Comparing predictive accuracy in small samples using fixed-smoothing asymptotics

Laura Coroneo*
University of York

Fabrizio Iacone
University of York

First draft: 26th August 2015
Current draft: 17th October 2016

Abstract

We consider fixed-smoothing asymptotics for the Diebold and Mariano (1995) test of predictive accuracy. We show that this approach allows to obtain predictive accuracy tests that are correctly sized even when only a small number of out of sample observations are available. We apply the fixed-smoothing asymptotics to the Diebold and Mariano (1995) test to evaluate the predictive accuracy of the Survey of Professional Forecasters (SPF) against a simple random walk. Our results show that the predictive ability of the SPF was partially spurious, especially in the last decade.

Keywords: Diebold and Mariano test, long run variance estimation, fixed-smoothing asymptotics, Heteroskedasticity Autocorrelation Robust (HAR) inference, SPF.
JEL classification: C12, C32, C53, E17

*The authors thank Jia Chen, David De Antonio Liedo, Bruce Hansen, Michael McCracken, Barbara Rossi and seminar participants at the University of York, the University of Nottingham, the University of Bologna, the 2015 International Conference on Computational and Financial Econometrics, the 2016 Royal Economic Society Annual Conference, the 2016 International Symposium of Forecasting and the Federal Reserve Bank of St. Louis Applied Time Series Econometrics Workshop for helpful suggestions. Corresponding author: Laura Coroneo, Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, UK. The support of the ESRC grant ES/K001345/1 is gratefully acknowledged.

1 Introduction

Good forecasts are key to good decision making. And being able to compare predictive accuracy is key to discriminate between good and bad forecasts. To this end, one of the most used tests to compare the predictive accuracy of two competing forecasts is the Diebold and Mariano (1995) [DM] test.

The DM test is based on a loss function associated with the forecast errors of each forecast, testing the null of zero expected loss differential of two competing forecasts. This framework allows to test for equal predictive accuracy using any loss function, and the test statistic is valid for contemporaneously correlated, serially correlated and non-normal forecast errors. The DM approach takes forecast errors as model-free and the test is valid also when the forecasts are produced from unknown models, as for example from forecast survey data.

When the forecasts are produced by estimated models, nested or non-nested, it is in general necessary to account for the impact of the model parameter estimation uncertainty on the distribution of the forecast accuracy test, see West (1996) and Clark and McCracken (2001). In this case, the limiting distribution of the test statistics depends on the specific modelling assumptions made for obtaining the forecast errors, see West (2006) and Clark and McCracken (2013). West (1996) shows that in some cases the DM approach is asymptotically valid even when forecasts are obtained from estimated models. This happens when the number of in sample observations is large relative to the number of out of sample observations or when a quadratic loss function is used to evaluate the accuracy of non-nested models estimated by ordinary least squares. In addition, in practice it is also not uncommon to compare forecasts produced by models for which it is not tractable to account for the model parameter estimation uncertainty. For these reasons, the DM test is still widely applied also when forecasts are obtained by estimated models, see Diebold (2015).

One reason for the success of the DM test is that the test statistic is simple to

compute and asymptotically normally distributed. As a consequence, the DM testing framework has been extended in several directions, see Diebold (2015), for example to test for conditional predictive ability, Giacomini and White (2006), and to deal with structural changes, Giacomini and Rossi (2010).

However, as also noted by DM, the test can be subject to large size distortions in small samples, which can be spuriously interpreted as superior predictive ability for one forecast. This is due to the fact that in the test statistic the long run variance is replaced by a consistent estimate, and standard limit normality is then still employed: this may be unsatisfactory when only a small number of out of sample observations are available. As remarked by Clark and McCracken (2013), *"one unresolved challenge in forecast test inference is achieving accurately sized tests applied at multi-step horizons – a challenge that increases as the forecast horizon grows and the size of the forecast sample declines"*.

In this paper, we consider two alternative asymptotics for testing assumptions about the expected loss differential of two competing forecasts. The first is the fixed-b approach of Kiefer and Vogelsang (2005), in which the limit properties of the weighted autocovariances estimate of the long run variance are derived assuming that the bandwidth to sample size ratio is constant. With this approach, the test to compare predictive accuracy has a non-standard limit distribution, that depends on the bandwidth to sample ratio b and on the kernel used to estimate the long run variance. The second alternative asymptotic that we consider is the fixed-m approach as in Sun (2013) and Hualde and Iacone (2015). In this case, the estimate of the long run variance is based on a weighted periodogram estimate with Daniell kernel and a truncation parameter m that is assumed to be constant as the sample size increases. The test to compare predictive accuracy has a t distribution with degrees of freedom that depends on the truncation parameter. This averaged periodogram estimate can be seen as one application of the orthonormal series variance estimate, see Phillips (2005).

Following Sun (2014a) and Sun (2014b) we refer to these two alternative asymptotics,

fixed-b and fixed-m, as "fixed-smoothing asymptotics". With these asymptotics, the assumptions on the bandwidth parameter implies that the estimate of the long run variance is not consistent. However, inference is more precise than with HAC standard asymptotics, and therefore it is often referred to as "Heteroskedasticity Autocorrelation Robust" (HAR).

We perform a Monte Carlo analysis and find that fixed-smoothing asymptotics deliver correctly sized predictive accuracy tests for highly correlated loss differentials even in small samples. Monte Carlo results also show that the power of the tests with fixed-smoothing asymptotics is comparable to the power of bootstrap tests. Overall, the findings of our Monte Carlo exercises are in line with the general literature on testing: the application of fixed-smoothing asymptotics to the DM test for predictive ability discussed here and the focus on small samples is novel. We apply fixed-smoothing asymptotics to evaluate the predictive accuracy of the Survey of Professional Forecasters' (SPF) forecasts for four core macroeconomic indicators: output growth, inflation, the unemployment rate and the three-month Treasury bill rate for the period from 1985:Q1 until 2014:Q4. Results show that part of the superior predictive accuracy indicated by the the DM test is spurious, especially in the most recent subsample.

For high frequency, large sample forecast evaluations, Patton (2015) and Li and Patton (2013) show that fixed-b asymptotics delivers considerable size improvements. For small samples, Harvey, Leybourne and Newbold (1997) propose a modified statistic and critical value: while this is only justified when the loss differential is an independent process, they find that their modified DM test alleviates the size distortion of the original test, even in presence of weak autocorrelation. The modifications of the DM test based on fixed-smoothing asymptotics that we propose have the advantage of being formally based on asymptotic theory, also when the loss differential is a dependent process. Harvey, Leybourne and Whitehouse (2015) perform an extensive Monte Carlo simulation exercise to examine the small sample size and power properties of different approaches. Their

results confirm that the fixed-m approach proposed in this paper outperforms standard approaches in small samples.

The paper is organized as follows. In Section 2 we introduce the test for equal predictive accuracy and in Section 3 we describe the DM estimate. The tests for equal predictive accuracy using fixed-b asymptotics and fixed-m asymptotics are described in Section 4. In Section 5 we present a Monte Carlo study and in Section 6 perform a Monte Carlo comparison with the bootstrap. Then in Section 7 we apply the testing methodology to analyse the predictive ability of the SPF and in Section 8 we conclude.

2 Comparing predictive accuracy

We consider the time series $y_1, ..., y_T$, for which we want to compare two h-step ahead forecasts $\widehat{y}^h_{1,t}$ and $\widehat{y}^h_{2,t}$ made at time $t-h$, with forecast errors $e^h_{1,t} = y_t - \widehat{y}^h_{1,t}$ and $e^h_{2,t} = y_t - \widehat{y}^h_{2,t}$, respectively. We denote by $L(e^h_{i,t})$, for $i = 1, 2$, the loss associated with the forecast error $e^h_{i,t}$; for example, a quadratic loss would be $L(e^h_{i,t}) = \left(e^h_{i,t}\right)^2$. The time-$t$ loss differential between the two forecasts is

$$d_t = L(e^h_{1,t}) - L(e^h_{2,t})$$

and it can be represented as

$$d_t = \mu + u_t$$

where u_t has $E(u_t) = 0$ and it is a weakly dependent process, with autocovariances $\gamma_j = E(u_t u_{t+j})$ and long run variance $\sigma^2 = \sum_{j=-\infty}^{\infty} \gamma_j$, with $0 < \sigma^2 < \infty$.

DM propose to test the hypothesis of equal predictive ability as $H_0 : \{\mu = 0\}$. Let

$$\overline{d} = \frac{1}{T} \sum\nolimits_{t=1}^{T} d_t$$

denote the sample mean of the loss differential. Under regularity conditions, it holds

that

$$\sqrt{T}\frac{\overline{d}-\mu}{\sigma} \rightarrow_d N(0,1). \tag{1}$$

Unfortunately, this statistic is unfeasible to test H_0, because σ^2 is unknown. However, the parameter σ^2 can be replaced with an appropriate estimate and, if a consistent estimate is used, then the limit normality is not affected by the replacement.

3 The DM estimate

A typical estimate for the long run variance is the Weighted AutoCovariances Estimate (WCE),

$$\widehat{\sigma}^2 = \widehat{\gamma}_0 + 2\sum\nolimits_{j=1}^{T-1} k(j/M)\,\widehat{\gamma}_j \tag{2}$$

where $\widehat{\gamma}_j = \frac{1}{T}\sum_{t=1}^{T-j}\widehat{u}_t\widehat{u}_{t+j}$, with $\widehat{u}_t = d_t - \overline{d}$, and $k(.)$ is a kernel function such that $k(0)=1$, $|k(\tau)|<1$, $k(\tau)=k(-\tau)$, $k(\tau)$ is continuous at $\tau=0$ and $\int_0^1 k(\tau)^2\,d\tau<\infty$. The parameter M is a bandwidth parameter (or a truncation lag), and for consistency of $\widehat{\sigma}^2$ regularity conditions include $M\rightarrow\infty$ and $M/T\rightarrow 0$ as $T\rightarrow\infty$. We refer to Hannan (1970) for a survey of these estimates, and for a discussion of which kernels ensure that $\widehat{\sigma}^2\geq 0$.

In a variation of this approach, DM note that if $\widehat{y}_{1t}^h$ is an optimal forecast h steps ahead, then e_{1t}^h is at most a MA($h-1$), and then propose to set $M=h-1$ and $k(j/M)=1$ if $j/M\leq 1$ and 0 otherwise, so

$$\widehat{\sigma}_{DM}^2 = \widehat{\gamma}_0 + 2\sum\nolimits_{j=1}^{h-1}\widehat{\gamma}_j. \tag{3}$$

This does not meet the condition $M\rightarrow\infty$, but the estimate is nevertheless consistent, because it exploits the assumption that u_t is MA($h-1$), thus ensuring

$$\sqrt{T}\frac{\overline{d}-\mu}{\widehat{\sigma}_{DM}} \rightarrow_d N(0,1). \tag{4}$$

The choice of $\widehat{\sigma}^2_{DM}$ may be very appealing, as it exploits information about the structure of u_t. However, the rectangular kernel used in (3) may generate negative estimates for $\widehat{\sigma}^2_{DM}$, which is undesirable. Moreover, the Monte Carlo exercise in DM suggests the possibility of large size distortions in small samples, which would be spuriously interpreted of superior predictive power for one forecast rule. DM mention the possibility of using alternative kernels and standard asymptotics, to avoid the risk of negative estimates of σ, but simulations in Clark (1999), in which a Bartlett kernel was used, do not suggest that simply replacing the kernel results in a definite improvement of the size distortion.

