Granger Centre Discussion Paper Series

Forecast evaluation tests and negative long-run variance estimates in small samples

by

David I. Harvey, Stephen J. Leybourne and
Emily J. Whitehouse

Granger Centre Discussion Paper No. 17/03

Forecast Evaluation Tests and Negative Long-run Variance Estimates in Small Samples*

David I. Harvey, Stephen J. Leybourne and Emily J. Whitehouse

School of Economics, University of Nottingham

May 3, 2017

Abstract

In this paper, we show that when computing standard Diebold-Mariano-type tests for equal forecast accuracy and forecast encompassing, the long-run variance can frequently be negative when dealing with multi-step-ahead predictions in small, but empirically relevant, sample sizes. We subsequently consider a number of alternative approaches to dealing with this problem, including direct inference in the problem cases and use of long-run variance estimators that guarantee positivity. The finite sample size and power of the different approaches are evaluated using extensive Monte Carlo simulation exercises. Overall, for multi-step-ahead forecasts, we find that the recently proposed Coroneo and Iacone (2016) test, which is based on a weighted periodogram long-run variance estimator, offers the best finite sample size and power performance.

Keywords: Forecast evaluation; Long-run variance estimation; Simulation; Diebold-Mariano test; Forecasting.

JEL Classification: C12, C22, C53.

*We are very grateful to the Editor, Michael McCracken, and an anonymous referee for their very helpful and constructive comments on an earlier version of this paper. Correspondence to: David Harvey, School of Economics, University of Nottingham, University Park, Nottingham, NG7 2RD, UK. Email: dave.harvey@nottingham.ac.uk

1 Introduction

Given the critical role that forecasting plays in economic and financial research and policy-making, the evaluation of competing forecasts of the same outcomes has become an extensive and prominent field in the econometric and empirical economic literatures. Within this field, the most common forecast evaluation exercise typically undertaken is to compare the accuracy of two or more sets of forecasts on the basis of some measure of loss associated with the forecast errors, such as mean squared forecast error. In a key contribution to the literature, Diebold and Mariano (1995) [DM] proposed an approach for testing equal forecast accuracy valid for potentially contemporaneously correlated, serially correlated and non-normal forecast errors, based on testing for a zero mean in a series defined as the difference between the two forecasts' error loss functions (the "loss differential"). Harvey *et al.* (1997) [HLNa] suggested two finite sample modifications to the DM statistic to improve size control in small samples, based on a finite sample bias correction to the test statistic, and using Student's t critical values rather than those from a standard normal. Application of the DM test or its HLNa variant have now become prevalent in empirical forecasting research, to the extent that it is now routine for the results of such forecast accuracy tests to be reported alongside any forecast comparisons.

Testing for equal forecast accuracy is just one approach to evaluating the predictive ability of rival forecasts. A second popular evaluation method is to test for whether one set of forecasts encompasses another, in the sense that the encompassed forecasts do not result in a reduction in forecast accuracy when used in combination with the encompassing set of forecasts. Harvey *et al.* (1998) [HLNb] proposed a forecast encompassing test based on a DM-type approach, where the loss differential is redefined to permit testing an encompassing null hypothesis, and the approach has become standard in cases where one abstracts from model parameter estimation uncertainty.

In this paper we focus on the behaviour of the DM/HLNa tests based on squared error loss, and the HLNb test for encompassing, in *small samples*. Our work is therefore in a similar vein to that of Ashley (2003) and Ashley and Tsang (2014) who investigate out-of-sample inference with limited data availability. The DM test statistic is fundamentally comprised of the loss differential series sample mean standardised by an estimate of the long-run variance. DM make use of the fact that optimal h-steps-ahead forecasts are at most $(h-1)$-order dependent to advocate use of a rectangular kernel in the long-run variance estimator which truncates at lag $h-1$. While this approach results in decent finite sample size and power properties for many sample size and h settings, the long-run variance estimator is not guaranteed to be positive whenever $h > 1$. DM note this possibility, but suggest that such an outcome would be rare; similarly, Clark (1999) finds a low occurrence of negative long-run variance estimates in his equal accuracy test simulations (always less than 3% of replications). However, these observations were made on the basis of results that considered predictions only up to two steps ahead. Our first contribution is to highlight that the prevalence of negative long-run variance estimates can

be much greater in small samples when longer horizon forecasts are considered. For example, when testing equal mean squared forecast error with $h = 6$, we find that negative variance estimates arise approximately 20% of the time for a sample size of 16, rising to over 40% of the time for a sample size of 8.

In practical applications, often due to the limitations of economic or forecast data, it is not uncommon for forecast evaluation to be conducted using sample size and forecast horizon settings that lie in the region where negative variance estimates occur frequently. For example, in the context of testing for equal forecast accuracy, the recent papers Caporale and Gil-Alana (2014), Dreger and Wolters (2014), Dib *et al.* (2008) and Qin *et al.* (2008) all implement the DM/HLNa tests in forecast samples smaller than 25 observations with horizons of $h = 6$ or greater. Further, Mehl (2009) and Chow and Choy (2006) have reported finding negative DM/HLNa long-run variance estimates when using samples of 18 and 24 forecasts at horizons of 6 and 5-6, respectively.

Given that negative variance estimates can arise frequently in situations of practical relevance, it is important to determine the best approach to deliver a reliable testing procedure in terms of small sample size and power properties. DM suggest treating a negative variance estimate as a zero, thereby automatically rejecting the null against a two-sided alternative in such cases. However, given the low occurrence of negative variance estimates in their simulations, the size implications of such an approach are not fully explored. In the simulation work of Clark (1999), the relatively few replications where negative variance estimates were obtained were excluded from the simulations, thereby abstracting from the effects of dealing with certain problematic cases. HLNa and HLNb simulated combinations of sample size and forecast horizon where negative variances can occur frequently, but their simulations failed to correctly deal with negative variance estimates, thus again the impact of negativity in the variance estimate is not clear. Of course, other long-run variance estimators exist which ensure non-negativity; for example, DM, Clark (1999) and others discuss the possible use of the Bartlett kernel, and in a recent paper on testing equal forecast accuracy, Coroneo and Iacone (2016) recommend use of the nonparametric periodogram estimator of Hualde and Iacone (2017), combined with use of bandwidth-dependent critical values. The second contribution of this paper is therefore to formally assess the behaviour of different strategies for dealing with the potential problem of negative long-run variance estimation in tests of equal accuracy and encompassing. We conduct an extensive set of Monte Carlo simulations to establish the small sample size and power properties of different approaches. Broadly, we find that for multi-step-ahead forecasts, the Coroneo and Iacone (2016) approach outperforms other methods, and the attractive finite sample properties reported in their paper for moderate sample sizes and forecast horizons extends to the small sample and longer horizon region under focus in this paper for both equal accuracy and encompassing tests, where the DM/HLNa and HLNb tests can suffer from negative variance estimates.

The outline of the paper is as follows. In section 2, we briefly outline the DM, HLNa

and HLNb tests for equal mean squared forecast error and forecast encompassing. Section 3 highlights by simulation the frequency with which negative long-run variance estimates can arise for different sample sizes and forecast horizons. In section 4, a number of ways of dealing with these cases are considered, including alternative long-run variance estimators that are guaranteed to be positive, and section 5 investigates the performance of these procedures using finite sample size and power simulations. In section 6 we conduct a related set of simulation experiments using a DGP calibrated to the empirical work of Dreger and Wolters (2014), while section 7 considers simulations for the case where forecasts are obtained from estimated models. Section 8 concludes.

2 Standard tests for equal accuracy and encompassing

Consider first the issue of evaluating whether two competing sets of forecasts are equally accurate according to some loss function-based accuracy measure, or whether one forecast outperforms the other in terms of that metric. Denote the actuals by y_t and the competing forecasts by f_{1t} and f_{2t}, $t = 1, ..., T$, and consider a given loss function $L(.)$ that depends on the forecast errors, so that the cost of error associated with the forecast f_{it} is $L(e_{it})$, $i = 1, 2$. Now define the loss differential series

$$d_t = L\left(e_{1t}\right) - L\left(e_{2t}\right), \quad t = 1, ..., T.$$

The null hypothesis of equal forecast accuracy, according to the specified loss function $L(.)$, can then be expressed as

$$H_0 : E(d_t) = 0.$$

For example, under squared error loss, $d_t = e_{1t}^2 - e_{2t}^2$ and the null hypothesis entails the equality of population mean squared forecast errors.

Under the assumptions that d_t is covariance stationary and short memory, DM propose a test of H_0 based on the asymptotic distribution of the sample mean loss differential

$$\sqrt{T}(\bar{d} - E(d_t)) \xrightarrow{d} N\left(0, \omega^2\right)$$

where $\bar{d} = T^{-1} \sum_{t=1}^{T} d_t$ and ω^2 denotes the long-run variance of d_t, i.e. $\omega^2 = \sum_{j=-\infty}^{\infty} \gamma_j$ with $\gamma_j = Cov(d_t, d_{t-j})$. Denoting a consistent long-run variance estimator by $\hat{\omega}^2$, the DM test statistic is then given by

$$DM = \sqrt{T}\left(\frac{\bar{d}}{\hat{\omega}}\right)$$

which has an asymptotic standard normal distribution under the null. DM suggest use of a long-run variance estimator comprised of a weighted sum of sample autocovariances, and, motivated by the fact that optimal h-steps-ahead forecast errors are at most $(h-1)$-dependent,

they advocate using a rectangular kernel truncated at lag $h-1$, i.e.

$$\hat{\omega}^2 = \hat{\gamma}_0 + 2\sum_{j=1}^{h-1}\hat{\gamma}_j \tag{1}$$

where $\hat{\gamma}_j = T^{-1}\sum_{t=j+1}^{T}\left(d_t - \bar{d}\right)\left(d_{t-j} - \bar{d}\right)$.[1] HLNa propose a modification of the DM statistic designed to improve the small sample size behaviour of the test. Their statistic is based on an approximate bias correction to the long-run variance estimator, and can be written as

$$MDM = \sqrt{T + 1 - 2h + T^{-1}h(h-1)}\left(\frac{\bar{d}}{\hat{\omega}}\right). \tag{2}$$

These authors also suggest use of t_{T-1} critical values in place of those from the standard normal, again to better control small sample size.

