Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ubes20

Testing for Equal Average Forecast Accuracy in Possibly Unstable Environments

David I. Harvey, Stephen J. Leybourne & Yang Zu

To cite this article: David I. Harvey, Stephen J. Leybourne & Yang Zu (02 Dec 2024): Testing for Equal Average Forecast Accuracy in Possibly Unstable Environments, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2024.2418835

To link to this article: https://doi.org/10.1080/07350015.2024.2418835

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2024, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/07350015.2024.2418835

Testing for Equal Average Forecast Accuracy in Possibly Unstable Environments

David I. Harvey[a], Stephen J. Leybourne[a], and Yang Zu[b]

[a]School of Economics, University of Nottingham, Nottingham, UK; [b]Department of Economics, University of Macau, Taipa, Macau

ABSTRACT

We consider the issue of testing the null of equal average forecast accuracy in a model where the forecast error loss differential series has a potentially nonconstant mean function over time. We show that when time variation is present in the loss differential mean, the standard Diebold and Mariano test, which was proposed for evaluating forecasts in a stable environment, has an asymptotic size of zero, and, whilst consistent, can have reduced local power. This arises due to inconsistent estimation of the implicit long run variance estimator, which diverges under a time varying mean. We suggest a modified statistic that replaces the standard long run variance estimator based on full-sample demeaning of the loss differential series with one based on nonparametric local demeaning. The new long run variance estimator is consistent under both the null and alternative when the mean function is time varying or constant, and in both cases, the modified test recovers the asymptotic size and power properties associated with the original test in the constant mean case. The modified test therefore provides a robust method for testing the equal average forecast accuracy null, allowing for instability in the loss differential mean. The benefits of our test are demonstrated via Monte Carlo simulation and two empirical applications.

ARTICLE HISTORY
Received October 2023
Accepted October 2024

KEYWORDS
Average forecast accuracy; Diebold-Mariano test; Kernel smoothing nonparametric estimation; Time varying loss differential mean

1. Introduction

The evaluation of the accuracy of competing economic and financial forecasts has assumed a pivotal role in the literature on predictability, allowing determination of which forecasting methods perform best using historic data on forecasts and actuals. Central to such evaluation are tests for equal forecast accuracy, with a rejection indicating the superior performance of one set of forecasts over another according to a chosen measure of accuracy based on forecast error loss. Modern approaches to equal accuracy testing stem from the seminal work by Diebold and Mariano (1995) [DM], whose proposed test, based on testing for a nonzero mean in a forecast error loss differential series, d_t $(t = 1, \ldots, n)$ in generic notation, allows testing to be conducted for a user-chosen loss function under weak statistical assumptions on the d_t.

The DM approach is designed for testing equal forecast accuracy where it is assumed that the relative performance between the competing methods, measured as the mean of their loss differentials, $E(d_t)$, does not vary over the time period under consideration. However, it has since been recognized that the relative performance of different forecasting methods might vary over time; for example, Stock and Watson (2003) highlight the prevalence of instabilities in forecast rankings over time. Giacomini and White (2006) [GW], Giacomini and Rossi (2010) and Odendahl, Rossi, and Sekhposyan (2023) consider this issue in their work on forecast evaluation in potentially unstable environments, allowing for the possibility of time varying $E(d_t)$ under their alternative hypotheses. One strand of the literature has developed to test for instability in relative forecast performance, beginning with Giacomini and Rossi (2010); see also Martins and Perron (2016) and Perron and Yamamoto (2021), inter alia.

What the aforementioned approaches have in common is an assumption of the constancy of $E(d_t)$ under the *null* hypothesis. What is missing in an environment of possible instabilities, therefore, is to permit time variation in $E(d_t)$ under the null, devising procedures that test the null of equal *average* forecast accuracy across the forecast evaluation period, that is $n^{-1}\sum_{t=1}^{n} E(d_t) = 0$, without imposing $E(d_t) = 0$ for all t as in GW, Giacomini and Rossi (2010) and Odendahl, Rossi, and Sekhposyan (2023). A rejection of such a null implies that one forecast is better than another *on average*, but not necessarily at every point in the evaluation period. The interests of testing average performance superiority derive naturally from the common practice of comparing average losses in forecast evaluation. Take the commonly used prediction mean square error (MSE) as an example. Here, $d_t = e_{1t}^2 - e_{2t}^2$ where e_{1t} and e_{2t} denote two forecast error series, and the comparison of the MSEs can be understood as considering the difference in the sample means $n^{-1}\sum_{t=1}^{n} e_{1t}^2 - n^{-1}\sum_{t=1}^{n} e_{2t}^2$. In a stable environment, where $E(e_{1t}^2)$ and $E(e_{2t}^2)$ are constant over t, this can be thought of as a sample proxy for evaluating the difference $E(d_t) = E(e_{1t}^2) - E(e_{2t}^2)$. In an unstable environment with $E(e_{1t}^2)$ and $E(e_{2t}^2)$ time varying, that is, where the forecast errors are heteroscedastic, the same practice is then understood as a proxy for comparing the difference between $n^{-1}\sum_{t=1}^{n} E(e_{1t}^2)$ and $n^{-1}\sum_{t=1}^{n} E(e_{2t}^2)$,

CONTACT David I. Harvey dave.harvey@nottingham.ac.uk School of Economics, University of Nottingham, University Park, Nottingham, NG7 2RD, UK.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

that is evaluating the *average* mean $n^{-1}\sum_{t=1}^{n} E(d_t)$. A forecast evaluation test devised for the average equality hypothesis is then able to put such a comparison of MSEs into a formal testing context, such that we know if an observed difference in the MSEs is statistically "significant".

The literature contains a number of papers that also consider the average type forecast performance measure. GW explicitly consider average-type alternative hypotheses for their conditional and unconditional forecast evaluation tests, although their null hypotheses are still concerned with uniform behavior of the loss differential series at all forecasting periods. Clark and McCracken (2015) consider testing average equal predictive ability in the context of comparing nested models, but their DGP imposes certain stationarity assumptions (i.e., existence of covariance functions for the observables).

Motivated by the above discussion, in this article we address the issue of testing the null of equal average forecast accuracy in a model where the loss differential series has a potentially nonconstant mean function $E(d_t)$ over time. Our approach is based on a t-statistic associated with the sample mean $\overline{d}$ (the average of d_t over the evaluation period), which, in the unstable case, is an estimate of the full sample quantity $E(\overline{d}) = n^{-1}\sum_{t=1}^{n} E(d_t)$. Such an approach is close to the original DM testing approach; although in their paper the assumption was that $E(d_t)$ was constant, we show that the sample mean $\overline{d}$ remains an appropriate sample statistic to focus on in the potentially time varying mean environment.

The key issue here is estimating the variance of the sample mean $\overline{d}$. In a stable environment, standard long run variance (LRV) estimators used in the DM statistic provide a consistent estimate for the variance of the (normalized) sample mean. However, when $E(d_t)$ is time varying, such LRV estimators cannot account for the varying mean feature in the data and can diverge to infinity. Since our testing problem features instability under both the null and the alternative, the LRV estimator could be divergent in both cases. We show that the divergent feature of such a LRV estimator will lead to zero asymptotic size for the original DM test under the null, along with the need for a rate slower than the stable environment rate $n^{-1/2}$ for the test to achieve nontrivial power.

We therefore proceed to propose a modification to the original DM statistic, replacing the standard LRV estimator with a LRV estimator based on locally demeaned d_t, employing a kernel smoothing nonparametric estimator for the time varying mean $E(d_t)$. We demonstrate that the new LRV estimator, using locally demeaned d_t, is consistent for the true variance of $\overline{d}$ under both the null and alternative, regardless of whether $E(d_t)$ is constant or time varying. We then show that our modified DM test obtains correct asymptotic size and achieves nontrivial power against local alternatives with a standard $n^{-1/2}$ rate, again under constant or time varying $E(d_t)$. As such, the modified test can deliver valid inference in an unstable forecast evaluation environment, in contrast to the standard DM test. Our approach can therefore be interpreted as robustifying the DM test to possible time variation in the $E(d_t)$, where the null and alternative are framed in terms of average performance.

We evaluate and compare the finite sample size and power performance of the standard and modified DM tests under the null and alternative hypotheses, considering both constant and time varying specifications for $E(d_t)$. These Monte Carlo simulations demonstrate the superiority of the modified DM test when $E(d_t)$ is time varying. We further illustrate the potential value of the new procedure in evaluating forecasts of UK house price growth and US real output growth.

The remainder of the article is organized as follows. In Section 2 we introduce our modeling framework and the hypotheses being tested. Section 3 discusses the standard DM test and determines its properties within our framework, while the modified DM test is introduced in Section 4 and its properties established. Section 5 compares the finite sample size and power properties of the two tests. Our empirical applications are presented in Section 6 and Section 7 concludes. Proofs of our asymptotic results are provided in an Appendix. In the remainder of the article we use the following notation: $\xrightarrow{d}$ denotes convergence in distribution and $\xrightarrow{p}$ convergence in probability.

2. The Model and Hypotheses

We consider a framework where a sequence of n loss differentials, $d_{t,q}$, associated with two sets of q-step-ahead forecasts, can be represented in the form

$$\{d_{t,q}\}_{t=1}^{n} \equiv \{L(y_{t+q}, f_{1t,q}) - L(y_{t+q}, f_{2t,q})\}_{t=1}^{n} \tag{1}$$

where y_{t+q} denotes the target variable being predicted, $f_{it,q}$ the forecasts and $L(.,.)$ a general measurable loss function, with $t = 1, \ldots, n$ denoting the forecast evaluation period. Without loss of generality we suppress the dependence on q in all the subsequent notation. We denote the mean of the loss differential series by $E(d_t) = m_t$, $t = 1, \ldots, n$, and we make the following specific assumptions regarding m_t and d_t:

Assumption 1. $m_t = m(t/n)$, where $m(.)$ is a bounded deterministic function on $[0, 1]$ that is Lipschitz continuous other than at a finite number of possible discontinuities.

Assumption 2. d_t is α-mixing of size $-r/(r-2)$ with $r > 2$ and $E(|d_t|^r) < \infty$.