4 Fixed-smoothing asymptotics

4.1 Fixed-b asymptotics

Following the approach of Kiefer and Vogelsang (2005) we consider alternative asymptotics for the estimate (2): for given M, the ratio M/T is taken as fixed as $T \to \infty$. As M/T is fixed, letting $b = M/T$, this alternative approach is referred to as fixed-b asymptotics. With this assumption, Kiefer and Vogelsang (2005) show that the estimate of σ is not consistent and not even asymptotically unbiased. This implies that the standardized sample mean has a non-standard limit distribution that depends on b and on the kernel. Kiefer and Vogelsang (2005) provide a formula to generate quantiles of the limit distribution, that can be used as critical values in tests.

For fixed-b asymptotics and assuming that the Bartlett kernel is used, we introduce the notation

$$\widehat{\sigma}^2_{BART} = \widehat{\gamma}_0 + 2\sum\nolimits_{j=1}^{T-1} k_{BART}\left(j/M\right)\widehat{\gamma}_j,\ M/T \to b, \tag{5}$$

$$k_{BART}\left(x\right) = \begin{cases} 1-\left|x\right|, & \text{if } \left|x\right| \leq 1; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Kiefer and Vogelsang (2005) show that

$$\text{if } b \in (0,1]\text{, then } \sqrt{T}\frac{\overline{d}-\mu}{\widehat{\sigma}_{BART}} \Rightarrow \Phi_{BART}(b)\,, \tag{7}$$

where $\Rightarrow$ denotes weak convergence in the in the $D[0,1]$ space with the Skorohod topology. They characterise the limit distribution $\Phi_{BART}(b)$ and provide formulas to compute quantiles. For the Bartlett kernel with $b \leq 1$, these can be obtained using the formula

$$q(b) = \alpha_0 + \alpha_1 b + \alpha_2 b^2 + \alpha_3 b^3$$

where

$$\begin{aligned} \alpha_0 &= 1.6449,\ \alpha_1 = 2.1859,\ \alpha_2 = 0.3142,\ \alpha_3 = -0.3427 \text{ for 0.950 quantile} \\ \alpha_0 &= 1.9600,\ \alpha_1 = 2.9694,\ \alpha_2 = 0.4160,\ \alpha_3 = -0.5324 \text{ for 0.975 quantile} \end{aligned}$$

The results of Kiefer and Vogelsang (2005) provide asymptotics that may be valid for any M, even $M = T$, but notice that Kiefer and Vogelsang (2005) do not automatically recommend using $M = \lfloor bT \rfloor$: rather, they provide alternative asymptotics for a user chosen bandwidth. So, for example, assuming $T = 128$ and $M = \left\lfloor T^{1/3} \right\rfloor = 5$, then $b = 5/128 = 0.039063$ and the 5% critical value for a two-sided test is 2.0766 instead of 1.96.

When testing assumptions about the sample mean, Kiefer and Vogelsang (2005) show in Monte Carlo simulations that the fixed-b asymptotics yields a remarkable improvement in size. However, while the empirical size improves (it gets closer to the theoretical size) as b is closer to 1, the power of the test worsens, implying that there is a size-power trade-off. These results are also confirmed analytically by Sun, Phillips and Jin (2008), who prove that the fixed-b limit distribution provides a higher-order correction.

4.2 Fixed-m asymptotics

We now consider an alternative estimate of the long run variance, a Weighted Periodogram Estimate (WPE). Letting $\lambda_j = 2\pi j/T$ for $j = 0, \pm 1, ..., \pm \lfloor T/2 \rfloor$ as the Fourier frequencies, and

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

as the periodogram of d_t, we consider estimates

$$\widetilde{\sigma}^2 = 2\pi \sum_{j=1}^{\lfloor T/2 \rfloor} K_M(\lambda_j) I(\lambda_j) \tag{8}$$

where $K_M(\lambda_j)$ is a kernel function that is symmetric and M is a bandwidth parameter.

Notice that as $\frac{1}{\sqrt{2\pi}} \sum_{t=1}^{T} \overline{d} e^{-i\lambda_j t} = \overline{d} \frac{1}{\sqrt{2\pi}} \sum_{t=1}^{T} e^{-i\lambda_j t}$ and, for $j \neq 0$, $\sum_{t=1}^{T} e^{-i\lambda_j t} = 0$, $I(\lambda_j)$ is also the periodogram of $\widehat{u}_t$ at these frequencies. Kernels $k(j/M)$ in (2) and $K_M(\lambda_j)$ in (8) are related, as $K_M(\lambda) := (2\pi)^{-1} \sum_{|l|<T} k(l/M) e^{-il\lambda}$, and the WCE in (2) has frequency domain representation

$$\int_{-\pi}^{\pi} K_M(\lambda) I^*(\lambda) d\lambda \tag{9}$$

where $I^*(\lambda)$ is the periodogram of $d_t - \overline{d}$. Weighted covariance estimation and weighted periodogram estimation are therefore very similar, and this suggests for WPE an alternative theory analogue to fixed-b for WCE.

The WPE of the long run variance using the Daniell kernel is

$$\widehat{\sigma}^2_{DAN} = 2\pi \frac{1}{m} \sum_{j=1}^{m} I(\lambda_j) \tag{10}$$

where m is a function of the bandwidth M (and, with slight abuse of notation, it is usually referred to as bandwidth itself). Regularity conditions, including $m \to \infty$, ensure that $\widehat{\sigma}^2_{DAN}$ is a consistent estimate of σ^2; for fixed m this is no longer the case,

but $\widehat{\sigma}^2_{DAN}$ is still asymptotically unbiased.

Using results from Hannan (1970), it is possible to show that, for fixed m,

$$\sqrt{T}\frac{\overline{d}-\mu}{\widehat{\sigma}_{DAN}} \rightarrow_d t_{2m}. \tag{11}$$

This result was anticipated in Sun (2013) and Müller (2014): we provide some details about the derivation of (11) in Appendix A. Monte Carlo simulations in Hualde and Iacone (2015) show that fixed-m asymptotics have the same size-power trade-off documented for fixed-b asymptotic: the smaller the value for m, the better the empirical size, but also the weaker the power.

5 A Monte Carlo study of the test for predictive accuracy under fixed-smoothing asymptotics

In this section we analyse the size and power properties of the proposed tests of equal predictive accuracy in small samples for both the case of equal predictive accuracy and the case of superior predictive accuracy of one forecasting model.

5.1 Size Analysis

We simulate forecast errors as in DM and Clark (1999). In particular, we first simulate a vector of forecast innovations from a bivariate standard normal, $(v_{1t}, v_{2t})' \sim N(0_2, I_2)$. We then introduce contemporaneous correlation by taking

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} = \begin{pmatrix} \sqrt{k} & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix}$$

and serial correlation by taking

$$\begin{aligned} e_{1t} &= \sum_{j=0}^{q} \theta^j u_{1t-j} / \sqrt{\sum_{j=0}^{q} \theta^{2j}} \\ e_{2t} &= \sum_{j=0}^{q} \theta^j u_{2t-j} / \sqrt{\sum_{j=0}^{q} \theta^{2j}} \end{aligned}$$

where $k = 1$, $\rho = 0.5$ and $\theta = 0.75$. DM, Clark (1999) and Harvey, Leybourne and Whitehouse (2015) fix q to 1, in our case instead q is set to range between 1 and 5. With this design, as q increases the processes e_{1t} and e_{2t} become similar to an AR(1) with parameter θ. Results in Clark (1999) suggest only limited sensitivity of size to ρ and θ, so we keep these fixed and investigate the effect of increasing the serial correlation with q. In Appendix B, we report a sensitivity analysis, including a size study for $\theta = 0.5$.

In Tables 1–2 we report results of the Monte Carlo, with theoretical size set to 5%. In all cases we use 10,000 replications (entries in the tables are rounded to the third decimal digit) and a quadratic loss function. We use $T = 40$ and $T = 120$ as these samples correspond to 10 years and 30 years of quarterly data, and therefore match the dimension of our sample in the empirical analysis. We consider three estimates of σ: the WCE using the DM estimate in (3) with $h - 1 = q$; the WCE using the Bartlett kernel in (5)–(6); the WPE using the Daniell kernel in (10). We refer to these three estimates as WCE-DM, WCE-B and WPE-D, respectively.

In the first part of the experiment, we study the size properties treating the estimates of σ as consistent and using standard asymptotics, i.e. the limit normal distribution, to compute the empirical size. In Table 1 we report the empirical size of the tests when the WCE-DM, WCE-B and WPE-D are used to estimate σ. When using the WCE-DM estimate, negative estimates are possible. We treat these instances as rejections of the null hypothesis. We discuss these occurrences in Appendix B.

For the WCE-B we use $M = \lfloor T^{1/3} \rfloor$ and $M = \lfloor T^{1/2} \rfloor$, and for the WPE-D we use $m = \lfloor T^{1/3} \rfloor$, $m = \lfloor T^{1/2} \rfloor$ and $m = \lfloor T^{2/3} \rfloor$. The choice of the first bandwidth for the

WCE-B is motivated by the fact that the optimal bandwidth, in minimum MSE sense, is obtained setting M proportional to $\lfloor T^{1/3} \rfloor$, see for example Newey and West (1994). We discuss here the naïve choice $M = \lfloor T^{1/3} \rfloor$, in Appendix B we also consider the automatic procedures from Newey and West (1994). The second bandwidth, $M = \lfloor T^{1/2} \rfloor$, is chosen because existing Monte Carlo evidence for fixed-b asyptotics suggests that longer bandwidths are associated with better empirical size.

As for the bandwidths for the WPE-D, Delgado and Robinson (1996), Phillips (2005) and Sun (2013) show that the optimal bandwidth, in MSE sense, is proportional to $\lfloor T^{4/5} \rfloor$, whereas Abadir, Distaso and Giraitis (2009) recommend $m = \lfloor T^{2/3} \rfloor$. However, in samples as small the ones of this exercise, even $m = \lfloor T^{2/3} \rfloor$ spans a substantial part of the interval $(0, \pi)$, and the estimate of σ with this bandwidth may therefore be subject to too much bias. The other two bandwidths are therefore chosen to limit this bias, and to allow comparison with the fixed-m asymptotics. We further explore this issue in Appendix B.

In general, Table 1 shows that, as the serial correlation increases with q, the size of the test deteriorates, although the size distortion is less serious in the larger sample. Comparing the results when WCE-B is used, on balance we find that $M = \lfloor T^{1/3} \rfloor$ yields better size properties, the only exception being for $q = 5$ in the large sample. The comparison between using the WCE-B with $M = \lfloor T^{1/3} \rfloor$ and the WCE-DM estimate is less clear cut in this instance. The DM estimate delivers better size properties in the large sample, but using the WCE with Bartlett kernel helps avoiding the very severe size distortion occurring in the small sample with $q = 4$ or $q = 5$ when the DM estimate is used.