Next consider investigating whether one set of forecasts encompasses another, in that the accuracy of one set of (encompassing) forecasts f_{1t} cannot be improved through linear combination with a second set of (encompassed) forecasts f_{2t}. HLNb develop a test for forecast encompassing based on the Bates and Granger (1969) forecast combination scheme, where the combination weights sum to one.[2] Denoting the combined forecast by f_{ct}, the combination is

$$f_{ct} = (1-\delta)f_{1t} + \delta f_{2t}$$

where δ $(0 \leq \delta \leq 1)$ determines the weights associated with the constituent forecasts. In this context, forecast f_{1t} encompasses forecast f_{2t} if the optimal mean squared error-minimising combination weight

$$\delta_{opt} = \frac{E(e_{1t}^2) - E(e_{1t}e_{2t})}{E(e_{1t}^2) + E(e_{2t}^2) - 2E(e_{1t}e_{2t})}$$

is equal to zero. The null of forecast encompassing can then be expressed in a DM-type form:

$$H_0 : E(d_t) = 0$$

with d_t in this case defined as

$$d_t = e_{1t}(e_{1t} - e_{2t}). \tag{3}$$

HLNb therefore propose applying the DM approach to this testing problem, along with the HLNa bias correction and use of t_{T-1} critical values. The test statistic is then (2) but with d_t given by (3). The test is conducted against the one-sided alternative $E(d_t) > 0$ (i.e. $\delta > 0$), given the assumption of a non-negative combination weight.

[1] Note that ARCH-type behaviour in the forecast errors induces additional autocorrelation into d_t, requiring use of higher order lags; see Harvey, Leybourne and Newbold (1999).

[2] Extensions of the test to allow for biased forecasts and combination weights that are not constrained to sum to one are discussed in Clements and Harvey (2009).

3 Frequency of negative long-run variance estimates

The long-run variance estimator (1), based on the rectangular kernel, is not guaranteed to be positive whenever $h > 1$. In practice, of course, a negative outcome is highly problematic since the MDM statistic for testing equal accuracy or encompassing cannot be computed. In such circumstances, a practitioner must then decide how to deal with such a result; suggestions in the literature include treating the estimate as zero or using an alternative long-run variance estimator that guarantees positivity. Whatever strategy is followed will have implications for the size and power of the resulting testing procedure, so it is therefore valuable to quantify how frequently negative long-run variance estimates are likely to be encountered in practice. While DM, Clark (1999) and Coroneo and Iacone (2016), *inter alios*, note that (1) can produce a negative result, little evidence has so far been provided as to the extent of this potential problem. To shed more light on the issue, in this section we report results from Monte Carlo simulation experiments to determine the frequency with which negative long-run variance estimates arise for different sample sizes and forecast horizons, both for equal accuracy and encompassing tests.

To begin, we consider the case of testing for equal forecast accuracy, adopting a standard simulation data generating process [DGP] consistent with the work of DM, HLNa and Clark (1999). We assume mean squared error loss, so that $d_t = e_{1t}^2 - e_{2t}^2$, $t = 1, ..., T$, generating the forecast errors according to the following DGP, which allows for h-steps-ahead forecasts to follow moving average [MA] processes of order $h - 1$:

$$\begin{aligned} e_{1t} &= v_{1t} + \sum_{j=1}^{h-1} \theta_j v_{1,t-j} \\ e_{2t} &= \sqrt{R}\left(v_{2t} + \sum_{j=1}^{h-1} \theta_j v_{2,t-j} \right) \end{aligned}$$

where $[v_{1t}, v_{2t}]' \sim N(0, I_2)$, $t = 1 - (h-1), ..., T$. The ratio of the variances of the two forecast errors is given by $R > 0$, with $R = 1$ giving the null and $R \neq 1$ the alternative. Focusing on the small samples that are often employed in forecast evaluation exercises, we simulate this DGP for $T = \{8, 16, 32, 64\}$, $h = \{2, 3, 4, 5, 6\}$, and calculate the frequency with which negative values of the long-run variance estimator (1) arise. We consider three settings for the MA parameters: (i) the case of no serial correlation with $\theta_j = 0\ \forall j$, (ii) a case of moderate serial correlation with $\theta_j = 0.9/(h-1)\ \forall j$, and (iii) a case of high degree serial correlation with θ_j set to the jth element of $\theta = (0.95, 0.9, 0.8, 0.65, 0.6)$, these values being drawn from the US inflation forecast error-based DGP 1 of Clark and McCracken (2013). Here and throughout the paper, simulations are conducted using 10,000 Monte Carlo replications. Table 1 reports the results under the null ($R = 1$), and under the alternative ($R > 1$), with the settings $R = 12$, $R = 7$, $R = 3$ and $R = 2$ for $T = 8$, $T = 16$, $T = 32$ and $T = 64$, respectively (chosen to ensure that the test powers considered in section 5 are roughly comparable across sample sizes).

As might be expected, we find negative long-run variance estimates occur with a frequency

that increases with the forecast horizon, and decreases with the sample size. While the occurrence of negative estimates is rare when $T = 64$, the problem can be substantial for the smaller sample sizes considered, particularly for longer forecast horizons where the frequency can rise above 40%. In such circumstances, a practitioner would be unable to compute the standard DM or MDM test statistics almost half the time. The pattern of frequencies for negative long-run variance estimates has very little dependence on whether the simulations are conducted under the null or alternative hypotheses, and while there is a reduction in the frequency of negative estimates as the degree of serial correlation increases, the overall features of the results are similar across the different dependence settings, particularly for the longer forecast horizons. We also considered simulations where the forecast errors were contemporaneously correlated, but this had little effect on the proportion of negative long-run variances obtained. These results highlight a potentially serious issue with the implementation of standard tests for equal forecast accuracy in small samples.

Turning now to testing for forecast encompassing, we let $d_t = e_{1t}(e_{1t} - e_{2t})$, $t = 1, ..., T$, where the forecast errors are generated according to the following DGP, again allowing for MA$(h-1)$-dependence in the errors of h-steps-ahead forecasts:

$$e_{it} = v_{it} + \sum_{j=1}^{h-1} \theta_j v_{i,t-j}, \quad i = 1, 2$$

where

$$\begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix} \sim N\left(0, \begin{bmatrix} 1 & \rho \\ \rho & \kappa^2 \end{bmatrix}\right), \quad t = 1 - (h-1), ..., T$$

with $\kappa^2 > \rho^2$. The null hypothesis that forecast f_{1t} encompasses f_{2t} is obtained by setting $\rho = 1$, while a setting of $\rho < 1$ gives the alternative. Under the alternative, it can be shown that power depends only on the single parameter $k = \sqrt{\kappa^2 - \rho^2}/(1-\rho)$. Table 2 reports results for the frequency of negative long-run variance estimates using the same settings for T, h and θ_j as in Table 1. Results are reported under both the null and alternative, with the settings for k under the alternative being $k = 1.25$, $k = 2.00$, $k = 3.00$ and $k = 4.50$ for $T = 8$, $T = 16$, $T = 32$ and $T = 64$, respectively (again chosen to broadly align the test power levels considered in section 5 across sample sizes).

The pattern of negative estimates for the long-run variance is very similar in the case of testing for forecast encompassing to that for testing for equal forecast accuracy. Indeed, on comparing Tables 1 and 2 for a given combination of T, h and θ_j, it is clear that the numerical frequencies are very close to each other, suggesting that the prevalence of negative long-run variance estimates is driven more by the interplay of sample size, serial correlation and the number of estimated autocovariances $(h-1)$, rather than by the precise form of d_t. We again see a rising incidence of negative estimates as T decreases and as h increases. As with the equal accuracy results, it makes little difference whether the long-run variance is being calculated under the null or alternative, and the rejection frequencies are highest for lower degrees of

serial correlation. The overall finding is that negative long-run variance estimates can occur with very high probability for equal accuracy and encompassing tests when using multi-step-ahead forecasts with small, yet practically relevant, sample sizes.

4 Adjusted Diebold-Mariano-type tests

Given the prevalence of negative long-run variance estimates that arise for multi-step-ahead forecasts in small samples when using the standard long-run variance estimator in the DM-type tests, it is important to establish methods for dealing with this potential problem. In this section we consider a number of possible approaches, all based on the DM-type tests for equal accuracy and encompassing. The following section then evaluates their relative performance in terms of finite sample size and power.

The first approach we consider is the suggested method of DM in the equal accuracy testing context, which is to treat any occurrence of a negative long-run variance as a zero, viewing the negative estimate as indicative of a very small long-run variance. This of course implies a test statistic of $\pm\infty$, depending on the sign of the numerator $\bar{d}$. In a two-sided testing context, as in DM, such a treatment induces an immediate rejection of the null hypothesis, so a negative long-run variance estimate always indicates evidence in favour of the alternative hypothesis under this approach. When testing against a one-sided alternative, as is common in applications of equal accuracy tests and always the case when testing for encompassing, treating a negative long-run variance as zero will either induce automatic rejection or non-rejection, depending on whether the implied test statistic value of $+\infty$ or $-\infty$ lies in the relevant one-tailed critical region. Applying this approach to the MDM tests of HLNa and HLNb, we can express the method as

$$MDM_{rej} = \begin{cases} MDM & \text{if } \hat{\omega}^2 > 0 \\ \text{sign}(\bar{d}) \times \infty & \text{otherwise} \end{cases}$$

with the test statistic to be compared with t_{T-1} critical values.

Given the frequency with which negative long-run variance estimates can occur, the MDM_{rej} approach will induce substantial over-size in two-sided equal accuracy testing procedures for $h > 1$ and small T, as all occurrences of a negative $\hat{\omega}^2$ trigger a rejection of the null. A similar, albeit reduced, feature of over-size would also be expected for one-sided equal accuracy tests and tests for forecast encompassing, with rejections of the null occurring whenever a negative $\hat{\omega}^2$ coincides with the appropriate sign of $\bar{d}$. A simple conservative approach which would avoid such properties is to treat the occurrence of a negative long-run variance estimate as a failure to correctly estimate the true long-run variance, and default to *non-rejection* of the null in such instances. One way of writing such a method would be to define the adjusted test statistic as

$$MDM_{non} = \begin{cases} MDM & \text{if } \hat{\omega}^2 > 0 \\ 0 & \text{otherwise} \end{cases}$$

with the test statistic again being compared with t_{T-1} critical values. A potential down-side of this approach is that the greater size control afforded by treating negative estimate cases as non-rejections is also likely to be associated with low power under the respective test alternative.

Another simple approach is to deal with a negative long-run variance estimate by replacing it with the corresponding short-run variance estimate $\hat{\gamma}_0$, thereby reducing the bandwidth in (1) from $h-1$ to zero. While this approach neglects the impact of autocorrelation terms, it can be argued that the very presence of a negative estimate indicates that estimation of such components is highly unreliable in these situations. When the short-run variance estimator is used, the appropriate bias correction in the MDM statistic is that for $h=1$, i.e.

$$MDM_0 = \sqrt{T-1}\left(\frac{\bar{d}}{\sqrt{\hat{\gamma}_0}}\right)$$

and the overall test statistic that adopts this statistic when a negative long-run variance is encountered can be written as

$$MDM_{SR} = \begin{cases} MDM & \text{if } \hat{\omega}^2 > 0 \\ MDM_0 & \text{otherwise} \end{cases}.$$

Critical values from the t_{T-1} distribution are again to be used.