Assumption 1 allows the loss differential mean function m_t to be time varying, permitting, for example, single or multiple smooth transition level or (bounded) trend changes, while also permitting a finite number of abrupt breaks in level/trend. The polynomially decaying α-mixing coefficient specified in Assumption 2 is standard for the central limit theorem to apply. Although heteroscedasticity is not explicitly mentioned, Assumption 2 permits very general heteroscedasticity in the d_t series. Unconditional heteroscedasticity is allowed to be fully flexible; and conditional heteroscedasticity is permitted to the extent that the mixing condition is satisfied. From Carrasco and Chen (2002), many commonly used conditional heteroscedasticity models such as the ARCH model of Engle (1982), the GARCH model of Bollerslev (1986) and the log normal stochastic volatility model of Andersen (1994), when stationary, are all α-mixing with coefficients decaying exponentially fast and thus permitted by Assumption 2. Also under Assumption 2, the LRV $\lim_{n\to\infty} n^{-1}V(\sum_{t=1}^{n} d_t)$ exists

(see, e.g., Theorem 2 of Harvey, Leybourne, and Zu (2024)), and we denote it in what follows by Ω.

Remark 1. Note that heteroscedasticity in d_t is closely linked to heteroscedasticity in the forecast errors. Heteroscedasticity in forecast errors can easily induce time variation in the mean of d_t, and also time variation in the variance of d_t. For example, suppose $e_{it} = \sigma_{it}u_{it}$, $i = 1, 2$, with σ_{it} deterministic, and $u_{it} \sim IIDN(0, 1)$ with u_{1t}, u_{2t} correlated with parameter ρ. Then, if $d_t = e_{1t}^2 - e_{2t}^2$, it can easily be shown that $E(d_t) = \sigma_{1t}^2 - \sigma_{2t}^2$ and $V(d_t) = 2(\sigma_{1t}^4 + \sigma_{2t}^4) - 4\rho^2\sigma_{1t}^2\sigma_{2t}^2$, hence, the loss differential has a time-varying mean and variance. Moreover, even in situations where heteroscedastic forecast errors do not induce time variation in the mean of d_t, time variation in the variance of d_t can arise, as occurs in the previous example when $\sigma_{1t} = \sigma_{2t}$. Hence, in an unstable environment, where forecast error variances are expected to change over time, it is important that we allow for time variation in both the mean and variance of d_t, as covered by Assumptions 1 and 2, respectively.

Remark 2. Our framework treats the forecasts as primitives, and makes assumptions directly on the loss differential series d_t. However, it is also compatible with a GW-type framework where forecasts are obtained from estimated models using a fixed number of observations (i.e., a rolling window estimation scheme). That is, where the forecasts f_{it}, $i = 1, 2$ in (1) are replaced with the forecast methods $f_i(W_t, \ldots, W_{t-w_i+1}; \hat{\beta}_{i,t})$ where W_t is a data vector partitioned as $W_t = (y_t, X_t')'$, with X_t is a vector of predictors, and where the model parameter estimates $\hat{\beta}_{i,t}$ are obtained over $t - w_i + 1, \ldots, t$ with w_i fixed. In this setting, the assumption for d_t in Assumption 2 would implicitly be applied to the data vector W_t.

The focus of our analysis in this article concerns testing based on the average forecast accuracy quantity $n^{-1}\sum_{t=1}^{n} E(d_t)$. Specifically, we will test the null hypothesis of $n^{-1}\sum_{t=1}^{n} E(d_t) = 0$ against the alternative $n^{-1}\sum_{t=1}^{n} E(d_t) \neq 0$. We will formalize this testing framework by considering the corresponding integral measure of average forecast accuracy based on $E(d_t) = m_t$:

$$\int_0^1 m(x)\mathrm{d}x.$$

A quantity that will also be important in our analysis is the variation in the mean function m_t over time, which we denote by

$$V_m = \int_0^1 m(x)^2\mathrm{d}x - (\int_0^1 m(x)\mathrm{d}x)^2$$

with $V_m = 0$ and $V_m > 0$ corresponding to the cases of constant m_t and time varying m_t, respectively.

For the purposes of specifying the null and (fixed and local) alternative hypotheses in our testing problem, we suppose that $m(x)$ can be written as

$$m(x) = m^0(x) + \gamma_n m^1(x)$$

where $m^0(.)$ and $m^1(.)$ are, similar to $m(.)$, bounded deterministic functions on $[0, 1]$ that are continuous other than a finite number of possible discontinuities. Here

$$\int_0^1 m(x)\mathrm{d}x = \int_0^1 m^0(x)\mathrm{d}x + \gamma_n \int_0^1 m^1(x)\mathrm{d}x$$

and we assume that

$$\int_0^1 m^0(x)\mathrm{d}x = \mu_0 \text{ with } \mu_0 = 0$$
$$\int_0^1 m^1(x)\mathrm{d}x = \mu_1 \text{ with } \mu_1 \neq 0.$$

Our null hypothesis of equal average forecast accuracy is consequently given by

$$H_0 : \gamma_n = 0, \text{ that is } \int_0^1 m(x)\mathrm{d}x = \int_0^1 m^0(x)\mathrm{d}x = 0.$$

The null clearly nests two possibilities: (i) when m_t is constant, $\int_0^1 m^0(x)\mathrm{d}x = 0$ implies that the forecasts have equal accuracy at all points in time (i.e., the standard DM-type null); (ii) when m_t is time varying, the null requires the forecasts to have equal accuracy on average, in which case clearly each forecast must be dominant over the other (in terms of accuracy) at different points during the forecast evaluation period.

As regards the alternative hypothesis, we first consider the fixed alternative

$$H_1 : \gamma_n = \gamma \neq 0, \text{ that is}$$
$$\int_0^1 m(x)\mathrm{d}x = \gamma \int_0^1 m^1(x)\mathrm{d}x = \gamma\mu_1, \ \mu_1 \neq 0.$$

Of course, the alternative can hold either when one forecast possesses relative superior accuracy throughout the whole evaluation period, or where each forecast has periods of dominance, but one outperforms the other on average.

It should be noted that the equal average forecast accuracy null, and corresponding alternatives, relate specifically to the sample period $t = 1, \ldots, n$ under investigation. For example, it is quite possible that the forecasts are equally accurate on average over $t = 1, \ldots, n$, but one outperforms the other on average if a different sample period is considered, or vice-versa.

2.1. Comparison with Extant Literature

To place our model and hypotheses in the context of previous work in the field of equal forecast accuracy testing, we first note that the original DM approach is a special case of our approach, with $E(d_t)$ constant. Their approach sets $E(d_t) = c$, with $c = 0$ under the null and $c \neq 0$ under the alternative. While the unconditional test statistic of GW is the same as the DM statistic, their alternative hypothesis is more flexible, allowing for time varying $E(d_t)$. Their null is that $E(d_t) = 0$ for all t (hence, $E(d_t)$ is actually constant at zero under the null, as in DM), while the alternative is that $n^{-1}\sum_{t=1}^{n} E(d_t) \neq 0$, with one forecast outperforming the other on average. Compared to our setup, it is clear that the alternative hypothesis of GW coincides with ours, but their null hypothesis is more restrictive, requiring the forecasts to have equal accuracy at all points in time, rather than simply equal on average.

Giacomini and Rossi (2010) also consider a framework where $E(d_t) = 0$ for all t under the null (as in DM and GW), but have a completely different focus to ours under the alternative, with their alternative simply requiring that $E(d_t) \neq 0$ for at least one point in time. Hence, while time variation in $E(d_t)$ under

the alternative is permitted, the emphasis is on any departure from the null, rather than specific forms such as a constant difference in forecast accuracy, as in DM, or a difference in forecast accuracy on average, as in GW and our setup. Obviously then, their alternative does not involve any concept of average behavior of time varying $E(d_t)$.

Of all the approaches outlined thus far, ours is the only one that permits time variation in $E(d_t)$ under the null, and also the only one to specify the null and alternative in terms of the same accuracy measure, namely the average forecast accuracy $n^{-1}\sum_{t=1}^{n} E(d_t)$. In a recent working paper, Richter and Smetanina (2020) also consider time variation under the null and alternative, with the hypotheses specified using a single accuracy measure. In the absence of serial correlation in d_t, their average forecast accuracy measure is based on the average of inverse coefficients of variation: $n^{-1}\sum_{t=1}^{n} E(d_t)/\sigma_t$, where $\sigma_t^2 = V(d_t)$. The null and alternative hypotheses are given by $n^{-1}\sum_{t=1}^{n} E(d_t)/\sigma_t = 0$ and $n^{-1}\sum_{t=1}^{n} E(d_t)/\sigma_t \neq 0$, respectively. By introducing σ_t in this way, the accuracy measure is based on the behavior of both the time varying mean $E(d_t)$ and unconditional heteroscedasticity in the loss differential series, with relative performance assessed by an interaction of the two quantities over time. Under homoscedasticity, that is $\sigma_t = \sigma$ for all t, their measure essentially coincides with ours, since $n^{-1}\sum_{t=1}^{n} E(d_t)/\sigma_t = \sigma^{-1}n^{-1}\sum_{t=1}^{n} E(d_t)$, with this quantity equal to zero if and only if $n^{-1}\sum_{t=1}^{n} E(d_t) = 0$. However, the measures are quite different under heteroscedasticity. Indeed, it is quite possible that two forecasts could be equally accurate in terms of average mean, that is $n^{-1}\sum_{t=1}^{n} E(d_t) = 0$, but due to changes in σ_t, $n^{-1}\sum_{t=1}^{n} E(d_t)/\sigma_t \neq 0$. In such a situation, our null hypothesis holds, but it is the alternative hypothesis that is true in the Richter and Smetanina (2020) framework. As such, examining the behavior of $n^{-1}\sum_{t=1}^{n} E(d_t)/\sigma_t$ is not informative for evaluating whether $n^{-1}\sum_{t=1}^{n} E(d_t)$ is zero or not. Our proposal is to use the unweighted loss differential mean in the accuracy measure, which is arguably a more natural or direct approach to examining average forecast accuracy, and is in keeping with the work of GW, who use precisely this measure under their alternative hypothesis.