For the WPE-D, we find that the bandwidth $m = \lfloor T^{2/3} \rfloor$ is too long for the small samples used in this investigation: the bandwidth $m = \lfloor T^{1/2} \rfloor$ yields better size in most cases, although a certain size distortion still occurs, especially in the smallest sample. Comparing the results for the three cases in which WPE-D is used, corresponding to

Table 1: Size of tests with standard asymptotics

T=40

		WCE		WPE		
q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
1	0.075	0.092	0.115	0.093	0.075	0.081
2	0.095	0.105	0.121	0.095	0.082	0.106
3	0.115	0.113	0.128	0.090	0.089	0.137
4	0.141	0.125	0.131	0.096	0.102	0.163
5	0.173	0.136	0.139	0.098	0.112	0.179

T=120

		WCE		WPE		
q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{2/3} \rfloor$
1	0.058	0.069	0.080	0.084	0.062	0.064
2	0.057	0.073	0.082	0.079	0.058	0.070
3	0.064	0.082	0.087	0.082	0.063	0.089
4	0.073	0.090	0.090	0.082	0.069	0.108
5	0.085	0.102	0.098	0.085	0.077	0.128

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size using standard normal asymptotics for various MA(q) processes with $\theta = 0.75$ and alternative estimates of the long run variance. For the WCE, DM is the WCE with the truncated kernel as in DM and $h - 1 = q$, $\lfloor T^{1/3} \rfloor$ and $\lfloor T^{1/2} \rfloor$ are the WCE with the Bartlett kernel and $M = \lfloor T^{1/3} \rfloor$ and $M = \lfloor T^{1/2} \rfloor$. For the WPE, we use the Daniell kernel with $m = \lfloor T^{1/3} \rfloor$, $m = \lfloor T^{1/2} \rfloor$ and $m = \lfloor T^{2/3} \rfloor$.

the three different bandwidths, the choice $m = \lfloor T^{1/2} \rfloor$ limits two alternative sources of size distortion: the lower order bias in the estimation of σ at higher frequencies, which affects $m = \lfloor T^{2/3} \rfloor$ most, and the high variance of the estimate, which is more a problem when the shortest bandwidth, $m = \lfloor T^{1/3} \rfloor$, is used. Bearing in mind that our focus is on small samples, the WPE estimate with bandwidth $m = \lfloor T^{1/2} \rfloor$ is overall the best choice.

In Table 2 we report results when the properties of the estimates of σ and of the test statistic are derived assuming fixed-smoothing asymptotics. In columns WCE, we use (5)–(6), with $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$, and $M = T$, and fixed-b asymptotics, with limit (7); in columns WPE, we use the estimate (10) with $m = \lfloor T^{1/4} \rfloor$, $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/2} \rfloor$ and asymptotics from (11). Bandwidths $M = \lfloor T^{1/3} \rfloor$ and $M = \lfloor T^{1/2} \rfloor$ for the WCE-B means that the same test statistic is used both in Table 1 and Table 2, and the difference in the empirical size in the two tables is then due only to the different critical values. Bandwidth $M = T$, on the other hand, has been proposed when fixed-b asymptotics is used, by Kiefer and Vogelsang (2002). Likewise, for the WPE-D estimate, bandwidths $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/2} \rfloor$ allow for a comparison with results from Table 1. The size distortion for $m = \lfloor T^{2/3} \rfloor$ documented in Table 1 is due to the bias in the estimation of the long run variance and therefore cannot be improved upon, with fixed-m asymptotics. Instead, we consider $m = \lfloor T^{1/4} \rfloor$: this is too short to be considered for standard asymptotics, as $m = 2$ when $T = 40$, but fixed-m asymptotics provides a useful justification for this choice. As the Monte Carlo exercise in Hualde and Iacone (2015) shows that the best size is achieved for the lowest bandwidths, $m = \lfloor T^{1/4} \rfloor$ is a very interesting choice.

Comparing Tables 1 and 2, we find that fixed-smoothing asymptotics always improves the empirical size, yielding results closer to the prescribed 5%. Moreover, with WCE-B the empirical size is better the larger is the bandwidth, whereas with the WPE-D the empirical size is more precise the smaller is m. Indeed, we find that the bandwidth $M = \lfloor T^{1/3} \rfloor$ in the WCE-B still yields some size distortion, even when fixed-b asymp-

Table 2: Size of tests with fixed-smoothing asymptotics

T=40

	WCE			WPE		
q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$
1	0.054	0.051	0.054	0.044	0.044	0.051
2	0.061	0.054	0.054	0.043	0.043	0.052
3	0.066	0.054	0.054	0.037	0.041	0.057
4	0.076	0.056	0.057	0.039	0.039	0.065
5	0.081	0.060	0.056	0.039	0.043	0.074

T=120

	WCE			WPE		
q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$
1	0.054	0.049	0.049	0.050	0.047	0.047
2	0.055	0.047	0.048	0.044	0.046	0.044
3	0.064	0.050	0.048	0.045	0.045	0.049
4	0.073	0.054	0.048	0.043	0.043	0.052
5	0.081	0.059	0.055	0.043	0.044	0.057

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size using fixed-smoothing asymptotics for various MA(q) processes with $\theta = 0.75$ and alternative estimates of the long run variance. For the WCE, we use the Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$ and $M = T$. For the WPE, we use the Daniell kernel with $m = \lfloor T^{1/4} \rfloor$, $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/2} \rfloor$.

totics is used; results for $m = \lfloor T^{1/2} \rfloor$ for the WPE-D are also not entirely satisfactory, especially in the T=40 sample. Overall, then, with fixed-b asyptotics it seems desirable to choose bandwidths M longer than what we would consider when standard asymptotics is used; this result is mirrored in case fixed-m asymptotics is used, in which case, the bandwidths could be shorter than what is usually recommended under standard asymptotics.

In summary, in our Monte Carlo exercise we find that the DM test with the WCE-DM may be subject to relevant size distortion in small samples, and that alternative estimates of the long run variance may help limiting this size distortion, but not completely restore the theoretical 5% size. Fixed-smoothing asymptotics alleviates the size distortion, and may eliminate it completely, when a long bandwidth is used for the WCE-B or when a short bandwidth is used for the WPE-D.

5.2 Power Analysis

In the previous exercise, we saw that some tests of equal predictive accuracy give rise to relevant size distortion, and we therefore do not recommend using those tests. To choose between the remaining tests, that are broadly correctly sized, in the second part of the Monte Carlo exercise, we study the power of the tests.

In this experiment, we only consider test statistics in which σ is estimated as the WCE-B or as WPE-D, and only use critical values from fixed-smoothing asymptotics. Notice that we also include two cases in which even the non-standard asymptotics does not completely eliminate the size distortion: when σ is estimated with $M = \lfloor T^{1/3} \rfloor$ for the WCE-B and $m = \lfloor T^{1/2} \rfloor$ for the WPE-D. In this way, we are able to observe the power loss associated to using $M = \lfloor T^{1/2} \rfloor$ for the WCE-B, instead of $M = \lfloor T^{1/3} \rfloor$. We keep $m = \lfloor T^{1/2} \rfloor$ for the WPE-D for a similar power comparison against the case in which the WPE-D with $m = \lfloor T^{1/3} \rfloor$ is used.

We test $H_0 : \{\mu = 0\}$ in processes with $\mu = cT^{-1/2}$, for c ranging between 0 and

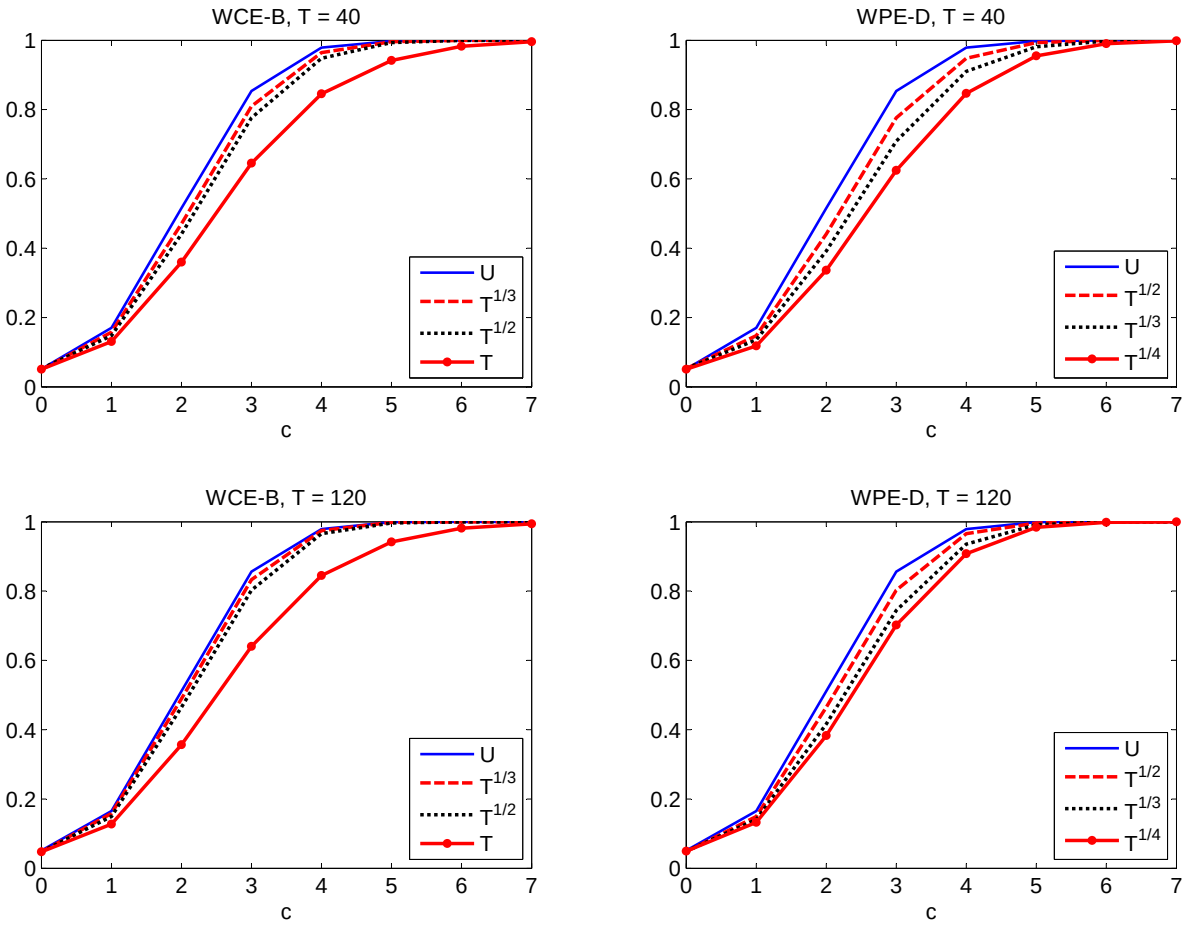
7. Since in this part of the exercise we are interested in power, rather than in size distortion, we use a time series of independent, standard normal distributed variates. As in the previous exercise, we use 10,000 repetitions and $T = 40$ and $T = 120$. We also compare the tests with fixed-smoothing asymptotics against a benchmark case in which σ is known. With samples as small as the ones used in our experiment, this benchmark is unfeasible. If a very large sample is available, this situation can be interpreted as a limit case of the test when a WCE-B with $b \to 0$ or a WPE-D with $m \to \infty$ are used, so that the replacement of σ^2 with its estimate is negligible and asymptotic normality is justified. Thus, in our experiment this benchmark should be the upper bound for the empirical power functions.

