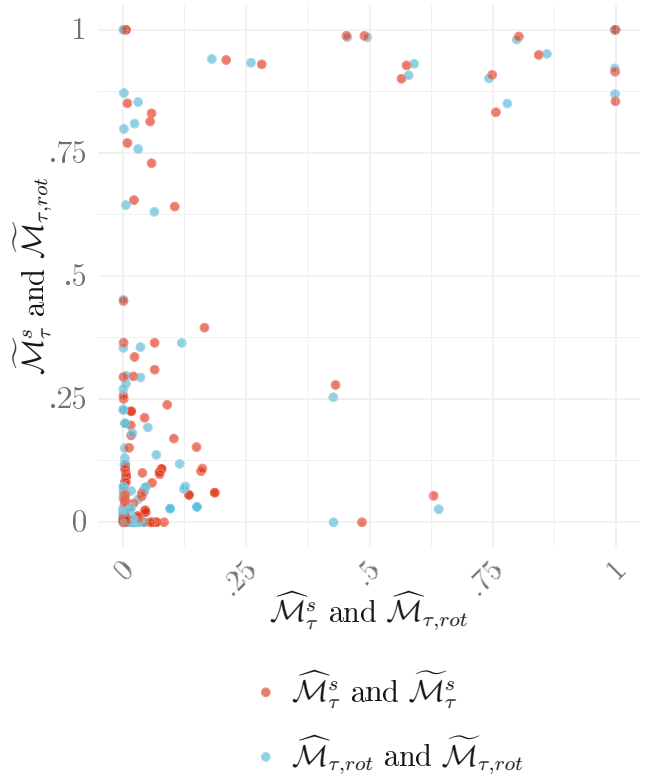

Figure G.16: Scatter plots of volatility-robust and default p-values

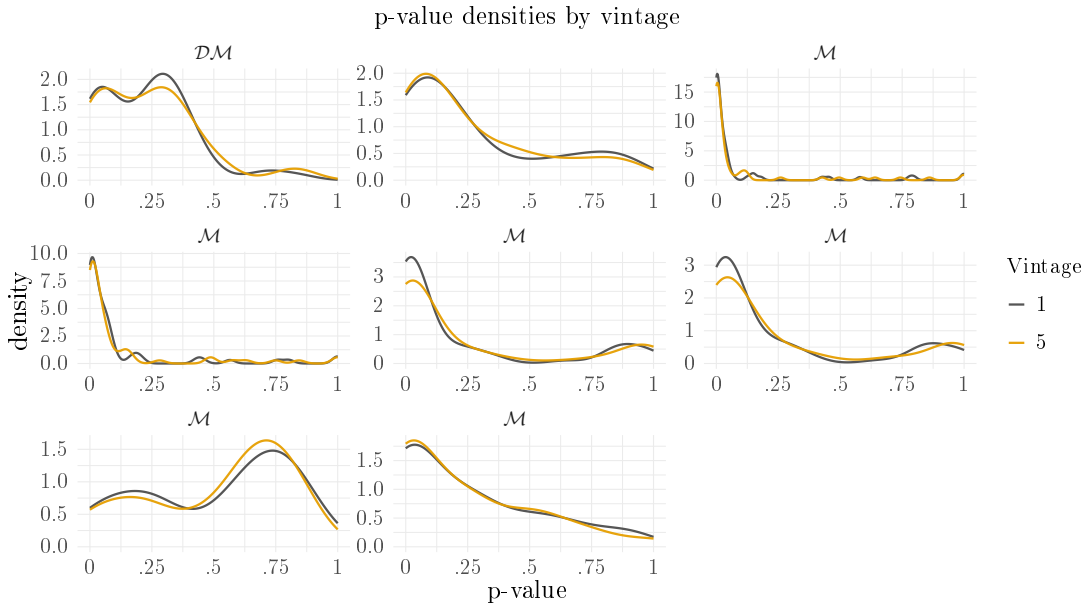

Figure G.17: The distribution of -values of the EPA tests by vintage

References

Bickel, P. J. and M. Rosenblatt (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics 1*(6), 1071–1095.

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

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

Kuelbs, J. and W. Philipp (1980). Almost sure invariance principles for partial sums of mixing -valued random variables. *The Annals of Probability 8*(6), 1003–1036.

Lan, W., B. Lei, L. Feng, and C.-L. Tsai (2024). Maximum-subsampling test of Equal Predictive Ability. *Journal of Business & Economic Statistics 42*(4), 1344–1355.

McCracken, M. W. (2020). Diverging tests of equal predictive ability. *Econometrica 88*(4), 1753–1754.

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