While the above methods replace negative long-run variances with simple decision rules or a short-run variance estimate, the next two approaches we consider retain a proper estimate of the long-run variance, but make use of estimators that impose positivity. An obvious possibility in this class is to replace the rectangular kernel in (1) with the Bartlett kernel, i.e.

$$\hat{\omega}^2_{Bart} = \hat{\gamma}_0 + 2\sum_{j=1}^{m}\left(1-\frac{j}{m+1}\right)\hat{\gamma}_j$$

where m denotes the bandwidth. Clark (1999) considered such an approach with Newey-West and pre-whitened Newey-West bandwidth selection. While Clark's simulations abstracted from issues of negative variance estimation, it was found that a Bartlett-based approach could result in greater finite sample over-size than when using the rectangular kernel, hence it would not be recommended to use the Bartlett kernel in all circumstances, particularly when the rectangular kernel does not have negative variance estimate problems. Here, we consider a hybrid approach, whereby the standard MDM test is used provided the long-run variance estimate is positive, but in the case of a negative estimate, the statistic switches to one based on the Bartlett kernel. For consistency with the optimal forecast-motivated choice of truncation $h-1$ in (1), along with the fact that use of the Bartlett kernel is most likely to arise in small samples, we set the Bartlett bandwidth to $m = h-1$. As the HLNa bias correction does not apply to $\hat{\omega}^2_{Bart}$ (and an equivalent bias correction is not possible to obtain without effectively reducing $\hat{\omega}^2_{Bart}$ to $\hat{\omega}^2$), we define the DM statistic that uses the Bartlett long-run variance estimator as

$$DM_{Bart} = \sqrt{T}\left(\frac{\bar{d}}{\hat{\omega}_{Bart}}\right).$$

We can then write the third testing approach as

$$MDM_B = \begin{cases} MDM & \text{if } \hat{\omega}^2 > 0 \\ DM_{Bart} & \text{otherwise} \end{cases}$$

with t_{T-1} critical values employed as before.

The original DM test, and the variants outlined above, all make use of weighted sample autocovariances in the long-run variance estimator. An alternative approach proposed by Coroneo and Iacone (2016) is to use a weighted periodogram estimator, and these authors recommend construction of a DM-type test using the estimator of Hualde and Iacone (2017). Denoting the periodogram of d_t for Fourier frequency $\lambda_j = 2\pi j/T$ by

$$I(\lambda_j) = \left| \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} d_t e^{-i\lambda_j t} \right|^2$$

with i the imaginary unit, they suggest use of the Daniell kernel with bandwidth m to construct the weighted periodogram estimator of the long-run variance

$$\hat{\omega}^2_{Dan} = 2\pi \frac{1}{m} \sum_{j=1}^{m} I(\lambda_j)$$

which is then used to construct the DM-type test statistic

$$DM_{CI} = \sqrt{T} \left(\frac{\bar{d}}{\hat{\omega}_{Dan}} \right).$$

If the bandwidth is treated as fixed, $\hat{\omega}^2_{Dan}$ is not a consistent estimator of ω^2, but is asymptotically unbiased, and under the null hypothesis of $E(d_t) = 0$, DM_{CI} follows an asymptotic t_{2m} distribution. This fixed-m treatment results in a test with appealing finite sample properties, offering better size control relative to the $m \to \infty$ treatment that results in standard normal limit theory. Coroneo and Iacone observe that the t_{2m} distribution can act as a better approximation of the true null distribution for a smaller bandwidth, whereas larger bandwidths can be associated with higher power, hence a size-power trade-off emerges. Following these authors, we consider two versions of the test, setting the bandwidths according to $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/4} \rfloor$ (where $\lfloor . \rfloor$ denotes the integer part of the argument), denoting the resulting test statistics by $DM_{CI,1}$ and $DM_{CI,2}$, respectively. Note that for any given sample size, m is then treated as a fixed number so that the fixed-m asymptotic theory can be applied, with critical values drawn from the t_{2m} distribution.

In addition to the above methods, we also experimented with other possible solutions to the negative variance estimate problem. We considered replacing a negative long-run variance estimate with a modified estimate based on reducing the rectangular kernel bandwidth sequentially until a positive estimate was obtained, and we investigated the exponential covariogram-based long-run variance estimator proposed in the spatial prediction context by Hering and Genton (2011). We also considered alternatives to the Bartlett long-run variance estimator with

bandwidth $h-1$, examining results for the Bartlett kernel using a larger bandwidth setting of $2(h-1)$, and also the standard and pre-whitened quadratic spectral long-run variance estimators of Andrews (1991) and Andrews and Monahan (1992) with automatic bandwidth selection. However, these alternatives did not deliver superior finite sample size and power performance relative to the better of the approaches considered above, hence we do not detail these tests and their results in this paper; full results are available from the authors on request.

5 Finite sample size and power

In this section we consider the finite sample performance of the different methods outlined in the previous section. We first consider testing for equal forecast accuracy, again focusing on mean squared error loss ($d_t = e_{1t}^2 - e_{2t}^2$), and simulate the empirical sizes of the MDM_{rej}, MDM_{non}, MDM_{SR}, MDM_B, $DM_{CI,1}$ and $DM_{CI,2}$ testing approaches, with the tests conducted against a two-sided alternative at the nominal 0.10-level. In addition to these six approaches, for comparison we also report results for the DM_{Bart} statistic compared with t_{T-1} critical values, which always employs the Bartlett kernel-based estimator $\hat{\omega}^2_{Bart}$ regardless of the sign of the rectangular kernel-based estimator $\hat{\omega}^2$. As with the earlier simulations in section 3, we use a standard simulation setup in line with DM, HLNa and Clark (1999). Table 3 reports the sizes for the same simulation DGPs that were considered in the negative long-run variance simulations of section 3 when the null hypothesis was imposed ($R = 1$). Note that $DM_{CI,1}$ and $DM_{CI,2}$ are identical when $T = 16$ since $\lfloor T^{1/3} \rfloor = \lfloor T^{1/4} \rfloor$ in this case.

When $h = 1$, the original MDM statistic cannot suffer from negative long-run variance estimation problems, so MDM_{rej}, MDM_{non}, MDM_{SR}, MDM_B all amount to simply conducting MDM. (Note also that when $h = 1$, no serial correlation is present in the DGP, hence the θ_j settings play no role.) Here, the test is well behaved, with sizes very close to the nominal level, with only modest under-size displayed for $T = 8$ and $T = 16$. A very similar pattern of size behaviour is also seen for $DM_{CI,1}$ and $DM_{CI,2}$, while DM_{Bart} exhibits some minor over-size but is also generally well behaved. All tests are therefore reliable for one-step-ahead forecasts and there is little to choose between them in terms of finite sample size.

For multi-step-ahead forecasts ($h > 1$), the possibility of negative long-run variance estimates arises and so the method of dealing with these problem cases results in different size properties for the overall procedures that we consider. The MDM_{rej} approach translates any negative long-run variance estimate into a rejection of the null, thus the high frequency of negative estimates for larger h and smaller T induces a high degree of over-size for this approach. In line with the results of Table 1, the size of MDM_{rej} reaches almost 0.50, and such large upward size distortions render this procedure invalid. The DM_{Bart} test can also exhibit severe over-size, consistent with the simulations of Clark (1999), with size rising to almost 0.50 in the worst cases. The MDM_B method achieves better size control through use of the Bartlett kernel only in problem cases, but is again subject to quite substantial over-size for moderate values

of h and T, with empirical size rising above 0.30 in the case of high degree serial correlation. The MDM_{SR} approach (which replaces negative long-run variance estimates with a short-run variance estimate) offers better size control for the cases of no serial correlation and modest serial correlation, but, as might be expected, when the degree of serial correlation is high, the simplification of using only a short-run variance results in substantial size distortions. Of the MDM-based approaches, the best performing method is MDM_{non} (which translates negative variance estimates into non-rejections of the null). However, the size can still be inflated above the nominal level, with sizes of around 0.16 occurring. In contrast, the $DM_{CI,1}$ and $DM_{CI,2}$ weighted periodogram approaches offer a much greater degree of size control across h and T. Apart from the case of $T = 8$ with high degree serial correlation, the two versions generally have size close to 0.10, with the worst upward size distortion being a size below 0.12, offering a clear improvement over the other methods considered. When $T = 8$, $h > 3$ and the errors are highly serially correlated, $DM_{CI,1}$ can suffer from more substantial over-size, while $DM_{CI,2}$ retains excellent size control. The attractive finite sample size results reported in Coroneo and Iacone (2016) for moderate sample sizes and forecast horizons therefore extend to the small sample and longer horizon region under focus here, particularly for $DM_{CI,2}$, suggesting a valuable role for the DM_{CI} approach in delivering forecast accuracy tests with reliable size in small samples.

When comparing results for the over-sized DM_{Bart} test and the well-behaved $DM_{CI,2}$ test, both of which always use a long-run variance estimator that is guaranteed to be positive yet have very different finite sample size properties, it is interesting to examine the differences between the tests, so as to ascertain the components of $DM_{CI,2}$ that are instrumental in achieving size control. The $DM_{CI,2}$ statistic makes use of a different form of long-run variance estimator (a weighted periodogram estimator with Daniell kernel) compared to the DM_{Bart} statistic (which uses a weighted autocovariance estimator with Bartlett kernel), and the $DM_{CI,2}$ test adopts critical values from the t_{2m} distribution (based on fixed-m asymptotic theory) while the DM_{Bart} test uses t_{T-1} critical values (based on a limiting standard normal distribution obtained from $m \to \infty$ asymptotic theory). To gain some insight into the relative contributions of the change in long-run variance estimator and the change in critical values, we computed the size of a hybrid test that compares the $DM_{CI,2}$ statistic with t_{T-1} critical values. Results from these unreported simulations (which are available from the authors on request) show that in the cases where DM_{Bart} is most over-sized ($h > 3$ with small T and moderate or high degree serial correlation), use of $DM_{CI,2}$ with t_{T-1} critical values roughly halves the extent of the size distortion, suggesting that the long-run variance estimator and critical values both play an important role in controlling small sample size. In situations where DM_{Bart} has size closer to the nominal level, comparing $DM_{CI,2}$ with t_{T-1} critical values results in relatively little size improvement (indeed in some cases the over-size is greater than that for DM_{Bart}), suggesting that it is the use of t_{2m} critical values that plays the dominant part in improving size in such cases.

In addition to evaluating the empirical sizes of the procedures, it is also important to assess

their relative powers. Table 4 reports the size-adjusted powers of the equal accuracy test procedures (with the exception of MDM_{rej}), where the critical values for each test are first obtained by simulation from the corresponding size experiment. The MDM_{rej} approach is not amenable to size-adjustment due to the high proportion of automatic rejections induced by negative long-run variance estimates being treated as zero, which cannot be corrected by adjusting critical values; regardless of this, the severe over-size properties of MDM_{rej} exclude it as a reliable procedure anyway. The DGPs are again those used in section 3, with R varied across T to keep the power levels broadly similar across different sample sizes. When $h = 1$, where the MDM_{non}, MDM_{SR} and MDM_B procedures simply reduce to MDM, and where DM_{Bart} only differs from MDM by $\sqrt{T/(T-1)}$ with the same size-adjusted power, it is clear that the original MDM test can offer decent power gains over $DM_{CI,1}$ and $DM_{CI,2}$, particularly for smaller samples where we see gains around 0.15 relative to $DM_{CI,1}$, and up to 0.35 relative to $DM_{CI,2}$. Thus in the one-step-ahead context, where the tests are correctly sized and no negative long-run variance estimation problems arise, use of MDM is to be recommended.