The simulated empirical power is in Figure 1. Previous simulations in Kiefer and Vogelsang (2005) and in Hualde and Iacone (2015) found that the power is higher the smaller is M or the larger is m, and our results are consistent with them. The test with statistic with known σ has the highest power, as expected. It is worth noticing, however, that the power loss due to estimating σ is minimal, especially when the WCE-B with $M = \lfloor T^{1/3} \rfloor$ or $M = \lfloor T^{1/2} \rfloor$ is used. Overall, the only case in which we observe a remarkable power loss is for $M = T$ when the WCE-B is used. For this bandwidth choice, the condition $b \to 0$ as $T \to \infty$ is certainly not justifiable so the power loss with respect to the unfeasible benchmark is not going to disappear as the sample size increases. We also verify that the power difference between using $M = \lfloor T^{1/2} \rfloor$ instead of $M = \lfloor T^{1/3} \rfloor$ for the WCE-B is very limited; to a sightly less extent, this is also true of using $m = \lfloor T^{1/2} \rfloor$ instead of $m = \lfloor T^{1/3} \rfloor$ for the WPE-D.

6 Monte Carlo comparison with the bootstrap

Bootstrap is a widely used alternative to using asymptotic approximations in tests for equal predictive ability. For this reason, in this section we perform a Monte Carlo

Figure 1: Finite sample local power



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $cT^{-1/2}$ and independent innovations. U refers to the unfeasible case in which the unknown variance is used and the test statistic has standard normal limit distribution. For the feasible tests, fixed-smoothing asymptotics is used. The alternative estimates of the long run variance are: WCE-B is for the WCE with Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$ or $M = T$; WPE-D for the WPE with Daniell kernel and $m = \lfloor T^{1/2} \rfloor$, $m = \lfloor T^{1/3} \rfloor$ or $m = \lfloor T^{1/4} \rfloor$.

analysis of the size and power of the tests for equal predictive ability using bootstrap critical values, and compare it with the results using fixed-smoothing asymptotics.

In the i-th Monte Carlo replication, we simulate forecast errors $e_{(1t)}^{(i)}$ and $e_{(2t)}^{(i)}$ as described in section 5.1 (for size analysis) or section 5.2 (for power analysis), compute the loss differential $d_t^{(i)}$ and the test statistic

$$t^{(i)} = \sqrt{T}\left(\overline{d}^{(i)}/\widehat{\sigma}^{(i)}\right).$$

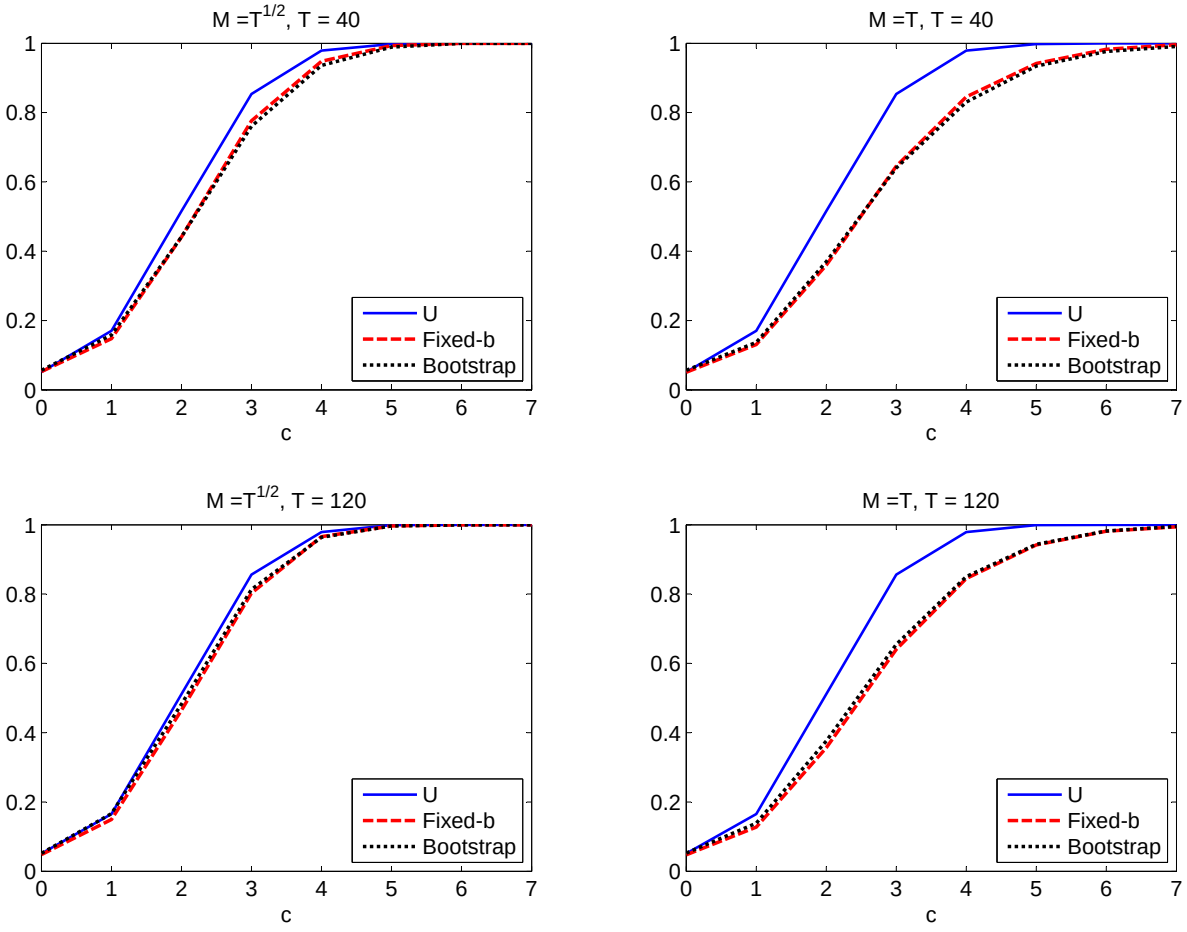
Then for each bootstrap replication b, we generate bootstrapped loss differentials $d_t^{(i,b)}$ using the overlapping stationary block-bootstrap of Politis and Romano (1994) with a circular scheme. In particular, we collate the loss differentials $(d_1^{(i)}, \ldots, d_T^{(i)}, d_1^{(i)}, \ldots, d_T^{(i)})$. We then draw block sizes $L_1, L_2, \ldots$ from a discrete uniform distribution with support on $\{1, \ldots, 2\left\lfloor T^{1/4}\right\rfloor\}$. We also draw random initial indices $I_1, I_2, \ldots$ from a discrete uniform distribution with support on $\{1, \ldots, T\}$. The series of bootstrapped loss differential $d_t^{(i,b)}$ is then given by the first T elements of $(d_{I_1}^{(i)}, \ldots, d_{I_1+L_1-1}^{(i)}, d_{I_2}^{(i)}, \ldots, d_{I_2+L_1-2}^{(i)}, \ldots)$. We finally construct the boostrapped test statistic as

$$t^{(i,b)} = \sqrt{T}\left((\overline{d}^{(i,b)} - \overline{d}^{(i)})/\widehat{\sigma}^{(i,b)}\right). \tag{12}$$

where $\overline{d}^{(i,b)}$ is the sample mean of $d_t^{(i,b)}$, and $\widehat{\sigma}^{(i,b)}$ is the estimate of its long run variance constructed using the same formula as in the original data (WCE-B or WPE-D). We perform $10,000$ bootstrap replications and use the 95% quantile of the bootstrap distribution of the test statistic, $(t^{(i,1)}, \ldots, t^{(i,10000)})$, as critical value $cv^{(i)}$. We then reject the null of equal predictive ability if $|t^{(i)}| > cv^{(i)}$. Notice that this is the naive bootstrap also performed by Kiefer and Vogelsang (2005) and Gonçalves and Vogelsang (2011) for the test with the WCE-B estimate of the long run variance using block-bootstrap.

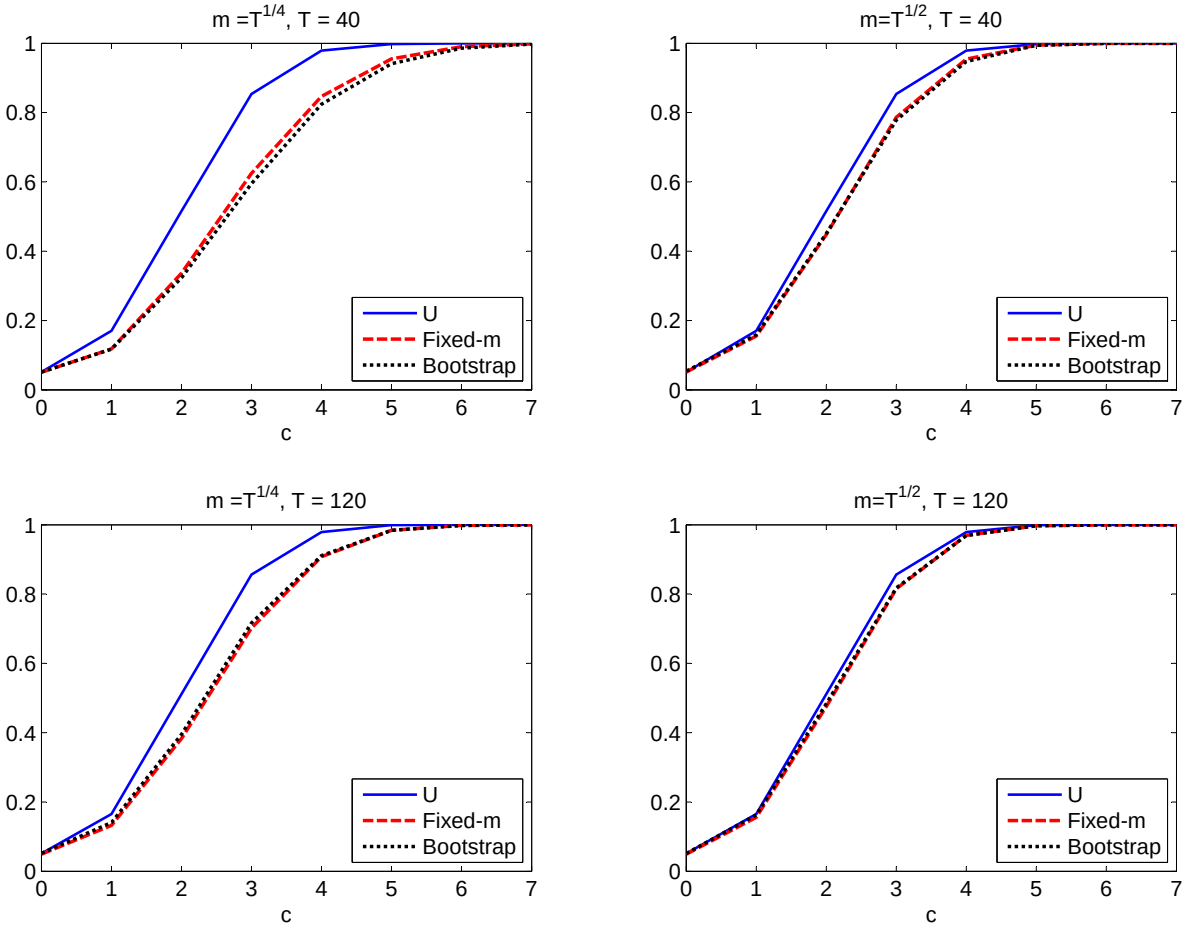
In Table 3, we report the size of tests of equal predictive ability using using bootstrap critical values for various MA processes with $\theta = 0.75$ and alternative estimates of

Figure 2: Finite sample local power: fixed-b vs bootstrap



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $cT^{-1/2}$ and independent innovations. U refers to the unfeasible case in which the unknown variance is used and the test statistic has standard normal limit distribution. For the feasible tests, fixed-b or bootstrap critical values are used. The long run variance is estimated using the WCE with Bartlett kernel with $M = \lfloor T^{1/2} \rfloor$ or $M = T$.