Another strand of related literature addresses the notion of conditional equality tests; for example, GW consider a conditional approach in addition to the unconditional test we discuss above. We wish to emphasize that the test considered in this article focuses on time variation in the *unconditional* mean of the loss differential series, rather than the *conditional* mean. Conditional mean variation can exist separately to unconditional mean variation; for example, even in a stationary and stable framework, the conditional mean of the loss differential can change. By way of illustration, consider an $AR(1)$ loss differential process $d_t = \phi d_{t-1} + \varepsilon_t$, where $|\phi| < 1$ and ε_t is a white-noise sequence. This is a case we would normally refer to as a stable environment, due to the covariance stationarity property of the process, with the unconditional mean of d_t, $E(d_t)$, being constant at zero. In contrast, the one-step-ahead conditional mean of the d_t series, $E(d_t|\mathcal{F}_{t-1})$, is ϕd_{t-1}, which is time varying. The conditional equality concept implies the loss differential is a martingale difference sequence, that is $E(d_t|\mathcal{F}_{t-1}) = 0$, and therefore any dependence in d_t represents conditional inequality in terms of forecast accuracy. This is a very strong equality concept between two forecasts, as Zhu and Timmermann (2020) discuss, and much stronger than the unconditional concept that we adopt in this article. As illustrated in Zhu and Timmermann (2022), conditional tests can be used to identify factors that explain the loss differentials, and to help devise a dynamic rotation strategy to obtain a more accurate combined forecast.

3. The DM (Unconditional GW) Test

The DM statistic, which is also GW's unconditional statistic, is given by

$$DM = \frac{\sqrt{n}\,\bar{d}}{\sqrt{\hat{\Omega}}}$$

where $\bar{d} = n^{-1}\sum_{t=1}^{n} d_t$ and $\hat{\Omega}$ is an estimate of Ω, the LRV of d_t. The LRV estimator $\hat{\Omega}$ implicitly assumes m_t is constant and is given by

$$\hat{\Omega} = n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n}(d_t - \bar{d})(d_s - \bar{d})k\left(\frac{t-s}{b}\right) \tag{2}$$

where $k(.)$ is a kernel function, and b is a lag truncation parameter. For the purposes of our analysis we make the following assumptions regarding $k(.)$ and b:

Assumption 3. The kernel function $k(.)$ is symmetric, satisfies $|k(.)| \leqslant 1$ and $k(0) = 1$, and is continuous at zero and almost everywhere else. The kernel function also satisfies $\int_{-\infty}^{\infty}|k(x)|\mathrm{d}x < \infty$, $\int_{-\infty}^{\infty}|xk(x)|\mathrm{d}x < \infty$ and $\int_{-\infty}^{\infty}|\phi_k(x)|\mathrm{d}x < \infty$ where $\phi_k(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}k(y)e^{-\mathrm{i}xy}\mathrm{d}y$. Let $\lambda_k = \int_0^{\infty}k(x)\mathrm{d}x$.

Assumption 4. The bandwidth b satisfies $b \to \infty$ and $n^{-1/2}b \to 0$ as $n \to \infty$.

These assumptions are similar to those used in De Jong and Davidson (2000). The condition $n^{-1/2}b \to 0$ is stronger than the $n^{-1}b \to 0$ assumption in De Jong and Davidson (2000), which we need here to deal with the effect of the time-varying mean. This condition is standard in the LRV estimation literature, and, as discussed in Andrews (1991), optimal bandwidths for all the commonly used kernels are permitted by this condition.

Our first result establishes the asymptotic behavior of the LRV estimator $\hat{\Omega}$ in the denominator of DM.

Theorem 1. Under Assumptions 1–4 and under H_0 or H_1,

$$\hat{\Omega} = \Omega + (1 + 2b\lambda_k)V_m + o_p(1). \tag{3}$$

Theorem 1 shows that when $m(x)$ is time varying ($V_m > 0$), $\hat{\Omega}$ is not a consistent estimator of Ω, but is diverging toward $+\infty$ at the rate of b; while when $m(x)$ is constant ($V_m = 0$), $\hat{\Omega}$ is consistent for Ω. The inconsistency of $\hat{\Omega}$ when $m(x)$ is time varying arises because, in (2), $\bar{d}$ is not the appropriate centering for d_t, since $E(\bar{d}) = n^{-1}\sum_{t=1}^{n} m_t$ while $E(d_t) = m_t \neq E(\bar{d})$ in general. A similar result to Theorem 1 was obtained in Hall (2005) (see eq. (4.28)), for the "uncentered HAC estimator" with strictly stationary, ergodic data.

Next, we proceed to consider the asymptotic size and power properties of the *DM* test. Recall that the (two-sided) α-level DM test is defined as rejecting H_0 if $|DM| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

Theorem 2. Under Assumptions 1–4:

(i) Under H_0,
  (a) when $m(x)$ is constant, $P(|DM| > z_{1-\alpha/2}) \to \alpha$;
  (b) when $m(x)$ is time varying, $P(|DM| > z_{1-\alpha/2}) \to 0$.

(ii) Under H_1, $P(|DM| > z_{1-\alpha/2}) \to 1$, when $m(x)$ is either constant or time varying.

It can be seen from Theorem 2(i) that, while the *DM* test is asymptotically correctly sized under H_0 when $m(x)$ is constant, whenever $m(x)$ is time varying it will have asymptotic size of zero. Under the alternative H_1, Theorem 2(ii) shows that the *DM* test is consistent, regardless of whether $m(x)$ is constant or time varying, hence, in this respect the *DM* test remains useful for detecting departures from the equal average forecast accuracy null.

In addition to the fixed alternative hypothesis H_1, to enable further evaluation of the large sample power properties of the tests, we also consider a sequence of local alternatives of the following form:

$$H_{1n} : \gamma_n \to 0 \text{ as } n \to \infty.$$

Here, the $m(x)$ function is comprised of the null hypothesis function $m^0(x)$ plus an asymptotically vanishing perturbation $\gamma_n m^1(x)$. The next theorem finds the sequences of local alternatives such that the *DM* test has nontrivial power, and derives its local power functions. It turns out that, depending on whether or not the function $m(x)$ is time varying, the *DM* test has nontrivial power for local alternative sequences converging to the null at different rates in n.

Theorem 3. Under Assumptions 1–4:

(i) Under H_{1n} when $m(x)$ is constant with $\gamma_n = n^{-1/2}$, $DM \xrightarrow{d} N(\Delta', 1)$, where $\Delta' = \int_0^1 m^1(x)\mathrm{d}x/\sqrt{\Omega}$, hence,

$$P(|DM| > z_{1-\alpha/2}) \to 1 - \Phi(z_{1-\alpha/2} - \Delta') + \Phi(-z_{1-\alpha/2} - \Delta')$$

where $\Phi(.)$ is the cdf of the standard normal distribution;

(ii) Under H_{1n} when $m(x)$ is time varying, with $\gamma_n = (n/b)^{-1/2}$, $DM \xrightarrow{p} \Delta$, where $\Delta = \int_0^1 m^1(x)\mathrm{d}x/\sqrt{2\lambda_k V_m}$, hence

$$P(|DM| > z_{1-\alpha/2}) \to \begin{cases} 0 & \text{if} \Delta \leqslant z_{1-\alpha/2} \\ 1 & \text{if} \Delta > z_{1-\alpha/2} \end{cases}.$$

From Theorem 3(i) we see that when $m(x)$ is constant, the *DM* test possesses nontrivial power against local alternatives converging to the null at the standard parametric rate $\gamma_n = n^{-1/2}$. Theorem 3(ii) shows that when $m(x)$ is time varying, the *DM* test only has nontrivial power against local alternatives converging to the null at a slower nonparametric rate $\gamma_n = (n/b)^{-1/2}$. Implicit in the proof of Theorem 3 is that in this case, the *DM* test will have zero asymptotic power against local alternatives with rates faster than $(n/b)^{-1/2}$ (smaller deviations from the null), such as the parametric rate $n^{-1/2}$. Therefore, the time variation in $m(x)$ compromises the ability of the *DM* test to detect departures from the null relative to the constant $m(x)$ case. When $\gamma_n = (n/b)^{-1/2}$, the local asymptotic distribution of the *DM* statistic is also nonstandard: it degenerates to a constant, such that the local power function of the *DM* test in this case is an indicator function with the condition that Δ is greater than the critical value $z_{1-\alpha/2}$.

4. A Modified DM Test

In this section we develop a new DM-type test, which, regardless of whether $m(x)$ is constant or time varying, has correct asymptotic size under H_0 and attractive large sample properties under fixed and local alternatives. The idea is simple: we replace $\hat{\Omega}$ in the *DM* statistic with a new LRV estimator, which is consistent for the LRV even when the loss differential series has a time varying mean. The new LRV estimator is constructed by replacing the $\bar{d}$ centering in $d_t - \bar{d}$ in (2) with a nonparametric estimate of m_t, to achieve correct demeaning under time variation given that $E(d_t) = m_t$. To this end, we consider the following nonparametric kernel smoothing estimator for m_t at each point in time t:

$$\hat{m}_t = \sum_{s=1}^{n} w_{t,s} d_s, \quad t = 1, \ldots, n$$

where $w_{t,s} = K\left(\frac{s-t}{nh}\right) / \sum_{s=1}^{n} K\left(\frac{s-t}{nh}\right)$ with $K(.)$ a kernel function and h the bandwidth. Replacing $\bar{d}$ in (2) with $\hat{m}_t$ then gives our modified LRV estimator:

$$\hat{\Omega}' = n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} (d_t - \hat{m}_t)(d_s - \hat{m}_s) k\left(\frac{t-s}{b}\right). \quad (4)$$

We make the following assumptions on $K(.)$ and h:

Assumption 5. The kernel function $K(.) : \mathbb{R} \to \mathbb{R}^+$ is bounded, Lipschitz continuous and satisfies $\int_{-\infty}^{\infty} |K(x)|\mathrm{d}x < \infty$, $\int_{-\infty}^{\infty} |K(x)x|\mathrm{d}x < \infty$ and $x^2 K(x) \to 0$ as $x \to \infty$.

Assumption 6. The bandwidth h satisfies $h \to 0$, $bh \to 0$ and $b/(nh) \to 0$, as $n \to \infty$.

The assumptions on the kernel function $K(.)$ in Assumption 5 are rather standard in the literature involving nonparametric kernel regression estimators. Assumption 4 implies that $b = O(n^{\tau_b})$ with $0 < \tau_b < 1/2$; given this constraint, Assumption 6 then implies that $h = O(n^{-\tau_h})$ with $\tau_b < \tau_h < 1 - \tau_b$. Although in principle the rates τ_b and τ_h are chosen simultaneously, in practice we recommend that τ_b is chosen first, and then a suitable value of τ_h is chosen subsequently. We discuss the choices of h and b in more detail in Section 5.

The following theorem establishes the consistency of $\hat{\Omega}'$ in (4).

Theorem 4. Under Assumptions 1–6 and under H_0 or H_1, when $m(x)$ is either constant or time varying, $\hat{\Omega}' \xrightarrow{p} \Omega$.