When $h > 1$, however, the power rankings change. DM_{Bart} often displays attractive levels of size-adjusted power, but given the very poor small sample size performance of this test, it would be difficult to justify its use in practice. Of the better size controlled tests, when $T = 8$, $DM_{CI,1}$ outperforms all the MDM-based procedures for all forecast horizons, with worthwhile power gains of up to 0.13 displayed. For larger sample sizes, $DM_{CI,1}$ power can dip a little below that of the MDM-based approaches when h is small, but for the longer forecast horizons, $DM_{CI,1}$ again offers decent power gains. The $DM_{CI,2}$ procedure is identical to $DM_{CI,1}$ when $T = 16$, but for the other sample sizes, the additional size robustness that $DM_{CI,2}$ delivers comes at some cost to size-adjusted power. This is most noticeable for $T = 8$ where the power of $DM_{CI,2}$ is markedly below that of $DM_{CI,1}$, while for the larger sample sizes, the differences between $DM_{CI,1}$ and $DM_{CI,2}$ are quite modest. Among the three MDM-based methods, power differences emerge for smaller values of T, becoming more exaggerated as h increases, with MDM_{non} displaying the lowest relative power (as expected given its conservative approach to dealing with negative long-run variance estimates), followed by MDM_B and then MDM_{SR}. Finally, considering the impact of serial correlation in the errors, the power levels of all the methods decrease as the degree of serial correlation increases, but the relative rankings of the tests are largely preserved.

Taking the multi-step-ahead size and power results together, with the exception of $T = 8$ in the case of $h > 3$, $DM_{CI,1}$ emerges as the best performing test overall, with reliable finite sample size performance and relatively high levels of power. When $T = 8$ and $h > 3$, $DM_{CI,2}$ offers better size control than $DM_{CI,1}$ for highly serially correlated errors, and although this comes with a loss in size-adjusted power, $DM_{CI,2}$ still has power that is generally a little higher than MDM_{non} (the best size controlled of the MDM-based methods), in addition to more reliable size. A possible role for MDM_{non} could be envisaged for practitioners who desire a simple approach that remains in the MDM-based framework. Otherwise, it is clear that $DM_{CI,1}$ or

$DM_{CI,2}$ should be employed whenever $h > 1$, with the choice between these procedures made on the basis of the sample size and forecast horizon.

We next consider testing for forecast encompassing, beginning by simulating the empirical sizes of the MDM_{rej}, MDM_{non}, MDM_{SR}, DM_{Bart}, MDM_B, $DM_{CI,1}$ and $DM_{CI,2}$ procedures, conducting one-sided tests at the nominal 0.10-level for the same simulation DGPs as in section 3. The results are reported in Table 5. As for the equal accuracy case, for one-step-ahead forecasts, we observe that the MDM-based procedures (which are identical for $h = 1$) and the DM_{CI} tests display good finite sample size control. Indeed, the MDM test has almost no size distortions even for small T, while only a very modest amount of under-size is displayed for $DM_{CI,1}$ and $DM_{CI,2}$. On the other hand, DM_{Bart} is over-sized, particularly for the smaller sample sizes. When $h > 1$, we find a similar picture of size behaviour to that in Table 3. Specifically, MDM_{rej} and DM_{Bart} can be substantially over-sized, although the over-size of MDM_{rej} is less severe than for the equal accuracy case, since here the encompassing test is conducted against a one-sided alternative, hence only a proportion of the negative long-run variances obtained induce a rejection of the null. Of the MDM-based approaches, MDM_{non} offers the best size control with size always below 0.14, while MDM_{SR} and MDM_B suffer from greater size distortion, although to a lesser extent than was found in the equal accuracy testing context. $DM_{CI,1}$ and $DM_{CI,2}$ again have very good size behaviour across most settings, the exceptions being when the errors are highly serially correlated and either $T = 16$ together with $h = 5$ or $h = 6$, where $DM_{CI,1}$ and $DM_{CI,2}$ can suffer from a small amount of upward size distortion, or $T = 8$ and $h > 2$, in which case $DM_{CI,1}$ can again be over-sized, while $DM_{CI,2}$ offers greater size control in these cases.

Turning to power for forecast encompassing tests, Table 6 gives results for the size-adjusted powers of MDM_{non}, MDM_{SR}, DM_{Bart}, MDM_B, $DM_{CI,1}$ and $DM_{CI,2}$ for the relevant DGPs of section 3, with k varying across T as specified in that section; the critical values used for the size-adjustment are again obtained by simulation from the corresponding size experiment. The relative power rankings of the tests are unchanged compared to tests for equal forecast accuracy, therefore the comments and conclusions outlined above are equally applicable in this context. We again find that MDM has a power advantage over the DM_{CI} tests for $h = 1$, while $DM_{CI,1}$ generally outperforms the other procedures for $T = 8$ and $h > 1$, and for the longer forecast horizons when T is larger. $DM_{CI,2}$ again has generally lower power than $DM_{CI,1}$, although this is only of real import when $T = 8$. Once again, therefore, MDM is to be recommended for one-step-ahead forecasts, but for multi-step-ahead forecasts, apart from a potential role for MDM_{non} when a simple MDM-based modification is desired, it is the $DM_{CI,1}$ and $DM_{CI,2}$ tests that are to be preferred. These tests offer the best finite sample performance in terms of size and relative power, with $DM_{CI,2}$ recommended for $T = 8$ when $h > 2$, and $DM_{CI,1}$ otherwise.

6 Simulations calibrated from empirical data

In order to ensure that our simulation results are representative of what is likely to be encountered in practical applications, we now consider a set of simulations for a DGP where the sample sizes, forecast horizons and forecast error serial correlation settings are all calibrated according to a particular application in the literature. Specifically, we follow the Dreger and Wolters (2014) application where Euro-area inflation is forecast one, two and three years ahead from an autoregressive model using quarterly data. We obtained HICP inflation data from the authors for the period 1981Q1–2010Q4, and, following Dreger and Wolters, we construct 1-, 4-, 8- and 12-quarter inflation rates as follows

$$\pi_t^h = \frac{4}{h}\log(pc_t/pc_{t-h}), \quad h = 1, 4, 8, 12$$

where pc_t denotes the consumer price index (HICP). The forecasting equation uses the benchmark model of Dreger and Wolters, based on prediction using current and lagged quarterly inflation:

$$\pi_{t+h}^h = \alpha_1\pi_t^1 + \alpha_2\pi_{t-1}^1 + \alpha_3\pi_{t-2}^1 + \alpha_4\pi_{t-3}^1 + \varepsilon_{t+h}. \tag{4}$$

Following Dreger and Wolters, at each forecast horizon we first estimate the model over the initial in-sample period 1983Q1–2002Q4. The parameter estimates are then used to produce the first $h = 4$, $h = 8$ and $h = 12$ forecasts for the periods 2003Q4, 2004Q4 and 2005Q4, respectively. The in-sample period is then extended by one observation and (4) is re-estimated, with the results used to obtain the next forecast for 2004Q1 at $h = 4$, 2005Q1 at $h = 8$ and 2006Q1 at $h = 12$. Continuing in this recursive manner, the final forecast for each forecast horizon is obtained for time 2010Q4, therefore producing a total of $T = 29$ forecasts for $h = 4$, $T = 25$ for $h = 8$, and $T = 21$ for $h = 12$. Denoting the forecast at a given time and horizon by $\hat{\pi}_{t+h}^h$, we can obtain three forecast error series

$$e_{t+h}^h = \pi_{t+h}^h - \hat{\pi}_{t+h}^h, \quad h = 4, 8, 12.$$

To determine the degree of serial correlation present in the forecast errors, we fit moving average processes to the three forecast error series, determining the order of MA process in each case according to the Akaike information criterion, selecting from MA processes up to order $h - 1$. We find the selected models to be $MA(3)$, $MA(2)$ and $MA(4)$ for $h = 4$, $h = 8$ and $h = 12$, respectively, with the fitted MA coefficients given in Table 7. Although these MA parameters have been estimated using a very small sample size, it can be seen that the values obtained are not inconsistent with the settings adopted in the earlier simulation exercises.

Given the calibrations obtained from the Dreger and Wolters application, we repeat the simulation experiments considered in sections 3 and 5, but now with the settings $T = \{29, 25, 21\}$, $h = \{4, 8, 12\}$ and the corresponding θ_j values from Table 7. Accordingly, Table 8 reports the frequency with which negative long-run variance estimates arise when using the standard rectangular kernel-based estimator (1), for both equal accuracy tests and encompassing tests,

under both the respective null and alternative hypotheses. The settings under the alternative for the three horizon/sample size pairings considered are $R = \{4, 6, 8\}$ (for testing equal accuracy) and $k = \{2, 1.8, 1.8\}$ (for encompassing testing), again chosen so that the test powers are broadly comparable across sample sizes. For $h = 4$, we observe a very low occurrence of negative long-run variance estimates, while for $h = 8$ the proportion of negative estimates across the simulations is in the region of 0.15, rising to around 0.33 for $h = 12$. These comments apply equally to tests for equal forecast accuracy and tests for forecast encompassing. The sample sizes considered in this empirically calibrated exercise lie inbetween the $T = 16$ and $T = 32$ settings used in the section 3 simulations, and two of the forecast horizons considered are greater than the range considered in section 3. However, it is clear that the pattern of frequencies for negative estimates is consistent with the earlier results, with a high incidence of problematic negative outcomes as the forecast horizon increases. This further demonstrates that the possibility of obtaining a negative long-run variance estimate is an empirically relevant issue when applying standard tests for equal accuracy and encompassing in small samples.

Table 9 presents the empirical sizes of the different test procedures for the empirically calibrated settings. Note that $DM_{CI,1}$ and $DM_{CI,2}$ are identical when $T = 25$ ($h = 8$) and $T = 21$ ($h = 12$), hence differences are only seen for the $h = 4$ results where $T = 29$. The DM_{CI} tests clearly offer the best size control of the procedures considered for tests for equal forecast accuracy and forecast encompassing, and the two bandwidth settings in $DM_{CI,1}$ and $DM_{CI,2}$ deliver similar size results. Given that here we consider longer forecast horizons than in section 5, it is reassuring to see that $DM_{CI,1}$ and $DM_{CI,2}$ retain good size control across h. As would be expected given the earlier simulations, substantial over-size is seen for MDM_{rej} and DM_{Bart}, particularly for the longer forecast horizons. Of the other MDM-based tests, MDM_{non} offers the best size control as before, with a maximum size around 0.15 observed, while MDM_{SR} and MDM_{B} can have size up to around 0.19 and 0.24 respectively. Table 10 reports the corresponding size-adjusted powers of the procedures, and, with the exception of the badly over-sized DM_{Bart} test, $DM_{CI,1}$ displays the best power performance, followed by $DM_{CI,2}$. In contrast, the best-sized MDM-based procedure, MDM_{non}, suffers from relatively low size-adjusted power for $h = 8$ and $h = 12$. These results clearly strengthen the case for use of $DM_{CI,1}$ or $DM_{CI,2}$ in practical applications.