Figure 3: Finite sample local power: fixed-m vs bootstrap



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $cT^{-1/2}$ and independent innovations. U refers to the unfeasible case in which the unknown variance is used and the test statistic has standard normal limit distribution. For the feasible tests, fixed-m or bootstrap critical values are used. The long run variance is estimated using the WPE with Daniell kernel with $m = \lfloor T^{1/4} \rfloor$ or $m = \lfloor T^{1/2} \rfloor$.

Table 3: Size of tests with bootstrap

T=40

	WCE			WPE		
q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$
1	0.047	0.043	0.042	0.037	0.040	0.042
2	0.047	0.042	0.044	0.036	0.036	0.040
3	0.048	0.041	0.044	0.036	0.036	0.040
4	0.050	0.040	0.043	0.031	0.032	0.042
5	0.053	0.039	0.043	0.030	0.031	0.043

T=120

	WCE			WPE		
q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	T	$\lfloor T^{1/4} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$
1	0.055	0.052	0.052	0.047	0.045	0.051
2	0.051	0.045	0.045	0.041	0.045	0.045
3	0.051	0.045	0.047	0.042	0.042	0.045
4	0.053	0.045	0.046	0.040	0.042	0.044
5	0.053	0.043	0.043	0.039	0.037	0.043

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size using bootstrap critical values for various MA(q) processes with $\theta = 0.75$ and alternative estimates of the long run variance. For the WCE, we use the Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$ and $M = T$. For the WPE, we use the Daniell kernel with $m = \lfloor T^{1/4} \rfloor$, $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/2} \rfloor$.

the long run variance. For the WCE, we use the Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$ and $M = T$. For the WPE, we use the Daniell kernel with $m = \lfloor T^{1/4} \rfloor$, $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/2} \rfloor$. Results in Table 3 indicate that the bootstrap test dominates standard asymptotics and is correctly sized regardless of the choice of M (for the test using WCE) or m (for the test using WPE). In Figures 2-3, we report the finite sample local power comparison of fixed-b and fixed-m asymptotics with the bootstrap. Both figures indicate that the bootstrap local power mimics the fixed-b and the fixed-m local power.

Results for the bootstrap test with the WCE-B estimate of the long run variance are in line with Gonçalves and Vogelsang (2011). They prove that the naive block-bootstrap

has the same limiting distribution as the fixed-b asymptotic distribution. They also find that the power of the naive block-bootstrap closely follows the power when using the fixed-b critical value. However, Kiefer and Vogelsang (2005) show that the size properties of the naive block-bootstrap test statistic depends on the choice of the block length.

7 Predictive Accuracy of the SPF

To illustrate the usefulness of fixed-smoothing asymptotics for equal predictive accuracy tests, we evaluate the predictive accuracy of the Survey of Professional Forecasters' (SPF) forecasts for output growth, output inflation, the unemployment rate and the three-month Treasury bill rate against a simple random walk.

Data on the SPF are provided by the Federal Reserve Bank of Philadelphia and are available at quarterly frequency. In particular, each quarter, the SPF asks panel members to make forecasts for a set of macroeconomic indicators for the current quarter and for the following four quarters. The deadline for panel members to submit their forecasts is the middle of the quarter. We focus on median responses for the period from 1985:Q1 until 2014:Q4 and consider four evaluation periods: the full sample and three 10-year subsamples, i.e. from 1985:Q1 to 1994:Q4, from 1995:Q1 to 2004:Q4 and from 2004:Q1 to 2014:Q4.

In the SPF, the output price index is the implicit price deflator for GNP in surveys conducted prior to 1992:Q1, the implicit deflator for GDP in surveys from 1992:Q1 to 1995:Q4, and the chain-weighted price index in surveys conducted thereafter. In the same way, real output is defined as fixed-weighted real GNP in surveys conducted before 1992:Q1, fixed-weighted real GDP in surveys from 1992:Q1 to 1995:Q4, and chain-weighted real GDP in surveys conducted thereafter. Real GNP/GDP growth and GNP/GDP inflation are constructed as the annualized quarter over quarter growth rates. For both variables, we define the corresponding benchmark forecasts and realized

values accordingly, as in Stark (2010). Finally, the three-month Treasury bill rate and the unemployment rate are expressed in levels.

For all the variables considered, we use as benchmark a naive random walk, i.e. a no change benchmark using the vintages of data that were available to the public before the survey's mid-quarter deadline. In particular, for GNP/GDP inflation, the unemployment rate and the three-month Treasury bill rate, we use as benchmark a random walk on the variable. For real GNP/GDP growth, Stark (2010) finds that a no change model performs poorly, thus we use as benchmark a random walk with drift on real GNP/GDP levels and estimate the drift parameter using a rolling average of real GNP/GDP growth with a window of 60 observations.

We compare forecasts and benchmarks with forecast horizons of 0 (current quarter) to 4 (four quarters in the future) using the first release as realised value and a quadratic loss function. To evaluate the performance of the SPF against the random walk, we perform the DM test, with WCE-DM, WCE-B and WCE-D estimates of the long run variance, using standard and fixed-smoothing asymptotics.

To compute the WCE-DM, we use truncation lags equal to the forecast horizon. To select the bandwidths for the WCE-B and for the WPE-D, we use the results of our Monte Carlo exercise. For the WPE-D, we use the bandwidths $m = \lfloor T^{1/4} \rfloor$ and $m = \lfloor T^{1/3} \rfloor$, as in our Monte Carlo they always returned good size properties for the DM test when fixed-m asymptotics were used. We omit $m = \lfloor T^{1/2} \rfloor$ as we still found some evidence of size distortion in the Monte Carlo exercise, even with fixed-m asymptotics. For the WCE-B, our choice is a little bit more delicate: we omit $M = T$ in view of its low power, but we keep $M = \lfloor T^{1/3} \rfloor$, alongside $M = \lfloor T^{1/2} \rfloor$, despite some residual size distortion for the DM test even under fixed-b asymptotics, when this estimate is used. This implies that, for the WCE-B, we should put more weight on $M = \lfloor T^{1/2} \rfloor$.

Tables 4–7 report the test statistics presented in Section 3 for the null hypothesis of equal predictive accuracy of the SPF' forecasts for real GNP/GDP growth, GNP/GDP

Table 4: Real GNP/GDP Growth: SPF vs. Random Walk in level

Evaluation period: 1985.Q1 - 2014.Q4, T=120

Forecast horizon	0	1	2	3	4
WCE-DM	2.55**	1.25	0.82	0.06	-0.05
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.83*	1.32	0.90	0.06	-0.06
WCE-B, $M = \lfloor T^{1/2} \rfloor$	1.83*	1.33	0.88	0.06	-0.05
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.66	1.16	0.77	0.05	-0.05
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.74	1.30	0.86	0.05	-0.05

Evaluation period: 1985.Q1 - 1994.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	1.96**	1.12	0.82	0.71	0.56
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.54	1.20	0.89	0.77	0.60
WCE-B, $M = \lfloor T^{1/2} \rfloor$	1.56	1.20	0.90	0.71	0.58
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.46	1.14	0.91	1.10	0.57
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.40	1.06	0.77	0.74	0.65

Evaluation period: 1995.Q1 - 2004.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	1.16	-0.64	-1.64	-1.58	-1.07
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.19	-0.72	-1.80	-1.64	-1.13
WCE-B, $M = \lfloor T^{1/2} \rfloor$	1.24	-0.82	-1.71	-1.53	-1.13
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.11	-0.97	-1.45	-1.30	-1.04
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.14	-0.72	-1.64	-1.46	-1.04

Evaluation period: 2005.Q1 - 2014.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	1.81*	1.09	1.01	0.60	0.51
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.32	1.17	1.12	0.52	0.47
WCE-B, $M = \lfloor T^{1/2} \rfloor$	1.29	1.18	1.16	0.59	0.50
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.09	1.01	1.01	0.54	0.44
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.12	1.02	1.01	0.53	0.42

Note: this table reports the predictive accuracy tests for the SPF forecasts of real GNP/GDP growth with respect to a random walk with drift on GNP/GDP levels. GNP/GDP growth is defined as the annualized quarter over quarter growth rates of fixed-weighted real GNP in the surveys conducted before 1992:Q1, fixed-weighted real GDP in the surveys from 1992:Q1 to 1995:Q4, and chain-weighted real GDP in the surveys thereafter. Random walk predictions and realized values are computed accordingly. The drift parameter is estimated using a rolling window of 60 observations. ** and * indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics for WCE-DM, fixed-b asymptotics for WCE-B and fixed-m asymptotics for WPE-D. ■ and ■ indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics and limit normality.