Theorem 3 shows that $\hat{\Omega}'$ is a consistent estimator of Ω irrespective of whether or not $m(x)$ is time varying, in contrast to the result for $\hat{\Omega}$ shown in Theorem 1. Note that Theorem 3 holds both in the case of $m(x)$ being continuous and in the case where $m(x)$ has points of discontinuity (satisfying Assumption 1). In the latter case, the nonparametric estimator $\hat{m}_t$ cannot be uniformly consistent, but crucially the consistency of the new HAC estimator $\hat{\Omega}'$ is not affected.

Remark 3. LRV estimators similar to ours which are consistent under a time-varying mean function have already been considered in the literature, with Altissimo and Corradi (2003) and Juhl and Xiao (2009) proposing very similar estimators to ours. However, Altissimo and Corradi (2003) consider a piecewise-constant alternative model, and the local mean estimator they consider is a simple average over local windows (i.e., $K(.)$ is a flat truncated kernel). Juhl and Xiao (2009) consider a smooth varying alternative model, whereas we permit a finite number of discontinuities; these authors also make a fourth order stationary β-mixing assumption for their stochastic component, thereby not allowing for unconditional heteroscedasticity.

Remark 4. In a recent paper, Chan (2022) analyses a general class of *difference-based* LRV estimators. As discussed in Example 2.5 of Chan (2022), the form of our proposed estimator $\hat{\Omega}'$ can be viewed as a special case of such difference-based estimators. However, Chan's results do not apply to our setting because the probabilistic assumptions on the stochastic component are different. In particular, Chan assumes strict stationarity, so no unconditional heteroscedasticity is permitted, and a dependence structure characterized by the *physical dependence measure* proposed by Wu (2007). As discussed on p.1379 of Chan (2022), the theoretical properties of difference-based estimators remain unknown under the mixing type assumptions that we consider in this article.

Remark 5. As highlighted by a referee, it is possible to conceive of kernels $k(.)$ for which $\lambda_k = \int_0^\infty k(x)\mathrm{d}x = 0$. Use of such a kernel would remove the divergence problem inherent in (3), since then $\hat{\Omega} = \Omega + V_m + o_p(1)$. In such a case, while $\hat{\Omega}$ no longer diverges, it remains inconsistent due to the bias term V_m. Of course, V_m can be consistently estimated using our nonparametric mean function estimator $\hat{m}_t$, that is

$$\hat{V}_m = \frac{1}{n}\sum_{t=1}^{n}\hat{m}_t^2 - \left(\frac{1}{n}\sum_{t=1}^{n}\hat{m}_t\right)^2 \tag{5}$$

and $\hat{\Omega} - \hat{V}_m$ would be a consistent estimator of Ω. However, issues arise when attempting to implement this approach in practice. An obvious choice of kernel for which $\lambda_k = 0$ would be a Bartlett kernel with its domain extended from $[-1, 1]$ (which gives $\lambda_k = 1$) to $[-2, 2]$ (which gives $\lambda_k = 0$), that is

$$k(x) = \begin{cases} 1 - |x| & -2 \leqslant x \leqslant 2 \\ 0 & \text{otherwise} \end{cases}.$$

The first difficulty with this approach is that the absolute integrability element $\int_{-\infty}^{\infty}|\phi_k(x)|\mathrm{d}x < \infty$ of Assumption 3 is violated, hence, stronger assumptions would be required to ensure its asymptotic validity. A similar problem arises when considering modifications to the Parzen and Tukey-Hanning kernels. Even abstracting from this issue, we found through simulation that convergence of the corresponding estimator $\hat{\Omega}$ to its limit $\Omega + V_m$ is very slow, with substantial upward biases observed in finite samples (unless $V_m = 0$). It appears that the inclusion of substantial negative kernel weights attached to high order sample autocovariances compromises the performance of the estimator, unless the sample size is very large. Use of such an estimator in *DM* then resulted in a procedure that was severely under-sized and manifested low power in finite samples (compared to what we subsequently observe in our modified test statistic that uses $\hat{\Omega}'$). In principle, entirely new kernels satisfying both $\lambda_k = 0$ and Assumption 3 could be devised, ideally with better finite sample properties, but we leave such developments for future investigation.

In view of the result in Theorem 4, we define our modified *DM*-type statistic as

$$DM' = \frac{\sqrt{n}\,\bar{d}}{\sqrt{\hat{\Omega}'}}.$$

We then obtain the following results regarding the behavior of DM' under H_0.

Theorem 5. Under Assumptions 1–6, when $m(x)$ is either constant or time varying:

(i) Under H_0, $DM' \xrightarrow{d} N(0,1)$.

(ii) Under H_1, $P(|DM'| > z_{1-\alpha/2}) \to 1$.

Theorem 5 shows that our new test, based on comparing DM' with standard normal critical values, has correct asymptotic size and is consistent against a fixed alternative, regardless of whether or not the $m(x)$ function is constant.

The next theorem studies the asymptotic power properties of the new test under local alternatives of the form H_{1n}.

Theorem 6. Under Assumptions 1–6 and under H_{1n} with $\gamma_n = n^{-1/2}$, when $m(x)$ is either constant or time varying, $DM' \xrightarrow{d} N(\Delta', 1)$, and hence

$$P(|DM'| > z_{1-\alpha/2}) \to 1 - \Phi(z_{1-\alpha/2} - \Delta') + \Phi(-z_{1-\alpha/2} - \Delta').$$

Theorem 6 shows that the DM' test has nontrivial power against the sequence of local alternatives with $\gamma_n = n^{-1/2}$, irrespective of whether $m(x)$ is constant or time varying. The *DM* test achieves power for this rate only when $m(x)$ is constant, having zero asymptotic power in the time varying case. Moreover, when $m(x)$ is constant or time varying, the DM' test has the exact same local power function as that for *DM* under constant $m(x)$. Therefore, the loss of power that occurs with *DM* for time varying $m(x)$ is fully restored by DM' to the level associated with constant $m(x)$. Implicit in Theorem 6 is the result that for local alternatives with rates slower than $n^{-1/2}$, the DM' test has asymptotic power of one.

Overall, the results of Theorems 2–3 and 5–6 suggest that *DM* and DM' would be expected to have similar finite sample properties under constant $m(x)$, while under time varying $m(x)$, DM' should have size closer to the nominal level and superior

finite sample power. The next section examines these finite sample properties in detail.

5. Finite Sample Size and Power Properties

We now compare the finite sample size and power properties of the DM and DM' tests using Monte Carlo simulation. For d_t, we consider the DGP

$$d_t = m_t + \nu_t \quad t = 1, \ldots, n. \tag{6}$$

Here ν_t is a mean zero stochastic process. In the deterministic mean function m_t we make use of a logistic smooth transition function of the form

$$S_t(c, g, \delta_1, \delta_2) = \frac{\delta_2 - \delta_1}{1 + \exp\{-g(\frac{t-1}{n-1} - c)\}} + \delta_1.$$

This function transitions from the value δ_1 to δ_2 over t, with midpoint fraction c and transition speed g. In what follows we set $\delta_1 = -1$, $\delta_2 = 1$ and $g = 30$. We further define two special cases of $S_t(c, 30, -1, 1)$. For $c = 0.5$, define S_t^0 as $S_t^0 = S_t(0.5, 30, -1, 1)$. Here

$$n^{-1} \sum_{t=1}^{n} S_t^0 = 0.$$

For $c \neq 0.5$ define $S_t^1(c) = S_t(c, 30, -1, 1)$. Now

$$n^{-1} \sum_{t=1}^{n} S_t^1(c) \neq 0$$

and for $c = 0.25, 0.75$, $n^{-1} \sum_{t=1}^{n} S_t^1(c) = 0.5, -0.5$, respectively.

We construct DM and DM' using their respective LRV estimators, $\hat{\Omega}$ and $\hat{\Omega}'$, employing the quadratic spectral (QS) kernel for $k(.)$ with bandwidth $b = b_0 n^{1/3}$ (that satisfies the requirement $b = O(n^{\tau_b})$ with $0 < \tau_b < 1/2$). Choice of the QS kernel is motivated by its optimality properties, in that it minimizes the asymptotic MSE among positive semidefinite LRV estimators (see Andrews (1991)) and is also optimal for testing purposes in the context of heteroscedasticity and autocorrelation robust inference problems (see Sun et al. (2008), and Lazarus, Lewis, and Stock (2021)). With respect to the bandwidth choice, both Sun et al. (2008) and Lazarus, Lewis, and Stock (2021) derive the optimal rate for the QS kernel lag truncation parameter to be $O(n^{1/3})$. Although there are different strategies to calculating b_0 in Sun et al. (2008) and Lazarus, Lewis, and Stock (2021), their approaches do not apply directly, and as suggested in Section 5 of Lazarus, Lewis, and Stock (2021), we use our own judgment to select b_0. For the local mean estimator $\hat{m}_t$ in $\hat{\Omega}'$ we use the Gaussian kernel for $K(.)$. Given that $b = b_0 n^{1/3}$, we require a rate setting for the bandwidth h that satisfies $h = O(n^{-\tau_h})$ with $1/3 < \tau_h < 2/3$. Through extensive simulations, we found that setting $b_0 = 1.5$ and $h = h_0 n^{-2/5}$ with $h_0 = 0.25$ delivered a good balance of finite sample size and power performance across a range of null and alternative model specifications, and we therefore adopt these throughout the remainder of the article. In what follows, our finite sample simulations are based on 50,000 replications. We perform nominal 0.05-level two-tailed tests. Because the tests' sizes can be sensitive to the choice of kernels and associated bandidths b and h in relatively small samples, it is important in practice to use finite sample null critical values for DM and DM'. We do this for our particular kernel choices and settings for b and h, based on simulation of (6) with $m_t = 0$ and ν_t generated as IID $N(0, 1)$ variates—see the supplementary appendix for further details.

5.1. Finite Sample Size

We first examine the finite sample sizes of DM and DM' under the null hypothesis when d_t is generated according to (6) with $m_t = aS_t^0$. Here, if $a = 0$, m_t is constant at 0, while if $a > 0$, m_t is time varying, moving smoothly from $-a$ to a, with $V_m = a^2 \int_0^1 m(x)^2 dx = 0.867a^2$. We consider the range of values $a \in \{0, 0.1, 0.2, \ldots, 0.5\}$.