7 Impact of model parameter estimation uncertainty

Beginning primarily with West (1996), much work on forecast evaluation testing has focused on cases where the forecasts have been produced by estimated models, either non-nested or nested, and more sophisticated methods have been proposed to properly account for the impact that model parameter estimation uncertainty can have on the distributions of DM-type forecast accuracy and encompassing tests in such circumstances. For reviews of this literature, see West (2006) and Clark and McCracken (2013). In some situations where forecasts have been obtained

from estimated models, the original DM approach is asymptotically valid without the need for any modification. Examples are where the forecast models are non-nested, linear and estimated by ordinary least squares (OLS), along with the loss function being mean squared forecast error, or when the number of forecast observations is small relative to the number of observations used for model estimation. In this section, we consider a set of simulations designed to examine the same issues of negative long-run variance estimation and test size performance in small samples, but now where the forecasts have first been obtained from estimated models. In order to focus on tests that are asymptotically valid, we restrict attention to tests for equal mean squared forecast error where the forecasts are obtained from non-nested linear models estimated by OLS.

Our forecasting exercise involves an in-sample period for model estimation, $t = 1, ..., N$, and an out-of-sample period for forecast evaluation, $t = N+1, ..., N+T$. We consider the following DGP

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad t = 1, ..., N+T$$

where, without loss of generality, we interpret x_{1t} and x_{2t} to be predictor variables useful for forecasting y_t at horizon h. We set $[x_{1t}, x_{2t}]' \sim N(0, I_2)$ and, as our focus here is on the impact of parameter estimation uncertainty rather than forecast error serial correlation, we simply generate $\varepsilon_t \sim N(0,1)$, $t = 1, ..., N+T$, independently of $[x_{1t}, x_{2t}]'$. As in the $\theta_j = 0$ simulations of sections 3 and 5, we do not assume knowledge of this lack of serial correlation when constructing the test statistics, so the results can be compared directly with the $\theta_j = 0$ sections of Tables 1 and 3. We consider two model-based forecasts, with the models given by

$$\begin{aligned} \text{Model 1} &: y_t = \beta_1 x_{1t} + e_{1t} \\ \text{Model 2} &: y_t = \beta_2 x_{2t} + e_{2t} \end{aligned}$$

which are first estimated by OLS over the period $t = 1, ..., N$ to give the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. The two forecast series are then specified as

$$\begin{aligned} f_{1t} &= \hat{\beta}_1 x_{1t}, \quad t = N+1, ..., N+T \\ f_{2t} &= \hat{\beta}_2 x_{2t}, \quad t = N+1, ..., N+T \end{aligned}$$

with the corresponding forecast errors

$$\begin{aligned} \hat{e}_{1t} &= y_t - \hat{\beta}_1 x_{1t} \\ \hat{e}_{2t} &= y_t - \hat{\beta}_2 x_{2t}. \end{aligned}$$

The tests for equal forecast accuracy are then defined exactly as in section 4, but with e_{1t} and e_{2t} replaced with $\hat{e}_{1t}$ and $\hat{e}_{2t}$, respectively. By setting $\beta_1 = \beta_2$, it is straightforward to show that $E(e_{1t}^2) = E(e_{2t}^2)$, so that the forecasts have equal accuracy in population, thereby giving the null hypothesis for our testing exercise. We set $\beta_1 = \beta_2 = 1$ and consider two in-sample period

sizes, $N = 40$ and $N = 80$, combined with the same set of out-of-sample sizes and forecast horizons employed in the earlier simulations of sections 3 and 5.

Table 11 reports results for the frequency of negative long-run variance estimates. On comparing these results (for both $N = 40$ and $N = 80$) with the $\theta_j = 0$ section of Panel A of Table 1, we find that the results are almost identical, hence the presence of forecast model parameter estimation uncertainty has almost no effect on the prevalence of negative estimates. Table 12 gives results for the empirical sizes of the test procedures, and while some minor differences are seen between the results for $N = 40$ and $N = 80$, the results for the different in-sample sizes are broadly similar to each other, and there is little difference between these results and those for $\theta_j = 0$ in Table 3. Once again, therefore, the impact of estimating the forecast models is very slight, and the same comments made in section 5 apply here. The fundamental findings of (i) a high frequency of negative long-run variance estimates when evaluating multi-step-ahead forecasts using small numbers of out-of-sample forecast errors, and (ii) the DM_{CI} tests offering the best size control among the alternative procedures considered, are therefore equally relevant in the context of forecasts obtained from estimated models.

8 Conclusion

In this paper, we have highlighted that application of the standard DM-based tests for equal forecast accuracy and forecast encompassing can often result in a negative long-run variance estimate when dealing with multi-step-ahead predictions and small, but empirically relevant, sample sizes. Having examined a number of possible approaches to dealing with this problem, we have found that the recently proposed testing approach of Coroneo and Iacone (2016), which uses a weighted periodogram long-run variance estimator combined with fixed-bandwidth asymptotics, offers the best overall finite sample size and power performance. Use of this test with a bandwidth setting of $\lfloor T^{1/3} \rfloor$ or $\lfloor T^{1/4} \rfloor$ (the choice being determined by the sample size and forecast horizon involved) results in only modest size distortions, while power levels are appealing relative to other approaches, permitting reliable inference even in the small sample/long horizon cases we consider. Aside from this preferred approach, a case could possibly be made for a strategy that uses the MDM tests of Harvey *et al.* (1997, 1998) when a positive long-run variance estimate is obtained, and defaulting to a non-rejection of the null hypothesis when a negative long-run variance arises; while this approach does not perform as well as the Coroneo and Iacone (2016) procedure, it does have the advantage of simplicity, since no additional computation beyond calculation of the MDM statistic is required. Finally, when the forecast evaluation is being done with one-step-ahead predictions, no negative long-run variance estimates can arise with the standard tests, and the MDM tests provide good size control and superior power to the Coroneo and Iacone (2016) test.

The simulations conducted in this paper considered a range of sample sizes and forecast horizons, as well as different degrees of serial correlation in the forecast errors. While we have

focused throughout on normally distributed forecast errors, we also considered simulations based on errors drawn from the t_6 distribution, given that forecast errors often appear to display fat-tailed behaviour. We found the results to be qualitatively similar to those based on normal errors, hence our conclusions would be unchanged under such a forecast error assumption. Finally, we note that the issue of negative long-run variance estimates would also be relevant in the recommended test of Harvey and Newbold (2000) for *multiple* forecast encompassing (where the null is that one forecast encompasses a number of competing predictors), since this test employs a multivariate version of the MDM approach. It would be expected that the variance-covariance estimator in the test statistic could fail to be positive definite for small samples and multi-step-ahead predictions, and in future work it would be interesting to consider extensions of the above techniques to that context.

References

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

Andrews, D.W.K. and Monahan, J.C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60, 953–966.

Ashley, R. (2003). Statistically significant forecasting improvements: how much out-of-sample data is likely necessary? *International Journal of Forecasting*, 19, 229–239.

Ashley, R. and Tsang, K.P. (2014). Credible Granger-causality inference with modest sample lengths: a cross-sample validation approach. *Econometrics*, 2, 72–91.

Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20, 451–468.

Caporale, G.M. and Gil-Alana, L. (2014). Long-run and cyclical dynamics in the US stock market. *Journal of Forecasting*, 33, 147–161.

Chow, H.K. and Choy, K.M. (2006). Forecasting the global electronics cycle with leading indicators: A Bayesian VAR approach. *International Journal of Forecasting*, 22, 301–315.

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

Clark, T.E. and McCracken, M. (2013). Advances in forecast evaluation. In *Handbook of Economic Forecasting, Vol. 2, Part B*, Elliott, G. and Timmerman, A. (eds.), pp. 1107–1201, Elsevier, Amsterdam.

Clements, M.P. and Harvey, D.I. (2009). Forecast combination and encompassing. In *Palgrave Handbook of Econometrics, Vol. 2: Applied Econometrics*, Mills, T.C. and Patterson, K. (eds.), pp. 169–198, Palgrave Macmillan, Basingstoke.

Coroneo, L. and Iacone, F. (2016). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. Discussion Paper, Department of Economics, University of York.

Dib, A., Gammoudi, M. and Moran, K. (2008). Forecasting Canadian time series with the New Keynesian model. *Canadian Journal of Economics*, 41, 138–165.

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

Dreger, C. and Wolters, J. (2014). Money demand and the role of monetary indicators in forecasting euro area inflation. *International Journal of Forecasting*, 30, 303–312.

Harvey, D.I., Leybourne, S.J. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13 281–291.

Harvey, D.I., Leybourne, S.J. and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, 16, 254–259.

Harvey, D.I., Leybourne, S.J. and Newbold, P. (1999). Forecast evaluation tests in the presence of ARCH. *Journal of Forecasting*, 18, 435–445.

Harvey, D.I. and Newbold, P. (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics*, 15, 471–482.

Hering, A.S. and Genton, M.G. (2011). Comparing spatial predictions. *Technometrics*, 53, 414–425.

Hualde, J. and Iacone, F. (2017). Fixed bandwidth asymptotics for the studentized mean of fractionally integrated processes. *Economics Letters*, 150, 39–43.

Mehl, A. (2009). The yield curve as a predictor and emerging economies. *Open Economies Review*, 20, 683–716.

Qin, D., Cagas, M.A., Ducanes, G., Magtibay-Ramos, N. and Quising, P. (2008). Automatic leading indicators versus macroeconometric structural models: A comparison of inflation and GDP growth forecasting. *International Journal of Forecasting*, 24, 399–413.

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

West, K.D. (2006). Forecast evaluation. In *Handbook of Economic Forecasting, Vol. 1*, Elliott, G., Granger, C.W.J. and Timmerman, A. (eds.), pp. 99–134, Elsevier, Amsterdam.

Table 1. Frequency of negative long-run variance estimates in tests for equal forecast accuracy.