Table 5: GNP/GDP Inflation: SPF vs. Random Walk

Evaluation period: 1985.Q1 - 2014.Q4, T=120

Forecast horizon	0	1	2	3	4
WCE-DM	4.20**	3.89**	2.27**	0.73	1.67*
WCE-B, $M = \lfloor T^{1/3} \rfloor$	3.28**	3.65**	2.39**	0.72	1.59
WCE-B, $M = \lfloor T^{1/2} \rfloor$	2.89**	3.71**	2.38**	0.62	1.57
WPE-D, $m = \lfloor T^{1/4} \rfloor$	2.59**	3.23**	1.81	0.50	1.24
WPE-D, $m = \lfloor T^{1/3} \rfloor$	2.69**	3.53**	2.07*	0.55	1.37

Evaluation period: 1985.Q1 - 1994.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	2.55**	3.08**	1.45	-0.66	0.38
WCE-B, $M = \lfloor T^{1/3} \rfloor$	2.73**	3.09**	1.50	-0.52	0.35
WCE-B, $M = \lfloor T^{1/2} \rfloor$	3.08**	3.11**	1.54	-0.50	0.35
WPE-D, $m = \lfloor T^{1/4} \rfloor$	3.59**	2.61*	1.82	-0.42	0.28
WPE-D, $m = \lfloor T^{1/3} \rfloor$	4.10**	2.63**	1.38	-0.45	0.30

Evaluation period: 1995.Q1 - 2004.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	1.20	1.17	0.56	0.29	0.32
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.04	1.19	0.59	0.33	0.38
WCE-B, $M = \lfloor T^{1/2} \rfloor$	0.94	1.16	0.56	0.31	0.36
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.01	1.16	0.50	0.28	0.33
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.04	1.21	0.54	0.30	0.31

Evaluation period: 2005.Q1 - 2014.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	3.27**	2.62**	1.89*	1.44	4.84**
WCE-B, $M = \lfloor T^{1/3} \rfloor$	2.42**	2.37**	2.04*	1.42	2.25**
WCE-B, $M = \lfloor T^{1/2} \rfloor$	2.35*	2.63**	2.05*	1.27	2.89**
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.85	2.59*	1.91	0.99	3.07**
WPE-D, $m = \lfloor T^{1/3} \rfloor$	2.07*	2.25*	1.70	1.20	2.97**

Note: this table reports the predictive accuracy tests for the SPF forecasts of GNP/GDP inflation with respect to a random walk. GNP/GDP inflation is defined as the implicit price deflator for GNP in surveys conducted prior to 1992:Q1, the implicit deflator for GDP in the surveys from 1992:Q1 to 1995:Q4, and the chain-weighted price index in the surveys thereafter. Random walk predictions and realized values are computed accordingly. ** and * indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics for WCE-DM, fixed-b asymptotics for WCE-B and fixed-m asymptotics for WPE-D. and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics and limit normality.

Table 6: Unemployment Rate: SPF vs. Random Walk

Evaluation period: 1985.Q1 - 2014.Q4, T=120

Forecast horizon	0	1	2	3	4
WCE-DM	3.77**	2.08**	1.89*	1.94*	2.14**
WCE-B, $M = \lfloor T^{1/3} \rfloor$	2.42**	1.98*	2.06*	2.22**	2.54**
WCE-B, $M = \lfloor T^{1/2} \rfloor$	2.31**	1.91*	1.95*	2.07*	2.30**
WPE-D, $m = \lfloor T^{1/4} \rfloor$	2.12*	1.80	1.82	1.87	2.03*
WPE-D, $m = \lfloor T^{1/3} \rfloor$	2.09*	1.79	1.79	1.87*	2.05*

Evaluation period: 1985.Q1 - 1994.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	3.24**	1.64	1.65*	2.00**	2.70**
WCE-B, $M = \lfloor T^{1/3} \rfloor$	2.69**	1.75	1.85*	2.18*	2.58**
WCE-B, $M = \lfloor T^{1/2} \rfloor$	2.82**	1.82	1.96	2.42**	2.89**
WPE-D, $m = \lfloor T^{1/4} \rfloor$	3.42**	1.73	2.09	2.92**	3.74**
WPE-D, $m = \lfloor T^{1/3} \rfloor$	2.32*	1.54	1.64	1.89	2.03*

Evaluation period: 1995.Q1 - 2004.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	2.03**	1.72*	1.26	1.11	1.04
WCE-B, $M = \lfloor T^{1/3} \rfloor$	2.00*	1.73	1.43	1.32	1.28
WCE-B, $M = \lfloor T^{1/2} \rfloor$	2.10*	1.71	1.35	1.25	1.18
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.80	1.47	1.16	1.14	1.04
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.86	1.50	1.17	1.05	0.99

Evaluation period: 2005.Q1 - 2014.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	2.84**	1.66*	1.57	1.69*	1.90*
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.86*	1.66	1.78	1.97*	2.27**
WCE-B, $M = \lfloor T^{1/2} \rfloor$	1.81	1.61	1.71	1.88	2.12*
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.50	1.32	1.39	1.54	1.72
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.58	1.38	1.47	1.63	1.85

Note: this table reports the predictive accuracy tests for the SPF forecasts of the unemployment rate with respect to a random walk. ** and * indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics for WCE-DM, fixed-b asymptotics for WCE-B and fixed-m asymptotics for WPE-D. and indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics and limit normality.

Table 7: Three-month Treasury Bill: SPF vs. Random Walk

Evaluation period: 1985.Q1 - 2014.Q4, T=120

Forecast horizon	0	1	2	3	4
WCE-DM	5.53**	4.26**	3.48**	2.35**	1.35
WCE-B, $M = \lfloor T^{1/3} \rfloor$	4.29**	4.31**	3.72**	2.48**	1.48
WCE-B, $M = \lfloor T^{1/2} \rfloor$	4.46**	4.46**	4.01**	2.61**	1.50
WPE-D, $m = \lfloor T^{1/4} \rfloor$	4.89**	5.08**	3.66**	2.21*	1.38
WPE-D, $m = \lfloor T^{1/3} \rfloor$	3.97**	4.44**	3.99**	2.55**	1.48

Evaluation period: 1985.Q1 - 1994.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	5.61**	3.67**	3.31**	1.34	0.69
WCE-B, $M = \lfloor T^{1/3} \rfloor$	5.02**	4.12**	3.28**	1.33	0.75
WCE-B, $M = \lfloor T^{1/2} \rfloor$	5.87**	4.75**	3.96**	1.32	0.72
WPE-D, $m = \lfloor T^{1/4} \rfloor$	9.34**	5.70**	6.21**	1.35	0.71
WPE-D, $m = \lfloor T^{1/3} \rfloor$	4.28**	3.24**	3.56**	1.57	0.83

Evaluation period: 1995.Q1 - 2004.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	2.63**	1.92*	1.46	1.23	0.56
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.94*	1.88*	1.58	1.20	0.40
WCE-B, $M = \lfloor T^{1/2} \rfloor$	1.83	1.78	1.59	1.35	0.49
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.55	1.53	1.44	1.50	0.63
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.58	1.53	1.38	1.13	0.40

Evaluation period: 2005.Q1 - 2014.Q4, T=40

Forecast horizon	0	1	2	3	4
WCE-DM	2.23**	2.12**	1.97**	1.59	1.08
WCE-B, $M = \lfloor T^{1/3} \rfloor$	1.93*	2.21**	2.11*	1.92*	1.46
WCE-B, $M = \lfloor T^{1/2} \rfloor$	1.81	2.08*	1.87	1.68	1.23
WPE-D, $m = \lfloor T^{1/4} \rfloor$	1.65	2.17*	1.82	1.56	1.06
WPE-D, $m = \lfloor T^{1/3} \rfloor$	1.63	1.96*	1.73	1.54	1.13

Note: this table reports the predictive accuracy tests for the SPF forecasts of the three-month Treasury Bill rate with respect to a random walk. ** and * indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics for WCE-DM, fixed-b asymptotics for WCE-B and fixed-m asymptotics for WPE-D. ■ and ■ indicate, respectively, two-sided significance at the 5% and 10% level using standard asymptotics and limit normality.

inflation, the unemployment rate and the three-month T-Bill rates with respect to the random walk. We denote by $e^h_{1,t}$ the h-steps ahead forecast error of the random walk and by $e^h_{2,t}$ the h-steps ahead forecast error from the SPF. Therefore, a positive entry in the tables means higher loss for the forecast made using the random walk, and viceversa for a negative entry. In the tables, we use asterisks to indicate two-sided significance using standard asymptotics for WCE-DM, fixed-b asymptotics for WCE-B and fixed-m asymptotics for WPE-D. We also use shades of gray to indicate two-sided significance using standard asymptotics and limit normality. This implies that for the WCE-DM the asterisks and the shades of grey coincide.

Results in Tables 4–7 show that overall the predictive ability of the SPF is stronger than the one of the random walk for the three-month Treasury bill rate and GNP/GDP inflation but not for real GNP/GDP growth and the unemployment rate. The tables also indicate that the subsample 1985.Q1 to 1994.Q4 is characterised by a strong predictive ability of the SPF with respect to the random walk, but this predictive ability sharply declined in the most recent subsample.

In particular, Table 4 shows that the SPF's forecasts for real GNP/GDP growth do not in general outperform the random walk. For the current quarter, the test with WCE-DM indicates significant outperformance of the SPF, but the tests with fixed-b and fixed-m asymptotics do not support this result, especially when looking at the three subsamples. As for GNP/GDP price inflation, Table 5 shows a much stronger predictive ability of the SPF, especially for short horizons and in the first and the last subsamples. Results in Table 6 indicate some predictive ability of the SPF's forecasts for the unemployment rate, but the evidence is much weaker when using the proposed tests with fixed-smoothing asymptotics. Finally, Table 7 provides strong evidence of superior predictive accuracy of the SPF's forecasts for the three month Treasury bill rate with respect the random walk, especially for short horizons. However, the predictive ability of the SPF for the three month Treasury bill rate sharply declined in the last two

subsamples.

As for the comparison of standard asymptotics with the fixed-smoothing asymptotics used in this paper, the tables show that the tests with standard asymptotics are more likely to reject the null of equal predictive ability than the tests that use fixed-smoothing asymptotics, especially in the subsamples (see for example the bottom panels in Tables 6-7). This is due to the fact that in the subsamples the tests are performed only on 40 observations, exacerbating the size distortions induced by standard asymptotics, see Table 1. For example, Table 6 shows that for inflation both the test with WCE-DM and test the with WCE-B and standard asymptotics reject at 10% significance level the null of equal predictive ability of the SPF and the random walk on the last subsample for almost all forecasting horizons. This could be interpreted as a clear indication of predictive ability of the SPF for the unemployment rate. However, the tests with fixed-smoothing asymptotics fail to reject the null of equal predictive ability for almost all forecasting horizons, especially when fixed-m asymptotics is used, indicating that the SPF did not have any significant predictive ability for the unemployment rate in this period.

8 Conclusion

We propose fixed-smoothing asymptotics to overcome the small sample size distortions of standard tests for predictive accuracy. Our Monte Carlo results show that these alternative asymptotics provide correctly sized tests for autocorrelated loss differentials even when only a small number of out of sample observations are available.

The methodology proposed in this paper is well-suited to evaluate the predictive accuracy of surveys with limited samples. As an illustrative example, and to facilitate comparison with other works, we apply our methodology to reassess the predictive accuracy of the Survey of Professional Forecasters (SPF). Other interesting applications

may include the ECB survey of professional forecasters, which has a short time series dimension and is thus well-suited for our setup.

In this paper, we focus on applying the fixed-b and fixed-m asymptotics to the Diebold and Mariano (1995) test. However, these methodologies are of broader applicability in the forecasting literature. For example, Harvey, Leybourne and Whitehouse (2015) apply the fixed-m approach to forecast encompassing tests. Future work includes applications of fixed-smoothing asymptotics to tests of equal predictive ability in presence of parameter estimation errors, see West (1996) and Clark and McCracken (2001); to forecast rationality tests, see Granger and Newbold (1986) and Diebold and Lopez (1996); forecast breakdown tests, see Giacomini and Rossi (2009); and forecast comparison in unstable environments, Giacomini and Rossi (2010).