Our first size simulations are for a homoscedastic case where we generate ν_t according to the $ARMA(1, 1)$ process

$$\nu_t = \phi \nu_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$$

with $\nu_1 = \varepsilon_1$, and the ε_t are IID $N(0, 1)$ variates. We consider the $ARMA(1, 1)$ parameter settings $\phi, \theta \in \{-0.5, 0, 0.5\}$ and the sample sizes $n \in \{75, 150, 300\}$. In order to provide an illustration of the amount of mean variation occurring relative to the stochastic component of the series, in Figure S1 of the supplementary appendix we plot the generated process d_t for the first Monte Carlo replication, along with the corresponding mean path m_t, for the three cases $a = 0.1$, 0.3, and 0.5 when $T = 150$ and $\phi = \theta = 0$. It is clear that as a increases, the amount of mean variation (i.e., V_m) also increases, but the movement in the mean never dominates the series, and even for the most extreme case of $a = 0.5$, the influence of the mean shift on the data is rather subtle. We therefore consider these settings to be a plausible representation of what might be encountered in practice.

The simulation results are shown in Table 1. In addition to the test sizes, we also report the values of the mean variation measure V_m that correspond to each value of a, and the values of the LRV Ω that correspond to each setting of ϕ and θ. The rows in the table are ordered by decreasing magnitude of Ω. Focusing first on results for the central serially uncorrelated case $\phi = \theta = 0$, we observe that, as a consequence of the method for finite sample critical value generation, DM is correctly sized when $V_m = 0$, but its size approaches zero as the amount of time variation V_m increases away from zero. This under-sizing also becomes more evident as the sample size n increases, in line with the limiting result established in Theorem 2(i)(b). In contrast, the size of DM' remains close to the nominal level across all V_m and sample sizes; some very modest under-sizing is apparent for the larger values of V_m when $n = 75$, but this under-sizing disappears rapidly as n increases.

Once we introduce serial correlation, for DM we observe a broadly similar pattern of results to the uncorrelated case: for given values of ϕ and θ, size is fairly close to the nominal level when $V_m = 0$ (some modest size distortions are observed with the direction and magnitude of these corresponding to the magnitude of Ω), but when $V_m > 0$, size decreases as the magnitude of V_m increases and as the magnitude of Ω decreases. These reductions in test size are also more exaggerated as n

Table 1. Finite sample size of nominal 0.05-level tests: $m_t = aS_t^0, v_t = \phi v_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}$.

				DM						*DM′*					
			a :	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5
ϕ	θ	Ω	V_m :	0.000	0.009	0.035	0.078	0.139	0.217	0.000	0.009	0.035	0.078	0.139	0.217
							Panel A. $n = 75$								
0.5	−0.5	9.000		0.067	0.066	0.063	0.059	0.054	0.047	0.088	0.088	0.088	0.088	0.087	0.087
0.5	0	4.000		0.065	0.063	0.058	0.049	0.039	0.029	0.083	0.083	0.082	0.082	0.082	0.081
0	−0.5	2.250		0.052	0.050	0.042	0.032	0.021	0.013	0.055	0.055	0.054	0.054	0.054	0.053
0	0	1.000		0.050	0.045	0.031	0.016	0.006	0.002	0.050	0.050	0.050	0.049	0.048	0.047
−0.5	0	0.444		0.048	0.036	0.015	0.004	0.000	0.000	0.044	0.044	0.044	0.042	0.041	0.039
0	0.5	0.250		0.036	0.022	0.005	0.000	0.000	0.000	0.026	0.026	0.025	0.024	0.022	0.021
−0.5	0.5	0.111		0.033	0.011	0.000	0.000	0.000	0.000	0.020	0.020	0.019	0.017	0.015	0.012
							Panel B. $n = 150$								
0.5	−0.5	9.000		0.061	0.059	0.056	0.050	0.044	0.037	0.074	0.074	0.074	0.074	0.074	0.074
0.5	0	4.000		0.059	0.057	0.049	0.040	0.029	0.019	0.071	0.071	0.071	0.071	0.071	0.070
0	−0.5	2.250		0.051	0.047	0.037	0.025	0.014	0.007	0.053	0.053	0.053	0.053	0.052	0.052
0	0	1.000		0.050	0.042	0.024	0.010	0.003	0.001	0.050	0.050	0.050	0.050	0.049	0.049
−0.5	0	0.444		0.049	0.032	0.010	0.002	0.000	0.000	0.047	0.047	0.046	0.046	0.045	0.044
0	0.5	0.250		0.040	0.020	0.003	0.000	0.000	0.000	0.032	0.032	0.032	0.031	0.030	0.029
−0.5	0.5	0.111		0.039	0.008	0.000	0.000	0.000	0.000	0.029	0.029	0.028	0.027	0.025	0.024
							Panel C. $n = 300$								
0.5	−0.5	9.000		0.057	0.055	0.051	0.044	0.037	0.029	0.065	0.065	0.065	0.065	0.065	0.065
0.5	0	4.000		0.056	0.053	0.044	0.032	0.022	0.013	0.063	0.063	0.063	0.063	0.063	0.063
0	−0.5	2.250		0.051	0.045	0.033	0.020	0.009	0.004	0.052	0.052	0.052	0.052	0.052	0.052
0	0	1.000		0.050	0.038	0.019	0.006	0.002	0.000	0.050	0.050	0.050	0.050	0.050	0.050
−0.5	0	0.444		0.049	0.029	0.006	0.001	0.000	0.000	0.049	0.048	0.048	0.048	0.048	0.048
0	0.5	0.250		0.044	0.017	0.001	0.000	0.000	0.000	0.038	0.038	0.038	0.038	0.038	0.037
−0.5	0.5	0.111		0.043	0.006	0.000	0.000	0.000	0.000	0.036	0.036	0.036	0.035	0.035	0.034

increases. The pattern of size results are as would be expected from the limit result (3), with the size distortions being most apparent when the term driving the inconsistency of $\hat{\Omega}$, that is $(2b\lambda_k+1)V_m$, is large relative to Ω. As regards DM', when $V_m = 0$ we observe a similar pattern of small sample size distortion to DM (again corresponding to the magnitude of Ω), although the magnitudes of the distortions are somewhat more pronounced. These features of size inflation and deflation are rapidly reduced with increasing sample size. When $V_m > 0$, we observe that DM' size again remains relatively unaffected by the magnitude of V_m.

Our second size simulations are for the heteroscedastic case where $\nu_t = \sigma_t\varepsilon_t$ where the ε_t are IID $N(0,1)$ variates. For σ_t we employ the smooth transition function $\sigma_t = S_t(c, 30, 1, \delta_2^*)$ with $\delta_2^* \in \{1/6, 1/3, 3, 6\}$ and $c \in \{0.25, 0.5, 0.75\}$. Here, for example $\sigma_t = S_t(0.5, 30, 1, 1/3)$ implies a downward transition in σ_t from 1 to 1/3 over t, with midpoint fraction 0.5 and transition speed 30. The results are shown in Table 2 for $n = 150$. The sizes of DM again tend to approach zero with increasing V_m. The sizes of DM', on the other hand, appear to be relatively well controlled across V_m, but do exhibit some modest over-sizing for the more extreme change patterns in σ_t.