	$\theta_j = 0$				$\theta_j = 0.9/(h-1)$				$\theta = (0.95, 0.9, 0.8, 0.65, 0.6)$			
h	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$
Panel A. $R = 1$ *(null)*												
2	0.074	0.015	0.001	0.000	0.036	0.003	0.000	0.000	0.035	0.004	0.000	0.000
3	0.175	0.064	0.013	0.000	0.131	0.032	0.003	0.000	0.101	0.019	0.002	0.000
4	0.275	0.120	0.031	0.002	0.231	0.077	0.016	0.001	0.182	0.047	0.004	0.000
5	0.352	0.175	0.055	0.008	0.325	0.141	0.034	0.004	0.297	0.081	0.012	0.001
6	0.422	0.217	0.085	0.017	0.406	0.186	0.060	0.008	0.367	0.121	0.024	0.001
Panel B. $R > 1$ *(alternative)*												
2	0.057	0.007	0.001	0.000	0.038	0.002	0.000	0.000	0.037	0.002	0.000	0.000
3	0.169	0.052	0.009	0.000	0.134	0.028	0.003	0.000	0.113	0.016	0.001	0.000
4	0.268	0.107	0.027	0.002	0.228	0.075	0.015	0.000	0.191	0.047	0.004	0.000
5	0.347	0.166	0.052	0.007	0.331	0.135	0.033	0.004	0.310	0.081	0.011	0.001
6	0.430	0.211	0.085	0.016	0.414	0.189	0.060	0.007	0.368	0.130	0.024	0.001

Table 2. Frequency of negative long-run variance estimates in tests for forecast encompassing.

	$\theta_j = 0$				$\theta_j = 0.9/(h-1)$				$\theta = (0.95, 0.9, 0.8, 0.65, 0.6)$			
h	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$
Panel A. $\rho = 1$ *(null)*												
2	0.070	0.015	0.001	0.000	0.037	0.003	0.000	0.000	0.036	0.003	0.000	0.000
3	0.182	0.067	0.012	0.001	0.130	0.030	0.002	0.000	0.104	0.018	0.001	0.000
4	0.275	0.126	0.029	0.003	0.226	0.083	0.014	0.001	0.177	0.045	0.005	0.000
5	0.360	0.169	0.054	0.009	0.330	0.132	0.033	0.002	0.306	0.081	0.010	0.000
6	0.424	0.220	0.089	0.017	0.406	0.180	0.061	0.008	0.360	0.126	0.021	0.002
Panel B. $\rho < 1$ *(alternative)*												
2	0.063	0.012	0.001	0.000	0.033	0.003	0.000	0.000	0.034	0.003	0.000	0.000
3	0.169	0.054	0.009	0.000	0.129	0.029	0.002	0.000	0.105	0.020	0.001	0.000
4	0.263	0.116	0.027	0.003	0.221	0.081	0.012	0.001	0.190	0.044	0.003	0.000
5	0.349	0.166	0.051	0.008	0.331	0.132	0.032	0.002	0.300	0.084	0.011	0.000
6	0.433	0.215	0.085	0.018	0.414	0.184	0.060	0.009	0.366	0.123	0.023	0.002

Table 3. Empirical size of nominal 0.10-level tests for equal forecast accuracy.

		$\theta_j = 0$				$\theta_j = 0.9/(h-1)$				$\theta = (0.95, 0.9, 0.8, 0.65, 0.6)$			
h		$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$
1	MDM_{rej}	0.085	0.091	0.102	0.103								
	MDM_{non}	0.085	0.091	0.102	0.103								
	MDM_{SR}	0.085	0.091	0.102	0.103								
	DM_{Bart}	0.107	0.103	0.108	0.105								
	MDM_{B}	0.085	0.091	0.102	0.103								
	$DM_{CI,1}$	0.085	0.087	0.099	0.098								
	$DM_{CI,2}$	0.086	0.087	0.097	0.097								
2	MDM_{rej}	0.203	0.145	0.124	0.109	0.165	0.123	0.116	0.107	0.164	0.122	0.116	0.107
	MDM_{non}	0.129	0.130	0.123	0.109	0.129	0.120	0.116	0.107	0.129	0.119	0.116	0.107
	MDM_{SR}	0.137	0.131	0.123	0.109	0.135	0.120	0.116	0.107	0.136	0.119	0.116	0.107
	DM_{Bart}	0.150	0.125	0.118	0.107	0.185	0.158	0.154	0.142	0.186	0.159	0.155	0.142
	MDM_{B}	0.153	0.133	0.123	0.109	0.144	0.121	0.116	0.107	0.144	0.120	0.116	0.107
	$DM_{CI,1}$	0.081	0.086	0.095	0.102	0.092	0.086	0.097	0.099	0.092	0.085	0.097	0.099
	$DM_{CI,2}$	0.086	0.086	0.097	0.101	0.079	0.086	0.096	0.095	0.079	0.085	0.096	0.094
3	MDM_{rej}	0.294	0.207	0.153	0.122	0.256	0.183	0.134	0.113	0.235	0.167	0.126	0.112
	MDM_{non}	0.119	0.143	0.140	0.122	0.125	0.150	0.131	0.112	0.134	0.148	0.124	0.112
	MDM_{SR}	0.138	0.150	0.142	0.122	0.145	0.156	0.132	0.112	0.159	0.153	0.124	0.112
	DM_{Bart}	0.193	0.151	0.130	0.114	0.234	0.193	0.162	0.146	0.261	0.209	0.175	0.159
	MDM_{B}	0.183	0.163	0.144	0.122	0.180	0.163	0.132	0.112	0.187	0.157	0.124	0.112
	$DM_{CI,1}$	0.082	0.092	0.094	0.098	0.109	0.099	0.100	0.097	0.120	0.094	0.094	0.095
	$DM_{CI,2}$	0.083	0.092	0.094	0.095	0.087	0.099	0.089	0.092	0.086	0.094	0.087	0.089
4	MDM_{rej}	0.366	0.263	0.179	0.131	0.340	0.229	0.158	0.117	0.321	0.207	0.139	0.115
	MDM_{non}	0.090	0.143	0.148	0.128	0.108	0.144	0.142	0.116	0.139	0.160	0.135	0.115
	MDM_{SR}	0.112	0.157	0.151	0.128	0.136	0.156	0.144	0.116	0.192	0.174	0.136	0.115
	DM_{Bart}	0.230	0.178	0.139	0.114	0.272	0.204	0.169	0.145	0.341	0.248	0.194	0.166
	MDM_{B}	0.183	0.183	0.157	0.129	0.196	0.173	0.147	0.117	0.233	0.182	0.137	0.115
	$DM_{CI,1}$	0.074	0.091	0.095	0.092	0.108	0.100	0.097	0.093	0.153	0.101	0.097	0.098
	$DM_{CI,2}$	0.079	0.091	0.097	0.094	0.085	0.100	0.088	0.089	0.079	0.101	0.086	0.088
5	MDM_{rej}	0.415	0.304	0.215	0.150	0.404	0.282	0.187	0.136	0.406	0.246	0.165	0.134
	MDM_{non}	0.063	0.130	0.160	0.142	0.079	0.141	0.153	0.132	0.110	0.164	0.153	0.133
	MDM_{SR}	0.091	0.147	0.165	0.143	0.118	0.162	0.157	0.132	0.218	0.198	0.157	0.133
	DM_{Bart}	0.269	0.191	0.150	0.123	0.313	0.229	0.176	0.153	0.410	0.278	0.217	0.186
	MDM_{B}	0.186	0.182	0.173	0.144	0.207	0.192	0.161	0.133	0.280	0.206	0.158	0.133
	$DM_{CI,1}$	0.079	0.093	0.097	0.097	0.107	0.106	0.103	0.099	0.198	0.108	0.103	0.102
	$DM_{CI,2}$	0.080	0.093	0.096	0.100	0.096	0.106	0.096	0.094	0.083	0.108	0.088	0.093
6	MDM_{rej}	0.469	0.340	0.239	0.165	0.463	0.320	0.216	0.148	0.452	0.277	0.187	0.136
	MDM_{non}	0.047	0.123	0.154	0.148	0.056	0.134	0.156	0.140	0.085	0.156	0.163	0.135
	MDM_{SR}	0.082	0.145	0.161	0.150	0.101	0.159	0.165	0.141	0.240	0.208	0.172	0.135
	DM_{Bart}	0.301	0.227	0.162	0.129	0.350	0.250	0.184	0.151	0.474	0.312	0.236	0.188
	MDM_{B}	0.193	0.196	0.176	0.153	0.221	0.202	0.174	0.142	0.311	0.218	0.173	0.135
	$DM_{CI,1}$	0.077	0.092	0.100	0.097	0.102	0.106	0.111	0.100	0.241	0.118	0.107	0.100
	$DM_{CI,2}$	0.081	0.092	0.096	0.097	0.096	0.106	0.095	0.096	0.098	0.118	0.089	0.088

Table 4. Size-adjusted power of nominal 0.10-level tests for equal forecast accuracy.