A Limiting fixed-m asymptotics

Let $x_t = \mu + u_t$, with $u_t = \sum_{l=0}^{\infty} A_l \varepsilon_{t-l}$ where ε_t is an independent, identically distributed process with $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = 1$, $E(\varepsilon_t^4) < \infty$, and $\sum_{l=0}^{\infty} j^{1/2} |A_l| < \infty$. Define the Fourier frequencies $\lambda_j = 0, \pm 1, ..., \lfloor T/2 \rfloor$ and the Fourier transform $w_x(\lambda) = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} x_t e^{i\lambda t}$, the periodogram $I_x(\lambda) = |w_x(\lambda)|^2$, the sample mean $\overline{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$ and the statistic $\tau = \frac{\overline{x} - \mu_0}{\sqrt{2\pi \frac{1}{m} \sum_{j=1}^{m} I_x(\lambda_j)}}$; then, under $H_0 : \{\mu = \mu_0\}$, as $T \to \infty$,

$$\tau \to_d t_{2m} \tag{13}$$

Proof. First, note that, for $j = 1, ..., m$, $\frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} e^{i\lambda_j t} = 0$, so $w_x(\lambda_j) = w_u(\lambda_j)$. Moreover, following Hannan (1970), page 247,

$$w_u(\lambda) = \left(\sum_{l=0}^{\infty} A_l e^{i\lambda l} \right) w_\varepsilon(\lambda) + r_T(\lambda)$$

where $r_T(\lambda) = o_p(1)$ uniformly in λ, so

$$w_u(\lambda_j) = \left(\sum_{l=0}^{\infty} A_l e^{i\lambda_j l} \right) w_\varepsilon(\lambda_j) + o_p(1) \tag{14}$$

Now let

$$\begin{aligned} s_T^2 &= \frac{1}{2\pi T} \sum_{t=1}^{T} \cos^2\left(\frac{2\pi j t}{T}\right) \\ \eta_{t,T} &= \frac{1}{\sqrt{2\pi T}} \varepsilon_t \cos\left(\frac{2\pi j t}{T}\right) \\ z_{t,T} &= s_T^{-1} \eta_{t,T} \end{aligned}$$

then sufficient conditions for the central limit theorem are that

$$
\begin{aligned}
&E\left(z_{t,T}\right)=0\ \forall t,T\\
&\sum_{t=1}^{T}V\left(z_{t,T}\right)=1\ \forall t,T\\
&z_{t,T}\text{ independent from }z_{s,T}\ \forall t,s,\ \forall T\\
&\sum_{t=1}^{T}E\left|z_{t,T}\right|^{2+\delta}\to 0\text{ for some }\delta>0
\end{aligned}
$$

The first three conditions are easy to establish; the Liapunov condition can be easily verified for $\delta=1$, noting that $E\left|\varepsilon_t\right|^3$ exists because $E\left(\varepsilon_t^4\right)<\infty$. Thus,

$$\sum_{t=1}^{T}z_{t,T}\to_d N\left(0,1\right)$$

i.e.,

$$
\begin{aligned}
\left(\frac{1}{2\pi T}\sum_{t=1}^{T}\cos^2\left(\frac{2\pi jt}{T}\right)\right)^{-1/2}\sum_{t=1}^{T}\frac{1}{\sqrt{2\pi T}}\varepsilon_t\cos\left(\frac{2\pi jt}{T}\right) &\to {}_dN\left(0,1\right)\\
\sum_{t=1}^{T}\frac{1}{\sqrt{2\pi T}}\varepsilon_t\cos\left(\frac{2\pi jt}{T}\right) &\to {}_dN\left(0,\frac{1}{2\pi}1/2\right)
\end{aligned}
$$

where we also used $\frac{1}{T}\sum_{t=1}^{T}\cos^2\left(\frac{2\pi jt}{T}\right)=\frac{1}{2}$ from Gradshteyn and Ryzhik (1994), equation (1.351.2), page 37, and

$$\left(1/2\frac{1}{2\pi}\right)^{-1}\left(\sum_{t=1}^{T}\frac{1}{\sqrt{2\pi T}}\eta_t\cos\left(\frac{2\pi jt}{T}\right)\right)^2\to_d\chi_1^2$$

The term $\frac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T}\varepsilon_t\sin\left(\frac{2\pi jt}{T}\right)$ may be discussed in the same way. The covariance of

$\sum_{t=1}^{T} \frac{1}{\sqrt{2\pi T}} \varepsilon_t \cos\left(\frac{2\pi jt}{T}\right)$ and $\sum_{t=1}^{T} \frac{1}{\sqrt{2\pi T}} \varepsilon_t \sin\left(\frac{2\pi jt}{T}\right)$ is

$$\frac{1}{2\pi T} \sum_{t=1}^{T} \sin\left(\frac{2\pi jt}{T}\right) \cos\left(\frac{2\pi jt}{T}\right) = 0$$

using Gradshteyn and Ryzhik (1994), equation (1.333.1), page 35, and then equation (1.342.1), page 36. Then, the joint convergence of $\sqrt{2\pi}\sqrt{2} \sum_{t=1}^{T} \frac{1}{\sqrt{2\pi T}} \varepsilon_t \cos\left(\frac{2\pi jt}{T}\right)$ and $\sqrt{2\pi}\sqrt{2} \sum_{t=1}^{T} \frac{1}{\sqrt{2\pi T}} \varepsilon_t \sin\left(\frac{2\pi jt}{T}\right)$ to a bivariate vector of independently normally distributed random variables with diagonal covariance matrix follows from an application of the Cramer-Wold device. Therefore,

$$2(2\pi) I_{\varepsilon}(\lambda_j) \rightarrow_d \chi^2_2.$$

Moreover, using

$$\sum_{t=1}^{T} e^{it(\lambda_j - \lambda_k)} = 0 \text{ for } j \neq k \tag{15}$$

for integers j, k such that $\lambda_j \in [0, \pi]$ and $\lambda_k \in [0, \pi]$, then, following Giraitis, Koul and Surgalis (2012), page 112, the formula (15) yields $E\left(w_{\varepsilon}(\lambda_j) w_{\varepsilon}(\lambda_k)^*\right) = 0$ for $j \neq k$; therefore, with an application of the Cramer Wold device, it is easy to conclude that

$$2\pi \frac{1}{m} \sum\nolimits_{j=1}^{m} I_{\varepsilon}(\lambda_j) \rightarrow_d C^2_{2m}/(2m)$$

where $C^2_{2m}/(2m)$ is a χ^2_{2m} distributed random variable divided by the number of degrees of freedom. Using (14) and $\sum_{l=0}^{\infty} A_l e^{i\lambda_j l} \rightarrow \sum_{l=0}^{\infty} A_l = \sigma$, it also follows that

$$2\pi \frac{1}{m} \sum\nolimits_{j=1}^{m} I_x(\lambda_j) \rightarrow_d \sigma^2 C^2_{2m}/(2m).$$

Finally, as in Phillips and Solo (1992), we use the Beveridge Nelson decomposition

$$u_t = \left(\sum_{l=0}^{\infty} A_l\right) \varepsilon_t + \widetilde{\varepsilon}_{t-1} - \widetilde{\varepsilon}_t$$

where

$$\widetilde{\varepsilon}_t = \sum_{l=0}^{\infty} \widetilde{A}_l \varepsilon_{t-l},\ \widetilde{A}_l = \sum_{k=l+1}^{\infty} A_k$$

and

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} u_t = \left(\sum_{l=0}^{\infty} A_l\right) \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t + \frac{1}{\sqrt{T}} \left(\widetilde{\varepsilon}_0 - \widetilde{\varepsilon}_T\right) \tag{16}$$

where $\frac{1}{\sqrt{T}} \left(\widetilde{\varepsilon}_0 - \widetilde{\varepsilon}_T\right) = o_p(1)$ as on Phillips and Solo (1992) page 978. In view of Remark 3.5 of Phillips and Solo (1992), the condition on the weights A_l is $\sum_{l=0}^{\infty} \widetilde{A}_l^2 < \infty$, as in equation (16) of Phillips and Solo (1992), and this is implied by $\sum_{l=0}^{\infty} l^{1/2} |A_l| < \infty$, Phillips and Solo (1992) page 973, so,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} u_t = \sigma \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t + o_p(1) \to_d N\left(0, \sigma^2\right).$$

Another application of the Cramer Wold device and of (15) allows to establish a central limit theorem for the vectors $\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t, \sqrt{2\pi}\sqrt{2} \sum_{t=1}^{T} \frac{1}{\sqrt{2\pi T}} \varepsilon_t \cos\left(\frac{2\pi j t}{T}\right)\right)'$ for integer $0 < j < m$ and conclude that $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t$ is asymptotically independent from $\sqrt{2\pi}\sqrt{2} \sum_{t=1}^{T} \frac{1}{\sqrt{2\pi T}} \varepsilon_t \cos\left(\frac{2\pi j t}{T}\right)$; asymptotic independence between $\sqrt{2\pi}\sqrt{2} \sum_{t=1}^{T} \frac{1}{\sqrt{2\pi T}} \varepsilon_t \sin\left(\frac{2\pi j t}{T}\right)$ and $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t$ is established in the same way. Therefore, (13) holds.

Remark. Condition $\sum_{l=0}^{\infty} l^{1/2} |A_l| < 0$ is fairly common in the literature, and it holds for any ARMA model. Many of these results are already known in the literature. For example, the limit normality for the Fourier transform is given in Hannan (1970) in page 225, also see Kokoszka and Mikosch (2000) page 51, where the asymptotic independence of the periodograms $I_{\varepsilon}(\lambda_j)$ at different frequencies is also discussed. A result similar to (13) is also in Sun (2013). The main reason of interest for this proof is then in the fact that, using the decompositions (14) and (16) we see that can treat most weakly

dependent processes as independent processes, and derive results from the latter ones. These results are then fairly intuitive and easy to establish.

B Additional Monte Carlo results

In this appendix, we report additional Monte Carlo results that include the frequency of negative estimates for the long run variance using the WCE-DM, the size of standard asymptotics for $\theta = 0.5$, for the WCE-B with automatic bandwidth selection and for the WPE with feasible minimum MSE bandwidth.

B.1 Negative estimates of the long run variance

In Table 8, we study the frequency of negative estimates for $\widehat{\sigma}^2_{DM}$, the WCE estimate with the rectangular kernel (WCE-DM) defined in (3). Table 8 shows that the risk of negative long-run variance estimates is higher in the small sample, at large forecasting horizons and for low values of θ. For $\theta = 0$, $q = 5$ and $T = 40$, the size distortion due just to a negative estimate $\widehat{\sigma}^2_{DM} < 0$ is actually larger than the nominal size.

Table 8: Frequency of negative estimates for the long run variance

	T=40			T=120		
q	$\theta = 0.00$	$\theta = 0.50$	$\theta = 0.75$	$\theta = 0.00$	$\theta = 0.50$	$\theta = 0.75$
1	0.001	0.000	0.000	0.000	0.000	0.000
2	0.005	0.001	0.000	0.000	0.000	0.000
3	0.014	0.007	0.003	0.000	0.000	0.000
4	0.033	0.017	0.007	0.000	0.000	0.000
5	0.060	0.037	0.019	0.002	0.001	0.000

Note: frequency of negative estimates of the long run variance using the WCE estimator with the truncated kernel as in DM for various MA(q) processes.