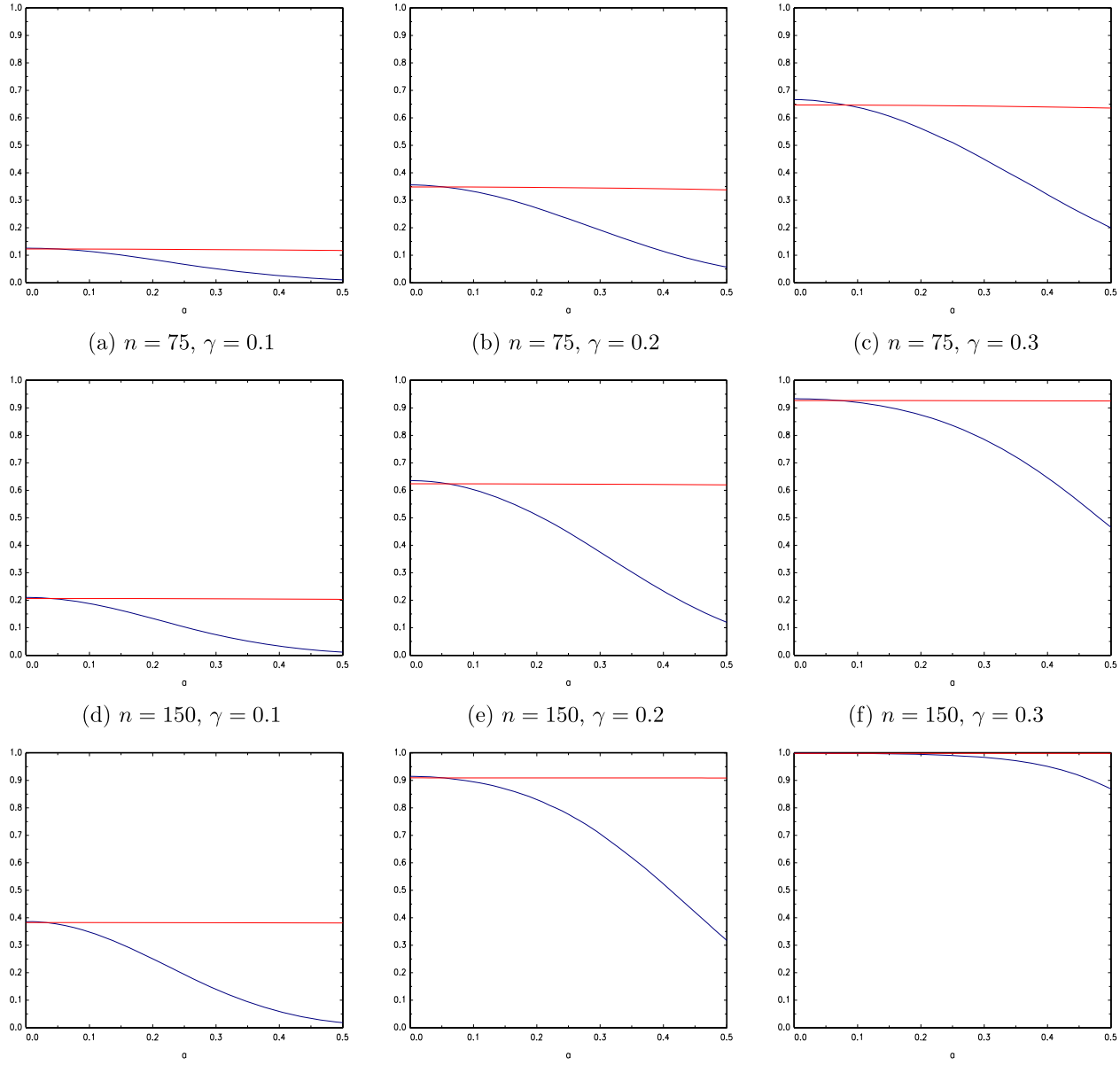
5.2. Finite Sample Power

To compare finite sample powers we focus on the uncorrelated, homoscedastic case where $\nu_t = \varepsilon_t$ in (6) with ε_t being IID $N(0,1)$. We first generate alternatives according to $m_t = aS_t^0 + \gamma$, so power is controlled by an additive offset γ to the null function aS_t^0 ($\gamma = 0$ representing size), with the mean changing smoothly from $\gamma - a$ to $\gamma + a$ (with transition mid-point 0.5). Here, $\int_0^1 m(x)\mathrm{d}x = \gamma$ and, as in the size simulations, $V_m = 0.867a^2$. In Figure 1, for each value in $\gamma \in \{0.1, 0.2, 0.3\}$ we plot power across $a \in \{0, 0.01, 0.02, \ldots, 0.5\}$. Figure 1(a)–(c) consider $n = 75$. Here we see that for each value of γ, the powers of DM and DM' are very similar when $a = 0$, but the powers of DM are diminishing as a (and hence, V_m) increases away from zero. Conversely, the powers of DM' stay pretty much constant across all V_m. Consequently, we observe quite marked differences in power levels between DM and DM' when $V_m > 0$, increasingly so as the magnitude of V_m increases. The same qualitative pattern of results is seen for $n = 150$ in Figure 1(d)–(f) and $n = 300$ in Figure 1(g)–(i), albeit at higher levels as the powers of both tests increase in n for a given value of γ.

In Figure 2 we consider the power behavior of DM and DM' from a perhaps more traditional perspective. Here, for each value in $a \in \{0, 0.25, 0.5\}$ we plot power across $\gamma \in \{0, 0.01, 0.02, \ldots, 0.5\}$. When $a = 0$ we see that the size and powers of DM and DM' are very similar throughout for each sample size. Hence. despite the fact that in this case, for each value of γ, $\hat{m}_t$ is estimating a nonexistent time variation in m_t, the cost in terms of power is practically zero. When $a = 0.25$, time variation in m_t is now present, and we observe the size and power of DM falling slightly below that of DM'. These differences become much more exaggerated when $a = 0.5$. We observe, therefore, that while the powers of both DM and DM' are monotonically increasing in γ, DM' can provide quite substantial power gains over DM when the degree of time variation in m_t is large.

Table 2. Finite sample size of nominal 0.05-level tests: $m_t = aS_t^0, \nu_t = \sigma_t\varepsilon_t, \sigma_t = S_t(c, 30, 1, \delta_2^*), n = 150$.

			DM					*DM′*					
		a :	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5
c	δ_2^*	V_m :	0.000	0.009	0.035	0.078	0.139	0.217	0.000	0.009	0.035	0.078	0.139	0.217
0.25	1/6		0.026	0.021	0.005	0.000	0.000	0.000	0.079	0.079	0.078	0.076	0.074	0.071
	1/3		0.039	0.025	0.008	0.001	0.000	0.000	0.064	0.064	0.063	0.062	0.061	0.058
	3		0.048	0.046	0.043	0.038	0.032	0.026	0.053	0.053	0.053	0.053	0.053	0.053
	6		0.046	0.046	0.045	0.043	0.042	0.039	0.054	0.054	0.054	0.054	0.054	0.054
0.50	1/6		0.040	0.035	0.022	0.008	0.002	0.000	0.063	0.063	0.063	0.062	0.061	0.060
	1/3		0.043	0.035	0.020	0.008	0.002	0.000	0.059	0.058	0.058	0.057	0.057	0.056
	3		0.044	0.042	0.039	0.035	0.030	0.025	0.057	0.057	0.057	0.057	0.057	0.057
	6		0.040	0.039	0.038	0.037	0.037	0.036	0.062	0.062	0.062	0.062	0.062	0.062
0.75	1/6		0.046	0.037	0.019	0.007	0.002	0.000	0.055	0.054	0.054	0.054	0.053	0.053
	1/3		0.047	0.038	0.019	0.007	0.002	0.000	0.053	0.053	0.053	0.052	0.052	0.051
	3		0.039	0.037	0.032	0.025	0.018	0.013	0.063	0.063	0.063	0.063	0.063	0.063
	6		0.026	0.026	0.025	0.025	0.024	0.023	0.078	0.078	0.078	0.078	0.078	0.078



(a) $n = 75$, $\gamma = 0.1$ (b) $n = 75$, $\gamma = 0.2$ (c) $n = 75$, $\gamma = 0.3$

(d) $n = 150$, $\gamma = 0.1$ (e) $n = 150$, $\gamma = 0.2$ (f) $n = 150$, $\gamma = 0.3$

(g) $n = 300$, $\gamma = 0.1$ (h) $n = 300$, $\gamma = 0.2$ (i) $n = 300$, $\gamma = 0.3$

Figure 1. Finite sample power of nominal 0.05-level tests, $m_t = aS_t^0 + \gamma, v_t = \varepsilon_t$. *DM*: ——; *DM′*: ——

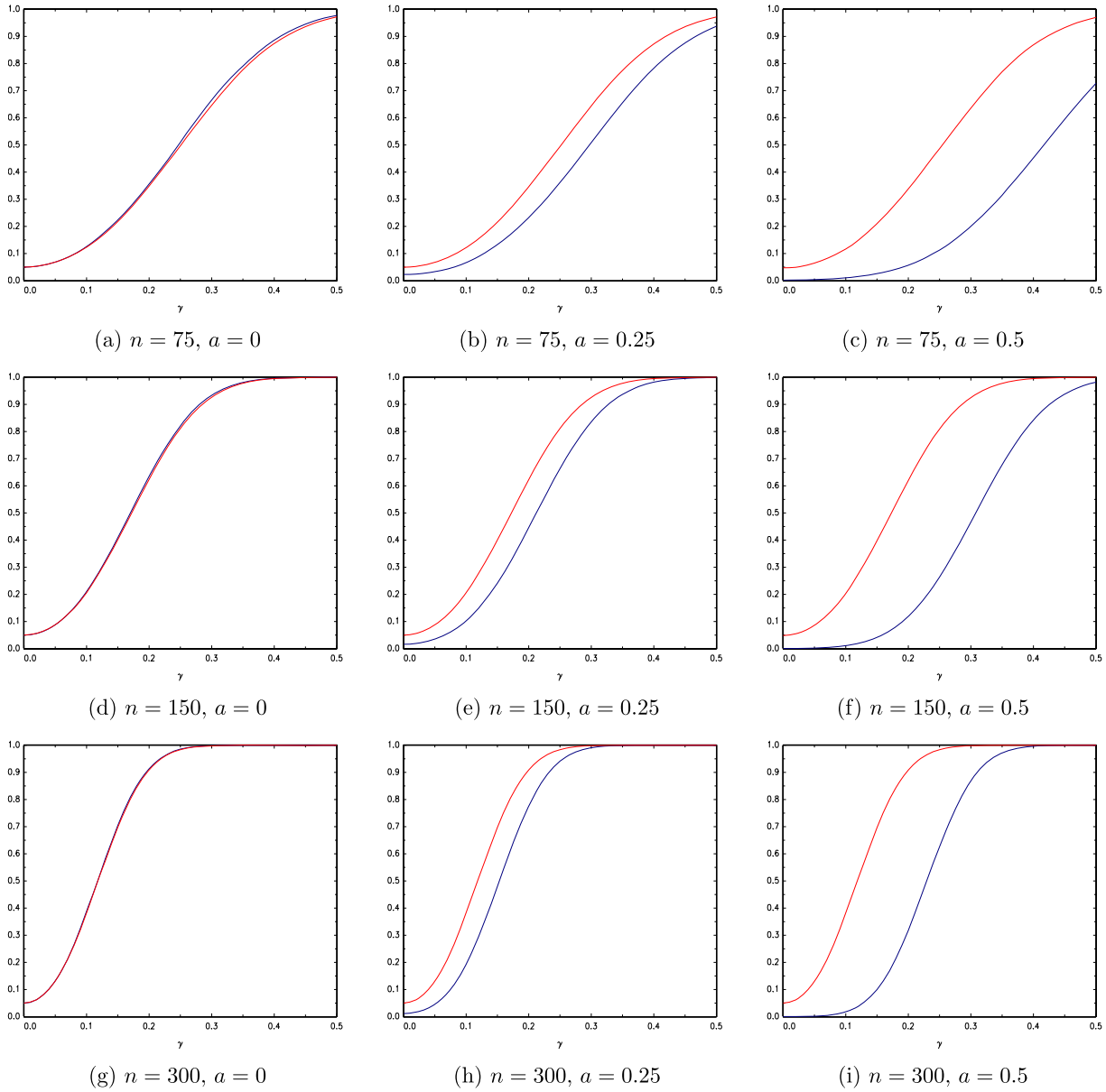
Next, we generate alternatives according to $m_t = \gamma S_t^1(c)$ with $c \in \{0.25, 0.75\}$; here, $\int_0^1 m(x)\mathrm{d}x = 0.5\gamma$ and -0.5γ for $c = 0.25$ and 0.75, respectively, while $V_m = 0.617\gamma^2$ for both values of c considered. Figure 3 gives the powers across $\gamma \in \{0, 0.02, 0.04, \ldots, 1.0\}$. Note that $\gamma = 0$ implies the null hypothesis holds, while simultaneously imposing $V_m = 0$, so both *DM* and *DM′* are correctly sized in this case due to the lack of mean variation for this setting. The power results show that it is again the case that *DM′* is much more powerful than *DM*, particularly when $n = 75$ where the power of *DM* is non-monotonic and barely rises above the nominal level. For larger n, the non-monotonicity in the power of *DM* disappears, but its power is still well below that of *DM′*. Notice that the power profiles of both tests are little affected by c, as might be expected given that $\left|\int_0^1 m(x)\mathrm{d}x\right| = 0.5$ in both cases and two-sided tests are being conducted.

6. Empirical Applications

To illustrate the potential performance differences between the *DM* and *DM′* tests in practice, we consider two forecast accuracy comparisons based on quarterly data. The first involves UK house price growth and the second US GDP growth.