		$\theta_j = 0$				$\theta_j = 0.9/(h-1)$				$\theta = (0.95, 0.9, 0.8, 0.65, 0.6)$			
h		$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$
1	MDM_{non}	0.710	0.820	0.677	0.699								
	MDM_{SR}	0.710	0.820	0.677	0.699								
	DM_{Bart}	0.710	0.820	0.677	0.699								
	MDM_{B}	0.710	0.820	0.677	0.699								
	$DM_{CI,1}$	0.564	0.655	0.591	0.642								
	$DM_{CI,2}$	0.357	0.655	0.528	0.555								
2	MDM_{non}	0.454	0.688	0.638	0.679	0.396	0.594	0.521	0.559	0.384	0.595	0.520	0.556
	MDM_{SR}	0.480	0.689	0.638	0.679	0.404	0.596	0.521	0.559	0.402	0.596	0.520	0.556
	DM_{Bart}	0.646	0.795	0.673	0.689	0.535	0.669	0.542	0.560	0.531	0.655	0.540	0.559
	MDM_{B}	0.443	0.686	0.638	0.679	0.389	0.596	0.521	0.559	0.384	0.593	0.520	0.556
	$DM_{CI,1}$	0.609	0.648	0.606	0.612	0.511	0.545	0.471	0.499	0.497	0.545	0.472	0.500
	$DM_{CI,2}$	0.374	0.648	0.533	0.538	0.326	0.545	0.416	0.438	0.330	0.545	0.416	0.437
3	MDM_{non}	0.385	0.574	0.601	0.656	0.340	0.479	0.466	0.509	0.283	0.399	0.390	0.420
	MDM_{SR}	0.433	0.596	0.605	0.656	0.376	0.490	0.466	0.509	0.305	0.401	0.390	0.420
	DM_{Bart}	0.565	0.757	0.679	0.689	0.466	0.596	0.515	0.525	0.363	0.483	0.419	0.431
	MDM_{B}	0.393	0.565	0.601	0.656	0.349	0.478	0.465	0.509	0.287	0.396	0.389	0.420
	$DM_{CI,1}$	0.564	0.635	0.619	0.643	0.468	0.526	0.460	0.487	0.394	0.431	0.387	0.400
	$DM_{CI,2}$	0.374	0.635	0.556	0.558	0.307	0.526	0.437	0.415	0.290	0.431	0.346	0.352
4	MDM_{non}	0.388	0.505	0.531	0.638	0.339	0.441	0.454	0.523	0.249	0.310	0.303	0.361
	MDM_{SR}	0.506	0.561	0.540	0.639	0.401	0.466	0.459	0.523	0.278	0.315	0.301	0.361
	DM_{Bart}	0.570	0.717	0.652	0.682	0.467	0.580	0.528	0.551	0.310	0.392	0.339	0.380
	MDM_{B}	0.428	0.524	0.535	0.638	0.353	0.429	0.456	0.523	0.249	0.308	0.301	0.361
	$DM_{CI,1}$	0.604	0.627	0.592	0.647	0.493	0.525	0.486	0.512	0.340	0.373	0.324	0.347
	$DM_{CI,2}$	0.389	0.627	0.529	0.571	0.332	0.525	0.451	0.460	0.256	0.373	0.300	0.310
5	MDM_{non}	0.389	0.476	0.483	0.599	0.319	0.413	0.436	0.516	0.197	0.235	0.259	0.294
	MDM_{SR}	0.513	0.556	0.507	0.603	0.425	0.462	0.448	0.518	0.232	0.241	0.257	0.294
	DM_{Bart}	0.560	0.698	0.647	0.679	0.467	0.580	0.531	0.554	0.266	0.308	0.298	0.314
	MDM_{B}	0.418	0.517	0.498	0.602	0.364	0.428	0.446	0.519	0.243	0.235	0.256	0.294
	$DM_{CI,1}$	0.592	0.629	0.598	0.626	0.503	0.516	0.507	0.524	0.281	0.314	0.288	0.305
	$DM_{CI,2}$	0.395	0.629	0.533	0.526	0.322	0.516	0.444	0.448	0.234	0.314	0.261	0.257
6	MDM_{non}	0.318	0.441	0.461	0.577	0.297	0.387	0.397	0.505	0.185	0.225	0.205	0.267
	MDM_{SR}	0.505	0.545	0.488	0.581	0.419	0.471	0.417	0.506	0.193	0.230	0.203	0.268
	DM_{Bart}	0.546	0.677	0.618	0.681	0.482	0.589	0.519	0.561	0.219	0.268	0.250	0.280
	MDM_{B}	0.433	0.506	0.482	0.583	0.374	0.429	0.405	0.505	0.202	0.227	0.201	0.267
	$DM_{CI,1}$	0.597	0.632	0.583	0.638	0.495	0.552	0.491	0.533	0.224	0.288	0.247	0.279
	$DM_{CI,2}$	0.377	0.632	0.531	0.554	0.334	0.552	0.458	0.447	0.209	0.288	0.227	0.239

Table 5. Empirical size of nominal 0.10-level tests for forecast encompassing.

		$\theta_j = 0$				$\theta_j = 0.9/(h-1)$				$\theta = (0.95, 0.9, 0.8, 0.65, 0.6)$			
h		$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$
1	MDM_{rej}	0.100	0.100	0.102	0.103								
	MDM_{non}	0.100	0.100	0.102	0.103								
	MDM_{SR}	0.100	0.100	0.102	0.103								
	DM_{Bart}	0.142	0.119	0.111	0.108								
	MDM_{B}	0.100	0.100	0.102	0.103								
	$DM_{CI,1}$	0.091	0.095	0.102	0.103								
	$DM_{CI,2}$	0.088	0.095	0.098	0.100								
2	MDM_{rej}	0.153	0.130	0.110	0.110	0.138	0.120	0.112	0.108	0.138	0.120	0.112	0.108
	MDM_{non}	0.118	0.123	0.110	0.110	0.119	0.118	0.112	0.108	0.119	0.118	0.112	0.108
	MDM_{SR}	0.125	0.125	0.110	0.110	0.124	0.119	0.112	0.108	0.125	0.119	0.112	0.108
	DM_{Bart}	0.174	0.136	0.116	0.112	0.170	0.151	0.136	0.131	0.170	0.151	0.137	0.131
	MDM_{B}	0.134	0.126	0.110	0.110	0.129	0.119	0.112	0.108	0.129	0.119	0.112	0.108
	$DM_{CI,1}$	0.095	0.097	0.098	0.102	0.110	0.101	0.105	0.107	0.111	0.101	0.106	0.107
	$DM_{CI,2}$	0.094	0.097	0.095	0.101	0.094	0.101	0.094	0.099	0.093	0.101	0.094	0.099
3	MDM_{rej}	0.184	0.151	0.131	0.111	0.171	0.145	0.121	0.115	0.163	0.140	0.121	0.114
	MDM_{non}	0.094	0.119	0.125	0.111	0.106	0.129	0.120	0.115	0.110	0.131	0.120	0.114
	MDM_{SR}	0.113	0.127	0.126	0.111	0.124	0.133	0.121	0.115	0.131	0.133	0.120	0.114
	DM_{Bart}	0.189	0.147	0.130	0.114	0.192	0.166	0.147	0.135	0.206	0.178	0.157	0.142
	MDM_{B}	0.135	0.133	0.127	0.111	0.139	0.136	0.121	0.115	0.141	0.135	0.120	0.114
	$DM_{CI,1}$	0.095	0.095	0.104	0.105	0.121	0.105	0.106	0.106	0.130	0.106	0.108	0.106
	$DM_{CI,2}$	0.088	0.095	0.101	0.099	0.090	0.105	0.099	0.098	0.089	0.106	0.100	0.097
4	MDM_{rej}	0.214	0.181	0.134	0.116	0.202	0.174	0.131	0.118	0.206	0.161	0.128	0.121
	MDM_{non}	0.076	0.116	0.121	0.115	0.095	0.132	0.123	0.118	0.118	0.137	0.125	0.121
	MDM_{SR}	0.105	0.132	0.123	0.116	0.123	0.144	0.125	0.118	0.154	0.147	0.127	0.121
	DM_{Bart}	0.207	0.169	0.128	0.116	0.212	0.180	0.146	0.140	0.251	0.208	0.166	0.154
	MDM_{B}	0.141	0.146	0.126	0.116	0.150	0.152	0.126	0.118	0.172	0.150	0.127	0.121
	$DM_{CI,1}$	0.093	0.106	0.100	0.098	0.123	0.116	0.107	0.108	0.166	0.117	0.111	0.113
	$DM_{CI,2}$	0.091	0.106	0.097	0.097	0.101	0.116	0.100	0.102	0.090	0.117	0.100	0.104
5	MDM_{rej}	0.235	0.187	0.144	0.120	0.225	0.182	0.140	0.116	0.227	0.178	0.138	0.119
	MDM_{non}	0.057	0.102	0.117	0.115	0.060	0.115	0.124	0.116	0.075	0.136	0.133	0.119
	MDM_{SR}	0.092	0.122	0.123	0.115	0.103	0.134	0.129	0.116	0.154	0.158	0.135	0.119
	DM_{Bart}	0.222	0.173	0.136	0.114	0.222	0.184	0.148	0.130	0.271	0.226	0.174	0.152
	MDM_{B}	0.142	0.141	0.130	0.116	0.146	0.148	0.133	0.116	0.178	0.163	0.135	0.119
	$DM_{CI,1}$	0.092	0.098	0.100	0.098	0.113	0.117	0.108	0.103	0.182	0.131	0.117	0.108
	$DM_{CI,2}$	0.086	0.098	0.098	0.092	0.099	0.117	0.101	0.095	0.097	0.131	0.100	0.099
6	MDM_{rej}	0.251	0.207	0.163	0.125	0.255	0.198	0.158	0.122	0.253	0.195	0.141	0.124
	MDM_{non}	0.044	0.095	0.116	0.117	0.051	0.107	0.126	0.117	0.070	0.128	0.131	0.123
	MDM_{SR}	0.086	0.118	0.125	0.119	0.103	0.135	0.135	0.118	0.170	0.165	0.135	0.123
	DM_{Bart}	0.248	0.183	0.141	0.115	0.248	0.196	0.160	0.132	0.307	0.243	0.182	0.159
	MDM_{B}	0.148	0.143	0.134	0.120	0.164	0.153	0.140	0.119	0.198	0.171	0.136	0.123
	$DM_{CI,1}$	0.095	0.099	0.099	0.095	0.116	0.117	0.116	0.103	0.214	0.144	0.125	0.111
	$DM_{CI,2}$	0.089	0.099	0.099	0.091	0.103	0.117	0.102	0.095	0.120	0.144	0.100	0.099

Table 6. Size-adjusted power of nominal 0.10-level tests for forecast encompassing.