B.2 Sensitivity to θ

In Table 9, we study the size properties of the DM test for various estimates of σ when $\theta = 0.5$ instead, assuming standard asymptotics. This exercise allows a comparison with Table 1 in which $\theta = 0.75$ was used, to appreciate the consequences of altering θ. Consistently with results in Clark (1999), the size when the WCE-DM is used does not seem to be sensitive to the different value of θ; on the other hand, the reduction in the dependence is associated with a slight improvement in the size properties when the WCE with Bartlett kernel (WCE-B) or the WPE with Daniell kernel (WPE-D) is used. Overall, in the case $\theta = 0.5$ the evidence that the test with statistic standardized by the WPE-D estimate (with $m = \lfloor T^{1/2} \rfloor$) gives best size is even more compelling.

Table 9: Size of tests with standard asymptotics for $\theta = 0.5$

	T=40			T=120		
q	WCE-DM	WCE-B	WPE-D	WCE-DM	WCE-B	WPE-D
1	0.077	0.085	0.075	0.059	0.068	0.061
2	0.099	0.090	0.076	0.058	0.066	0.060
3	0.124	0.096	0.078	0.068	0.072	0.062
4	0.157	0.097	0.081	0.074	0.075	0.062
5	0.196	0.097	0.080	0.086	0.075	0.065

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size using standard normal asymptotics for various MA(q) processes and alternative estimates of the long run variance: WCE-DM is for the WCE with the truncated kernel as in DM, WCE-B is for the Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$, and WPE-D for the WPE with Daniell kernel and $m = \lfloor T^{1/2} \rfloor$.

B.3 Automatic bandwidth selection

In Table 10, we study the application of the automatic bandwidth selection of Newey and West (1994), when $\theta = 0.75$. We compare the performance for the naïve $M = \lfloor T^{1/3} \rfloor$ bandwidth (already available in Table 1) against the NW estimate with prewhitening as in Newey and West (1994), and against a third estimate in which the same procedure is applied, but without prewhitening. In general, using the NW estimate without

prewhitening does not yield size as good as when the naïve $M = \lfloor T^{1/3} \rfloor$ estimate is employed: the prewithening on the other hand does provide some size correction, but the better size for larger q is mostly offset by worse size when $q = 1$: this suggests that the automatic NW procedure would not fare well when the dependence is relatively weak, and actually size properties deteriorating for larger θ are documented also in Clark (1999). Table 10 therefore shows that even the automatic bandwidth selection with prewhitening from Newey and West (1994) does not offer a complete correction of the size distortion, when standard asymptotics is used.

Table 10: Automatic bandwidth selection for WCE-B with standard asymptotics

	T=40			T=120		
q	$\lfloor T^{1/3} \rfloor$	Prew	No Pre	$\lfloor T^{1/3} \rfloor$	Prew	No Pre
1	0.092	0.129	0.125	0.069	0.074	0.079
2	0.105	0.122	0.130	0.073	0.066	0.082
3	0.113	0.110	0.129	0.082	0.067	0.085
4	0.125	0.107	0.135	0.090	0.065	0.092
5	0.136	0.108	0.140	0.102	0.068	0.096

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size using standard normal asymptotics for various MA(q) processes with $\theta = 0.75$ and alternative bandwidths for the WCE using the Bartlett kernel: $\lfloor T^{1/3} \rfloor$, the Newey and West (1994) estimate with prewhitening (Prew) and the Newey and West (1994) against the same procedure without prewhitening (No Pre).

B.4 Minimum MSE bandwidth

In Table11 we report the empirical size of equal predictive ability tests when the WPE estimate of the long run variance with feasible minimum MSE bandwidth is used.

To derive the minimum MSE bandwidth, we follow Phillips (2005) and Sun (2013). For the average periodogram with bandwidth m, the bias is

$$\text{Bias} = \left(\frac{m}{T}\right)^2 B, \quad \text{where } B = -\frac{\pi^2}{6} \sum_{j=-\infty}^{\infty} j^2 \gamma_j.$$

Using the fact that $\frac{2\pi I(\lambda_j)}{\sigma^2} \to_d \frac{1}{2}\chi^2_2$, $Var\left(\frac{2\pi I(\lambda_j)}{\sigma^2}\right) \to \frac{2\times 2}{4} = 1$ then for fixed m

$$Var\left(\frac{\frac{1}{m}\sum_{j=1}^{m} 2\pi I\left(\lambda_j\right)}{\sigma^2}\right) \to \frac{1}{m}$$

and the asymptotic MSE is $\frac{m^4}{T^4}B^2 + \frac{1}{m}\sigma^4$. Thus, $\frac{\partial}{\partial m}\left(\frac{m^4}{T^4}B^2 + \frac{1}{m}\sigma^4\right) = \left(4\frac{m^3}{T^4}B^2 - \frac{1}{m^2}\sigma^4\right)$ and from $4\frac{m^3}{T^4}B^2 - \frac{1}{m^2}\sigma^4 = 0$ we get $4\frac{m^5}{T^4}B^2 = \sigma^4$ and $m_{MSE} = T^{4/5}\left(\frac{\sigma^4}{4B^2}\right)^{1/5}$.

The bias factor B is usually unknown, but when $u_t = \phi u_{t-1} + \varepsilon_t$ with $|\phi| < 1$ and ε_t iid$(0, \omega)$, then $\sigma^2 = \frac{\omega^2}{(1-\phi)^2}$ and $B = -\frac{\pi^2}{6}\frac{2\phi}{(1-\phi)^4}\omega^2$, so we approximated σ^4/B^2 with a common plug in method: we assume such AR(1) model, estimate ϕ and then replace the estimated value in the formula for m_{MSE}.

Finally the feasible MSE bandwidth $\widehat{m}_{MSE}$ is given by the integer part of m_{MSE}, when this is between 1 and $T/2$, and by 1 or $T/2$ otherwise.

Result in Table 11 indicate that, as for the NW automatic bandwidth selection, the test is oversized both when standard and fixed-m asymptotics are used. This is due to the fact that the feasible minimum MSE bandwidth is longer than $\lfloor T^{1/4} \rfloor$, $\lfloor T^{1/3} \rfloor$ or $\lfloor T^{1/2} \rfloor$ used in Table 2, resulting in a larger bias.

Table 11: Size of tests with minimum MSE bandwidth

	T=40		T=120	
q	Standard	Fixed-m	Standard	Fixed-m
1	0.083	0.071	0.072	0.066
2	0.095	0.075	0.068	0.061
3	0.104	0.081	0.074	0.066
4	0.115	0.087	0.082	0.073
5	0.121	0.092	0.092	0.078

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size for various MA(q) processes with $\theta = 0.75$ using the WPE estimator of the long run variance and the feasible minimum MSE bandwidth. The test with standard asymptotics uses standard normal critical values and the test with fixed-m asymptotics uses critical values from a t_{2m}, where m is the feasible minimum MSE bandwidth.

C References

Abadir, K. M., W. Distaso, and L. Giraitis, 2009. Two estimators of the long-run variance: beyond short memory, *Journal of Econometrics* 150, 56-70.

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

Clark, T. E., and M. W, McCracken, 2001. Tests of equal forecast accuracy and encompassing for nested models, *Journal of Forecasting* , 105, 85-110.

Clark, T. E., and M. W. McCracken, 2013. Advances in Forecast Evaluation, in Handbook of Economic Forecasting (Vol. 2), eds. G. Elliott and A. Timmerman, Amsterdam: Elsevier, 1107-1201.

Delgado, M. A., and P. M. Robinson, 1996. Optimal spectral bandwidth for long memory, *Statistica Sinica* 6, 97-112.

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

Diebold, F. X. and J. A. Lopez, 1996. Forecast Evaluation and Combination, in G. S. Maddala and C. R. Rao (eds.), Handbook of Statistics, Vol. 14 (Amsterdam: North-Holland) 241268.

Diebold, F. X., and R. S. Mariano, 1995. Comparing predictive accuracy, *Journal of Business and Economic Statistics*, 20, 134-144.

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

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

Giacomini, R., and H. White, 2006, Tests of conditional predictive ability, *Econometrica*, 74, 1545-1578.

Gonçalves, S., and T. J. Vogelsang, 2011, Block bootstrap HAC robust tests: the sophistication of the naive bootstrap, *Econometric Theory*, 27:4, 745791.

Gradshteyn, I.S., and I. M. Ryzhik, 1994. Table of Integrals, Series, and Products, Fifth Edition. Academic Press, Boston.

Granger, C. W. J., and P. Newbold, 1986. Forecasting Economic Time Series, 2nd edition (Academic Press).

Hannan, E. J., 1970. *Multiple time series,* Wiley, New York.

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

Harvey, D., S. J. Leybourne, and E. Whitehouse, 2015. Testing forecast accuracy in small samples, Mimeo, University of Nottingham.

Hualde, J., and F. Iacone, 2015. Autocorrelation robust inference using the Daniell kernel with fixed bandwidth, Discussion Papers in Economics No. 15/14, Department of Economics, University of York.

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

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

Kokoszka, P., and T. Mikosch, 2000. The periodogram at the Fourier frequencies, *Stochastic Processes and their Applications*, 86, 49–79.

Li, J., and A. J. Patton, 2013, Asymptotic inference about predictive accuracy using high frequency data, working paper, Duke University.

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

Newey, W. K. and K. D. West, 1994. Automatic lag selection in covariance matrix estimation, *Review of Economic Studies*, 61, 631-653.

Patton, A. J., 2015. Comment, *Journal of Business and Economic Statistics*, 33, 22-24.

Phillips, P. C. B., 2005. HAC estimation by automated regression, *Econometric Theory*, 21, 116-142.

Phillips, P. C. B., and V. Solo, 1992. Asymptotics for linear processes, *Annals of Statistics*, 20, 971-1001.

Politis, D. N., and J. P. Romano, 1994. The Stationary Bootstrap, *Journal of the American Statistical Association*, 89, 1303-1313.

Stark, T., 2010. Realistic evaluation of real-time forecasts in the survey of professional forecasters. Research Rap Special Report, Federal Reserve Bank of Philadelphia.

Sun, Y., 2013. A heteroskedasticity and autocorrelation robust F test using orthonormal series variance estimator, *Econometrics Journal*, 16, 1-26.

Sun, Y., 2014a. Lets fix it: fixed-b asymptotics versus small-b asymptotics in heteroscedasticity and autocorrelation robust inference, *Journal of Econometrics*, 178, 659-677.

Sun, Y., 2014b. Fixed-smoothing asymptotics in a two-step generalized method of moments framework, *Econometrica*, 82, 2327-2370.

Sun, Y., P. C. B. Phillips, and S. Jin, 2008. Optimal bandwidth selection in heteroskedasticity autocorrelation robust testing, *Econometrica* 76, 175-194.

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

West, K. D., 2006, Forecast evaluation, Handbook of Economic Forecasting (Vol. 2), eds. G. Elliott and A. Timmerman, Amsterdam: Elsevier 100-132.