6.1. Forecasting UK House Price Growth

Here we evaluate the accuracy of forecasts of house price growth from a distributed lag (DL) model involving growth of the rent-price ratio, relative to a benchmark autoregressive (AR) model using direct q-step ahead forecasts. Specifically, we wish to forecast quarter-over-quarter house price growth at target date $t+q$, $p_{t+q} = \Delta \ln(P_{t+q})$ where P_t is the price of housing at time t. Denoting the rent-price ratio as R_t and letting $r_t = \Delta \ln(R_t)$ we fit fourth order DL and AR models



(a) $n = 75$, $a = 0$ (b) $n = 75$, $a = 0.25$ (c) $n = 75$, $a = 0.5$

(d) $n = 150$, $a = 0$ (e) $n = 150$, $a = 0.25$ (f) $n = 150$, $a = 0.5$

(g) $n = 300$, $a = 0$ (h) $n = 300$, $a = 0.25$ (i) $n = 300$, $a = 0.5$

Figure 2. Finite sample power of nominal 0.05-level tests, $m_t = aS_t^0 + \gamma$, $v_t = \varepsilon_t$. *DM*: ——; *DM′*: ——

$$p_s = \hat{\beta}_0 + \hat{\beta}_1 r_{s-q} + \hat{\beta}_2 r_{s-q-1} + \hat{\beta}_3 r_{s-q-2} + \hat{\beta}_4 r_{s-q-3} + error,$$
$$p_s = \hat{\phi}_0 + \hat{\phi}_1 p_{s-q} + \hat{\phi}_2 p_{s-q-1} + \hat{\phi}_3 p_{s-q-2} + \hat{\phi}_4 p_{s-q-3} + error$$

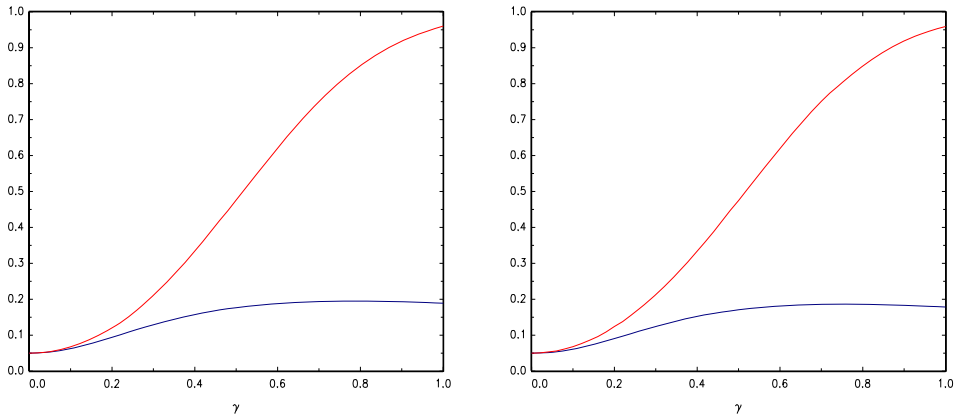
via OLS over rolling windows of N observations $s = t - N + 1, \ldots, t$. We then obtain direct forecasts of p_{t+q} according to

$$f_{1t,q} = \hat{\beta}_0 + \hat{\beta}_1 r_{t-q} + \hat{\beta}_2 r_{t-q-1} + \hat{\beta}_3 r_{t-q-2} + \hat{\beta}_4 r_{t-q-3},$$
$$f_{2t,q} = \hat{\phi}_0 + \hat{\phi}_1 p_{t-q} + \hat{\phi}_2 p_{t-q-1} + \hat{\phi}_3 p_{t-q-2} + \hat{\phi}_4 p_{t-q-3}$$

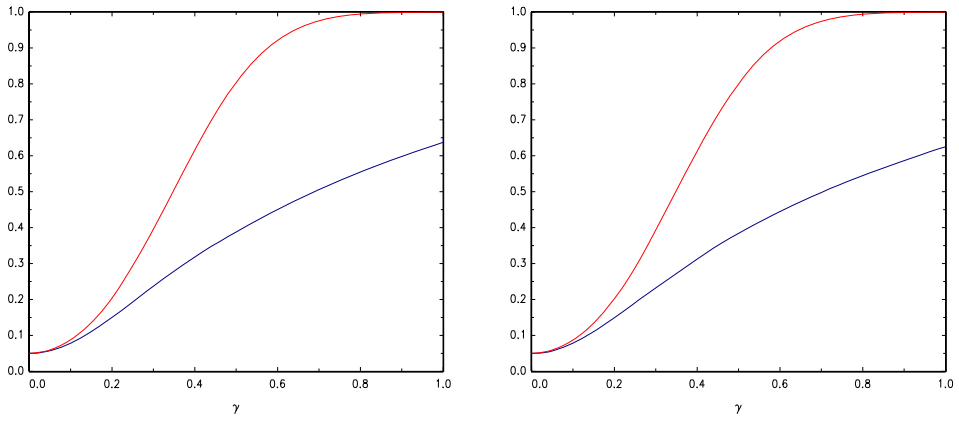
and construct the two sets of forecast errors $e_{1t,q} = p_{t+q} - f_{1t,q}$ and $e_{2t,q} = p_{t+q} - f_{2t,q}$. The (log) rent-price ratio R_t has been studied extensively as a predictor of house price growth (see Ghysels et al. 2013, and the references therein). Here we use the difference of the (log) rent-price ratio as the predictor to avoid potential biased estimation problems when the rent-price ratio exhibits substantial autocorrelation and endogeneity is present.

We compare the accuracy of the two sets of forecasts using both mean absolute error (MAE) and MSE loss functions; that is, we construct the loss differentials $d_t = |e_{1t}| - |e_{2t}|$ and $d_t = e_{1t}^2 - e_{2t}^2$, respectively (suppressing notational dependence on q). Quarterly residential property price and rental price indices are obtained from the OECD Data Explorer website. Using rolling windows of $N = 48$ observations, we examine forecasts $q = 1, 2, \ldots, 5$ quarters ahead for the target dates 1991:1-2023:4, giving $n = 132$. The *DM* and *DM′* statistics are computed using exactly the same kernel and bandwidth choices as used for our finite sample simulations in Section 5. The results are given in Table 3. The entries here are one-sided p-values for the statistics based on simulating the finite sample null distributions in the manner of Section 5, so that they are applicable for the exact sample size and kernel and bandwidth settings employed. Also reported is $\hat{V}_m/\hat{\Omega}'$, with $\hat{V}_m$ as defined in (5), which provides a standardized measure of the mean function variation.

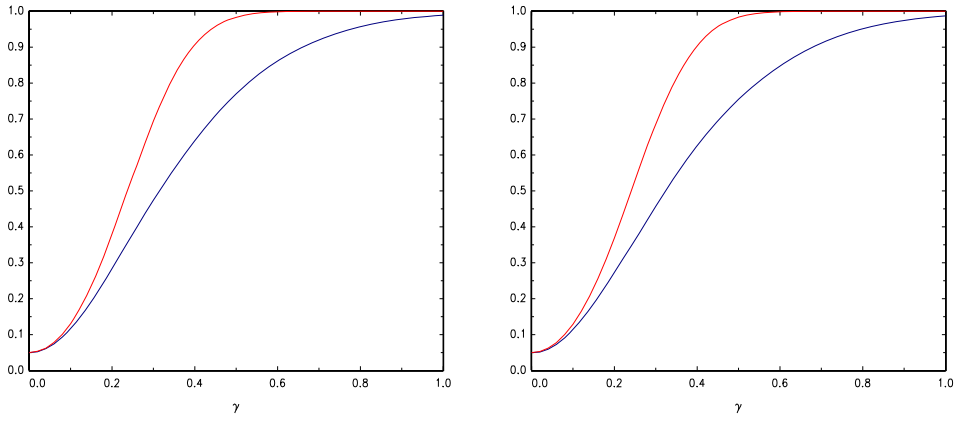
Considering first the MAE loss results, for $q = 1$, while *DM* finds no evidence whatsoever against the null of equal average forecast accuracy, we see strong evidence of the superiority of



(a) $n = 75$, $c = 0.25$ (b) $n = 75$, $c = 0.75$



(c) $n = 150$, $c = 0.25$ (d) $n = 150$, $c = 0.75$



(e) $n = 300$, $c = 0.25$ (f) $n = 300$, $c = 0.75$

Figure 3. Finite sample power of nominal 0.05-level tests, $m_t = \gamma s_t^1(c)$, $v_t = \varepsilon_t$. *DM*: ——; *DM*′: ——

the benchmark AR forecast when DM' is considered. For $q = 2, 3, 4$ both DM and DM' find evidence of superiority of the DL forecast, although the strength of rejection is stronger in each case for DM' than for DM, especially for $q = 3$ where a rejection is obtained at the 0.05-level as opposed to the 0.10-level. For $q = 5$, evidence of superiority of the DL forecast is provided by DM' while DM does not reject the equal average forecast accuracy null. Examining the value of the mean variation measure $\hat{V}_m/\hat{\Omega}'$, there is a clear pattern between the magnitude of mean variation and the differences in inference associated with the two tests. When $q = 1$ or $q = 5$, DM' returns a rejection of the null but DM does not, and in these cases we observe the largest values of $\hat{V}_m/\hat{\Omega}'$. When $q = 3$, both tests reject the null but DM' rejects more strongly, and here $\hat{V}_m/\hat{\Omega}'$ assumes an intermediate magnitude. When $q = 2, 4$, we find the smallest values of $\hat{V}_m/\hat{\Omega}'$, and these are associated with DM and DM' giving qualitatively the same inference. This pattern of results is in line with our theoretical and simulation results, where it is found that variation in the mean function of d_t can reduce the capability of DM to reject when the alternative of unequal

Table 3. Application of tests to quarterly UK house price growth forecasts, 1991:1-2023:4.

	MAE			MSE		
q	DM	DM'	$\hat{V}_m/\hat{\Omega}'$	DM	DM'	$\hat{V}_m/\hat{\Omega}'$
1	*0.211*	*0.036***	2.684	*0.463*	*0.440*	1.347
2	0.085*	0.078*	0.358	0.023**	0.011**	0.494
3	0.074*	0.032**	0.666	0.046**	0.014**	0.741
4	0.099*	0.074*	0.490	0.055*	0.024**	0.655
5	0.116	0.052*	0.846	0.109	0.055*	0.758

NOTE: MAE and MSE denote mean absolute error and mean squared error, respectively. Entries in italics are upper tail p-values, non-italicized entries are lower tail p-values. *, **, and *** denote rejection of the null of equal average forecast accuracy at the 0.10-, 0.05-, and 0.01-level, respectively.