		$\theta_j = 0$				$\theta_j = 0.9/(h-1)$				$\theta = (0.95, 0.9, 0.8, 0.65, 0.6)$			
h		$T=8$	$T=16$	$T=32$	$T=64$	$T=8$	$T=16$	$T=32$	$T=64$	$T=8$	$T=16$	$T=32$	$T=64$
1	MDM_{non}	0.649	0.689	0.693	0.678								
	MDM_{SR}	0.649	0.689	0.693	0.678								
	DM_{Bart}	0.649	0.689	0.693	0.678								
	MDM_{B}	0.649	0.689	0.693	0.678								
	$DM_{CI,1}$	0.602	0.614	0.640	0.643								
	$DM_{CI,2}$	0.495	0.614	0.615	0.598								
2	MDM_{non}	0.529	0.629	0.680	0.657	0.460	0.537	0.555	0.542	0.461	0.536	0.556	0.542
	MDM_{SR}	0.553	0.633	0.680	0.657	0.463	0.539	0.555	0.542	0.467	0.537	0.556	0.542
	DM_{Bart}	0.630	0.670	0.691	0.660	0.515	0.559	0.561	0.542	0.512	0.558	0.561	0.544
	MDM_{B}	0.528	0.633	0.680	0.657	0.452	0.538	0.555	0.542	0.449	0.537	0.556	0.542
	$DM_{CI,1}$	0.603	0.622	0.659	0.634	0.516	0.518	0.530	0.519	0.512	0.519	0.529	0.517
	$DM_{CI,2}$	0.489	0.622	0.628	0.589	0.409	0.518	0.519	0.486	0.413	0.519	0.517	0.484
3	MDM_{non}	0.496	0.578	0.635	0.653	0.424	0.478	0.531	0.513	0.375	0.411	0.457	0.449
	MDM_{SR}	0.542	0.597	0.637	0.653	0.449	0.486	0.531	0.513	0.381	0.411	0.458	0.449
	DM_{Bart}	0.603	0.663	0.672	0.666	0.517	0.541	0.549	0.518	0.424	0.461	0.469	0.459
	MDM_{B}	0.505	0.583	0.635	0.653	0.421	0.479	0.532	0.513	0.350	0.409	0.458	0.449
	$DM_{CI,1}$	0.598	0.621	0.639	0.630	0.510	0.506	0.524	0.507	0.446	0.438	0.456	0.438
	$DM_{CI,2}$	0.500	0.621	0.611	0.599	0.431	0.506	0.506	0.475	0.379	0.438	0.433	0.414
4	MDM_{non}	0.467	0.537	0.617	0.647	0.410	0.461	0.531	0.529	0.290	0.354	0.393	0.390
	MDM_{SR}	0.541	0.560	0.626	0.649	0.458	0.480	0.535	0.530	0.338	0.358	0.392	0.390
	DM_{Bart}	0.574	0.635	0.678	0.672	0.505	0.535	0.565	0.540	0.374	0.397	0.413	0.394
	MDM_{B}	0.479	0.532	0.625	0.649	0.418	0.467	0.535	0.530	0.330	0.355	0.392	0.390
	$DM_{CI,1}$	0.610	0.592	0.646	0.658	0.513	0.512	0.548	0.520	0.379	0.397	0.399	0.382
	$DM_{CI,2}$	0.497	0.592	0.611	0.612	0.434	0.512	0.513	0.484	0.359	0.397	0.376	0.352
5	MDM_{non}	0.436	0.523	0.599	0.639	0.401	0.464	0.521	0.538	0.297	0.298	0.341	0.345
	MDM_{SR}	0.539	0.568	0.619	0.643	0.489	0.499	0.529	0.539	0.315	0.298	0.343	0.345
	DM_{Bart}	0.571	0.644	0.668	0.669	0.520	0.543	0.577	0.559	0.335	0.335	0.370	0.354
	MDM_{B}	0.472	0.536	0.609	0.641	0.428	0.474	0.525	0.539	0.301	0.289	0.344	0.345
	$DM_{CI,1}$	0.599	0.610	0.653	0.650	0.532	0.528	0.564	0.543	0.344	0.341	0.366	0.348
	$DM_{CI,2}$	0.509	0.610	0.614	0.624	0.448	0.528	0.532	0.510	0.327	0.341	0.350	0.326
6	MDM_{non}	0.389	0.491	0.580	0.625	0.363	0.446	0.514	0.551	0.248	0.272	0.332	0.320
	MDM_{SR}	0.530	0.576	0.615	0.634	0.464	0.482	0.527	0.555	0.274	0.265	0.335	0.321
	DM_{Bart}	0.577	0.634	0.675	0.684	0.510	0.544	0.575	0.576	0.295	0.309	0.351	0.335
	MDM_{B}	0.468	0.532	0.600	0.634	0.407	0.456	0.523	0.554	0.271	0.259	0.335	0.321
	$DM_{CI,1}$	0.590	0.612	0.664	0.662	0.526	0.533	0.565	0.567	0.297	0.318	0.346	0.331
	$DM_{CI,2}$	0.488	0.612	0.626	0.622	0.450	0.533	0.552	0.531	0.286	0.318	0.336	0.310

Table 7. Fitted moving average process parameter estimates using Dreger-Wolters forecast errors.

	$T = 29$, $h = 4$	$T = 25$, $h = 8$	$T = 21$, $h = 12$
θ_1	0.939	0.638	0.364
θ_2	1.220	0.799	0.837
θ_3	0.457		−0.233
θ_4			0.351

Table 8. Frequency of negative long-run variance estimates in tests for equal forecast accuracy and forecast encompassing using Dreger-Wolters-calibrated forecast errors.

	$T = 29$, $h = 4$	$T = 25$, $h = 8$	$T = 21$, $h = 12$
Panel A. Tests for equal forecast accuracy			
$R = 1$ (null)	0.007	0.158	0.336
$R > 1$ (alternative)	0.004	0.150	0.333
Panel B. Tests for forecast encompassing			
$\rho = 1$ (null)	0.008	0.147	0.333
$\rho < 1$ (alternative)	0.005	0.151	0.332

Table 9. Empirical size of nominal 0.10-level tests for equal forecast accuracy and forecast encompassing using Dreger-Wolters-calibrated forecast errors.

	Tests for equal forecast accuracy			Tests for forecast encompassing		
	$T = 29$, $h = 4$	$T = 25$, $h = 8$	$T = 21$, $h = 12$	$T = 29$, $h = 4$	$T = 25$, $h = 8$	$T = 21$, $h = 12$
MDM_{rej}	0.155	0.314	0.436	0.132	0.197	0.239
MDM_{non}	0.148	0.155	0.100	0.128	0.122	0.077
MDM_{SR}	0.150	0.191	0.165	0.129	0.149	0.130
DM_{Bart}	0.200	0.243	0.334	0.164	0.186	0.225
MDM_{B}	0.150	0.216	0.237	0.130	0.159	0.161
$DM_{CI,1}$	0.103	0.091	0.108	0.112	0.109	0.114
$DM_{CI,2}$	0.086	0.091	0.108	0.098	0.109	0.114

Table 10. Size-adjusted power of nominal 0.10-level tests for tests for equal forecast accuracy and forecast encompassing using Dreger-Wolters-calibrated forecast errors.

	Tests for equal forecast accuracy			Tests for forecast encompassing		
	$T = 29$, $h = 4$	$T = 25$, $h = 8$	$T = 21$, $h = 12$	$T = 29$, $h = 4$	$T = 25$, $h = 8$	$T = 21$, $h = 12$
MDM_{non}	0.538	0.386	0.325	0.556	0.471	0.424
MDM_{SR}	0.540	0.454	0.473	0.558	0.515	0.518
DM_{Bart}	0.625	0.586	0.504	0.592	0.598	0.560
MDM_{B}	0.537	0.418	0.394	0.557	0.496	0.477
$DM_{CI,1}$	0.591	0.558	0.535	0.584	0.593	0.580
$DM_{CI,2}$	0.537	0.558	0.535	0.563	0.593	0.580

Table 11. Frequency of negative long-run variance estimates in tests for equal forecast accuracy under the null using estimated models.

	$N = 40$				$N = 80$			
h	$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$
2	0.071	0.018	0.002	0.000	0.068	0.017	0.001	0.000
3	0.177	0.062	0.010	0.000	0.179	0.066	0.010	0.001
4	0.266	0.119	0.032	0.003	0.273	0.117	0.032	0.003
5	0.351	0.172	0.055	0.010	0.359	0.166	0.056	0.009
6	0.425	0.214	0.089	0.016	0.423	0.215	0.088	0.018

Table 12. Empirical size of nominal 0.10-level tests for equal forecast accuracy using estimated models.

		$N = 40$				$N = 80$			
h		$T = 8$	$T = 16$	$T = 32$	$T = 64$	$T = 8$	$T = 16$	$T = 32$	$T = 64$
1	MDM_{rej}	0.083	0.102	0.105	0.107	0.082	0.093	0.101	0.099
	MDM_{non}	0.083	0.102	0.105	0.107	0.082	0.093	0.101	0.099
	MDM_{SR}	0.083	0.102	0.105	0.107	0.082	0.093	0.101	0.099
	DM_{Bart}	0.104	0.115	0.112	0.110	0.103	0.103	0.105	0.102
	MDM_{B}	0.083	0.102	0.105	0.107	0.082	0.093	0.101	0.099
	$DM_{CI,1}$	0.082	0.093	0.098	0.104	0.076	0.087	0.098	0.101
	$DM_{CI,2}$	0.083	0.093	0.097	0.104	0.081	0.087	0.096	0.101
2	MDM_{rej}	0.199	0.155	0.127	0.124	0.197	0.155	0.125	0.108
	MDM_{non}	0.128	0.137	0.125	0.124	0.129	0.138	0.124	0.108
	MDM_{SR}	0.138	0.139	0.125	0.124	0.136	0.139	0.124	0.108
	DM_{Bart}	0.148	0.137	0.121	0.123	0.148	0.131	0.114	0.106
	MDM_{B}	0.154	0.141	0.125	0.124	0.151	0.143	0.124	0.108
	$DM_{CI,1}$	0.081	0.093	0.099	0.110	0.077	0.092	0.095	0.097
	$DM_{CI,2}$	0.086	0.093	0.097	0.107	0.080	0.092	0.091	0.096
3	MDM_{rej}	0.296	0.203	0.155	0.128	0.299	0.207	0.153	0.125
	MDM_{non}	0.119	0.141	0.145	0.127	0.120	0.141	0.142	0.125
	MDM_{SR}	0.138	0.149	0.145	0.127	0.137	0.148	0.143	0.125
	DM_{Bart}	0.197	0.149	0.132	0.120	0.196	0.149	0.124	0.116
	MDM_{B}	0.184	0.161	0.147	0.127	0.182	0.160	0.144	0.125
	$DM_{CI,1}$	0.080	0.089	0.097	0.102	0.080	0.089	0.097	0.101
	$DM_{CI,2}$	0.084	0.089	0.096	0.099	0.083	0.089	0.094	0.100
4	MDM_{rej}	0.356	0.268	0.178	0.139	0.364	0.260	0.182	0.140
	MDM_{non}	0.090	0.148	0.146	0.136	0.091	0.143	0.151	0.137
	MDM_{SR}	0.111	0.160	0.149	0.137	0.115	0.155	0.154	0.137
	DM_{Bart}	0.234	0.181	0.135	0.125	0.235	0.177	0.141	0.124
	MDM_{B}	0.183	0.189	0.153	0.137	0.186	0.179	0.159	0.137
	$DM_{CI,1}$	0.080	0.094	0.088	0.104	0.079	0.093	0.099	0.105
	$DM_{CI,2}$	0.080	0.094	0.088	0.100	0.086	0.093	0.097	0.103
5	MDM_{rej}	0.420	0.303	0.211	0.159	0.419	0.300	0.212	0.155
	MDM_{non}	0.070	0.131	0.156	0.149	0.060	0.134	0.156	0.145
	MDM_{SR}	0.097	0.147	0.162	0.150	0.087	0.149	0.161	0.146
	DM_{Bart}	0.277	0.200	0.154	0.135	0.274	0.194	0.152	0.134
	MDM_{B}	0.192	0.187	0.171	0.151	0.183	0.186	0.170	0.148
	$DM_{CI,1}$	0.082	0.089	0.103	0.106	0.073	0.091	0.099	0.102
	$DM_{CI,2}$	0.078	0.089	0.100	0.104	0.081	0.091	0.097	0.101
6	MDM_{rej}	0.476	0.339	0.244	0.169	0.475	0.335	0.242	0.176
	MDM_{non}	0.051	0.125	0.155	0.154	0.052	0.121	0.154	0.157
	MDM_{SR}	0.085	0.147	0.165	0.155	0.090	0.142	0.164	0.159
	DM_{Bart}	0.308	0.227	0.168	0.142	0.314	0.214	0.161	0.140
	MDM_{B}	0.206	0.194	0.181	0.158	0.207	0.191	0.180	0.162
	$DM_{CI,1}$	0.074	0.095	0.103	0.107	0.079	0.086	0.097	0.107
	$DM_{CI,2}$	0.076	0.095	0.102	0.104	0.081	0.086	0.096	0.103