Table 4. Application of tests to quarterly US real output growth forecasts, 1982:1-2019:4.

	MAE			MSE		
q	DM	DM'	$\hat{V}_m/\hat{\Omega}'$	DM	DM'	$\hat{V}_m/\hat{\Omega}'$
1	0.020**	0.024**	0.233	0.028**	0.029**	0.253
2	0.085*	0.087*	0.241	0.033**	0.027**	0.306
3	0.202	0.206	0.247	0.106	0.109	0.244
4	0.079*	0.076*	0.276	0.061*	0.056*	0.302
5	0.040**	0.016**	0.579	0.065*	0.008***	1.177

NOTE: MAE and MSE denote mean absolute error and mean squared error, respectively. Entries are lower tail p-values. *, **, and *** denote rejection of the null of equal average forecast accuracy at the 0.10-, 0.05-, and 0.01-level, respectively.

average forecast accuracy is true, while DM' remains robust to such mean variation.

Under MSE loss, neither test rejects the null with $q = 1$ and both tests suggest the DL forecast is superior for $q = 2, 3, 4$. Only for $q = 5$ do we see the two tests provide differing inference; here DM' is alone in suggesting superiority of the DL forecast. It is noteworthy that this coincides with the largest value of $\hat{V}_m/\hat{\Omega}'$ for which a rejection by either test is obtained, again suggesting the mean variation in d_t is compromising the power of DM relative to DM'.

In Figures S2 and S3 of the supplementary appendix, plots of d_t and $\hat{m}_t$ are provided under both MAE and MSE. In the case of MAE, for $q = 1$ the mean function appears positive early in the sample, then close to zero for the majority of the remaining time period, before dropping to a negative value close to the sample end. In contrast, for $q = 2, 3, 4, 5$, the mean function is noticeably negative for about the first third of the sample, then close to zero, before becoming negative again toward the end of the period. In all cases, therefore, it appears that the two sets of forecasts have approximately equal accuracy in the central part of the time period, but differences in forecast ranking emerge in the earlier and later portions of the data, which in most cases translate to a significant difference in average forecast accuracy over the period according to DM'. Broadly similar comments apply in the case of MSE.

6.2. Forecasting US GDP Growth

We evaluate the accuracy of median consensus forecasts from the Survey of Professional Forecasters (SPF) relative to a benchmark autoregressive model. The SPF was introduced by the American Statistical Association and the National Bureau of Economic Research and is currently maintained by the Federal Reserve Bank of Philadelphia; forecast data is published on a quarterly basis and is available from their website. The target variable for our forecast evaluation exercise is US real GNP/GDP, with the forecasts being for quarter-over-quarter growth rates, expressed in annualised percentage points. The actual values we use are the Bureau of Economic Analysis first revised values (one quarter after the initial release). Benchmark forecasts with which we compare those from the SPF are also reported by the Federal Reserve Bank of Philadelphia and the one we select is again an AR model using direct q-step ahead forecasts; these are calculated from estimated autoregressive models using AIC model selection for a rolling window of 60 observations. The data is available from *https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/error-statistics*, and Stark (2010) provides detailed information on the forecasts and benchmark methods. Now e_{1t} denotes the SPF forecast error and e_{2t} denotes the corresponding error from the AR benchmark forecast. We construct the loss differentials as in the previous section for $q = 1, 2, \ldots, 5$ quarters ahead for the target dates 1982:1-2019:4, giving $n = 152$ (we end our sample before start of the Covid epidemic; growth forecasts being unreliable during this period). The DM and DM' statistics are also constructed using the same settings as in the previous sub-section. The results are shown in Table 4.

Under MAE loss the inference from DM and DM' is very similar throughout: both suggest the SPF forecasts are more accurate than the AR benchmark forecasts for all horizons except $q = 3$, where neither test rejects the null of equal average forecast accuracy. The same pattern of results is found when using MSE loss, with the exception of $q = 5$ where a considerably stronger rejection of the null is obtained by DM' than by DM. Interestingly, examining the values of $\hat{V}_m/\hat{\Omega}'$, we find that there is generally considerably less mean variation than in the house price growth examples of the previous section. It is perhaps not surprising, therefore, that the DM and DM' tests return very similar inference for these data. It is also noteworthy that the largest value of $\hat{V}_m/\hat{\Omega}'$ is obtained in the case of MSE with $q = 5$, which is the one occasion where DM rejects significantly less strongly than DM', again illustrating that relatively high levels of mean variation can compromise the power of DM to reject the null of equal average accuracy, in contrast to DM'.

Finally, in Figures S4 and S5 of the supplementary appendix, we again plot d_t and $\hat{m}_t$. The mean function paths in the MAE case display more oscillating behavior than in the house price growth example, but the majority of each mean path is clearly below zero, in line with the lower tail rejections obtained by the tests. An exception to this is $q = 3$ where the mean path appears to oscillate about zero; unsurprisingly, neither test rejects in this case. For MSE loss, the patterns of the mean paths are broadly similar to those for MAE loss, but with the movements less exaggerated.

7. Conclusion

We have considered testing the null of equal average forecast accuracy in a model where the loss differential series has a potentially nonconstant mean function over time. We have shown that the standard DM test has an asymptotic size of zero under the

null of equal average forecast accuracy when time variation is present in the loss differential mean. We have also shown that, although remaining consistent, the local power of the DM test can be impacted in the time varying mean case. The source of the size and power problems in the DM statistic is inconsistency of the usual LRV estimator. We proposed a modified test statistic that replaces the standard LRV estimator (that uses full-sample demeaning) with one based on local demeaning of the loss differential series, employing a kernel smoothing nonparametric estimator for the time varying mean. We have demonstrated that the new LRV estimator is consistent for the true LRV under both the null and alternative, regardless of whether the mean function is constant or time varying. This results in the modified DM test being asymptotically correctly sized, and we have shown that it achieves nontrivial power against local alternatives with a standard $n^{-1/2}$ rate, again under both a constant or time varying mean function. The new test therefore provides a robust approach to testing the equal average forecast accuracy null, allowing for instability in the loss differential mean. Monte Carlo simulations attested to the benefits of the new approach in finite samples. The modified test behaves similarly to the original DM test when the mean function is constant, while offering valuable power gains in the time varying case. Empirical illustrations further highlight the potential for the new test to provide some improved detectability of the alternative hypothesis in practical forecast evaluation exercises. Overall we consider that extension of the equal accuracy hypothesis to the more flexible specification of an equal average accuracy hypothesis provides a valuable generalization to forecast evaluation techniques. Within this generalized setting, our modification to the DM test robustifies it to variation in the mean function, providing a useful addition to the set of evaluation tools available to practitioners.

Supplementary Materials

The supplementary appendix contains proofs of the theorems, detail on the method for simulating finite sample critical values, and the additional simulation and empirical application figures referred to in Sections 5 and 6.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

Altissimo, F., and Corradi, V. (2003), "Strong Rules for Detecting the Number of Breaks in a Time Series," *Journal of Econometrics*, 117, 207–244. [6]

Andersen, T. G. (1994), "Stochastic Autoregressive Volatility: A Framework for Volatility Modeling," *Mathematical Finance*, 4, 75–102. [2]

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858. [4,7]

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327. [2]

Carrasco, M., and Chen, X. (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39. [2]

Chan, K. W. (2022), "Optimal Difference-based Variance Estimators in Time Series: A General Framework," *The Annals of Statistics*, 50, 1376–1400. [6]

Clark, T. E., and McCracken, M. (2015), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," *Journal of Econometrics*, 186, 160–177. [2]

De Jong, R. M., and Davidson, J. (2000), "Consistency of Kernel Estimators of Heteroscedastic and Autocorrelated Covariance Matrices," *Econometrica*, 68, 407–423. [4]

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263. [1]

Engle, R. F. (1982), "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1007. [2]

Ghysels, E., Plazzi, A., Valkanov, R., and Torous, W. (2013), "Forecasting Real Estate House Prices," in *Handbook of Economic Forecasting* (Vol. 2, Part A), eds. G. Elliott and A. Timmermann, pp. 509–580. Amsterdam: Elsevier. [11]

Giacomini, R., and Rossi, B. (2010), "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 25, 595–620. [1,3]

Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578. [1]

Hall, A. R. (2005), *Generalized Method of Moments*, Oxford: Oxford University Press. [4]

Harvey, D. I., Leybourne, S. J., and Zu, Y. (2024), "Tests for Equal Forecast Accuracy Under Heteroskedasticity," *Journal of Applied Econometrics*, 39, 850–869. [3]

Juhl, T., and Xiao, Z. (2009), "Tests for Changing Mean with Monotonic Power," *Journal of Econometrics*, 148, 14–24. [6]

Lazarus, E., Lewis, D. J., and Stock, J. H. (2021), "The Size-Power Tradeoff in HAR Inference," *Econometrica*, 89, 2497–2516. [7]

Martins, L. F., and Perron, P. (2016), "Improved Tests for Forecast Comparisons in the Presence of Instabilities," *Journal of Time Series Analysis*, 37, 650–659. [1]

Odendahl, F., Rossi, B., and Sekhposyan, T. (2023), "Evaluating Forecast Performance with State Dependence," *Journal of Econometrics*, 237, 105220. [1]

Perron, P., and Yamamoto, Y. (2021), "Testing for Changes in Forecasting Performance," *Journal of Business and Economic Statistics*, 39, 148–165. [1]

Richter, S., and Smetanina, E. (2020), "Forecast Evaluation and Selection in Unstable Environments," Discussion Paper, University of Chicago. [4]

Stark, T. (2010), "Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters," Special Report, Federal Reserve Bank of Philadelphia. [13]

Stock, J. H., and Watson, M. W. (2003), "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 41, 788–829. [1]

Sun, Y., Phillips, P. C. B., and Jin, S. (2008), "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing," *Econometrica*, 76, 175–194. [7]

Wu, W. B. (2007), "Strong Invariance Principles for Dependent Random Variables," *Annals of Probability*, 35, 2294–2320. [6]

Zhu, Y., and Timmermann, A. (2020), "Can Two Forecasts have the Same Conditional Expected Accuracy?" Discussion Paper, arXiv:2006.03238. [4]

——— (2022), "Conditional Rotation between Forecasting Models," *Journal of Econometrics*, 231, 329–347. [4]