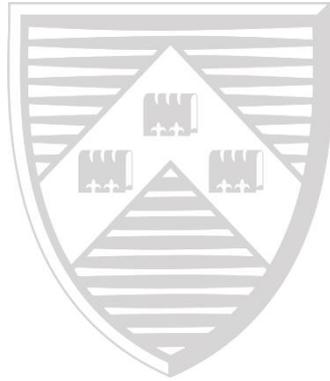# UNIVERSITY *of York*

Discussion Papers in Economics

No. 15/15

Comparing predictive accuracy in small samples

Laura Coroneo and Fabrizio Iacone

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Comparing predictive accuracy in small samples

Laura Coroneo[*]   Fabrizio Iacone
University of York   University of York

September 14, 2015

Abstract

We consider fixed-b and fixed-m 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 in small samples. We apply the alternative asympotics for 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-b and fixed-m asymptotic theory, SPF.
JEL classification: C12, C32, C53, E17

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) test. The test is based on a loss function associated with the forecast error of each forecast and tests the null of zero expected loss differential of two competing forecasts.

The Diebold and Mariano (1995) test statistic has the advantage of being simple and asymptotically normally distributed. However, as also noted by Diebold and Mariano (1995), 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 Diebold and Mariano (1995) test, the long run variance is replaced by a consistent estimate and standard limit normality is then employed, but this may be unsatisfactory in relatively small sample. The test has been extended in several directions, see Diebold (2015), for example to deal with model comparisons, Clark and McCracken (2001), and structural changes, Giacomini and Rossi (2010). However, as noted 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). They use alternative asymptotics in which the limit properties of the 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 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. With this approach, the test to compare predictive accuracy has a t-distribution with degrees of freedom that depend on the truncation parameter.

For high frequency, large sample forecast evaluations, Patton (2015) and Li and Patton (2013) show that fixed-b asymptotics deliver 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 Diebold and Mariano (1995) test alleviates the size distortion of the original test, even in presence of weak autocorrelation. The modifications of the Diebold and Mariano (1995) test based on fixed-b and fixed-m asymptotics that we propose have the advantage of being formally based on asymptotic theory, also when the loss differential is a dependent process.

We perform a Monte Carlo analysis and find that both the fixed-b and the fixed-m approaches deliver correctly sized predictive accuracy tests for highly correlated loss differentials even in small samples. We then apply our methodology 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 Diebold and Mariano (1995) test is spurious, especially in the most recent subsample.

The paper is organized as follows. Section 2 introduces the test for equal predictive accuracy and Section 2.1 describes the Diebold and Mariano (1995) test statistic. The tests for equal predictive accuracy using fixed-b asymptotics and fixed-m asymptotics are described, respectively, in Section 2.2 and Section 2.3. Section 3 contains a Monte Carlo study and Section 4 applies the testing methodology to analyse the predictive ability of the SPF. Finally, Section 5 concludes.

2   Comparing Predictive Accuracy

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

$$d_t = L(e_{1t}) - L(e_{2t})$$

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 covariance $E(u_t u_{t+j}) = \gamma_j$ and $\sigma^2 = \sum_{j=-\infty}^{\infty} \gamma_j$. The parameter σ^2 is usually referred to as the long run variance, and it is such that $0 < \sigma^2 < \infty$.

Diebold and Mariano (1995) propose to test the hypothesis of equal predictive ability

3

as $H_0 : \{\mu = 0\}$. Let

$$\bar{d} = \frac{1}{T} \sum_{t=1}^{T} d_t$$

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

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

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.

2.1   The Diebold and Mariano test

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

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

where $\widehat{\gamma}_j = \frac{1}{T} \sum_{t=1}^{T-j} \widehat{u}_t \widehat{u}_{t+j}$, with $\widehat{u}_t = d_t - \bar{d}$, 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 $M \to \infty$ and $M/T \to 0$ as $T \to \infty$. We refer to Hannan (1970) for a survey of these estimates, and for a discussion of which kernels ensure that $\widehat{\sigma}^2 > 0$.

In a variation of this approach, Diebold and Mariano (1995) note that if $\widehat{y}_{1t}$ is an optimal forecast h steps ahead, then e_{1t} is at most a MA($h - 1$), and then propose to set $M = h - 1$ and $k(j/M) = 1$ if $j/M \le 1$ and 0 otherwise, so

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

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

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

The choice of $\widehat{\sigma}_{DM}^2$ may be very appealing, as it exploits information about the structure of u_t. However, the kernel used in (3) may generate negative estimates for $\widehat{\sigma}_{DM}^2$, which is undesirable. Moreover, the Monte Carlo exercise in Diebold and Mariano (1995) suggests the possibility of large size distortions in small samples, which would be spuriously

interpreted of superior predictive power for one forecast rule. Diebold and Mariano (1995) 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.

2.2 Fixed-b asymptotics

We follow the approach of Kiefer and Vogelsang (2005) and consider alternative asymptotics for the estimate (2): for given M, the ratio M/T is taken as fixed as $T \rightarrow \infty$, instead. 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

$$\widehat{\sigma}^2_{BART} = \widehat{\gamma}_0 + 2 \sum_{j=1}^{T-1} k_{BART}(j/M) \widehat{\gamma}_j, \; M/T \rightarrow b, \tag{5}$$

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

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, these can be obtained using the formula

$$q(b) = a_0 + a_1 b + a_2 b^2 + a_3 b^3$$

where

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

The results of Kiefer and Vogelsang (2005) provide asymptotics that may be valid for any M, even $M = T$, but note 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 $M = \lfloor T^{1/3} \rfloor = 5$, 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, they show 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.

2.3   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

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

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 (and subject to other regularity conditions).

Note 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 indeed 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 very similar, and this suggests for WPE an alternative theory analogue to fixed-b for WCE. For the Daniell kernel, this is

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

6

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,

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

We provide more details about the derivation of (11) in the Appendix. Simulations in Hualde and Iacone (2015) find the same size-power trade off documented for the WCE: the smaller the value for m, the better the empirical size, but also the weaker the power.

In is worth mentioning that this approach is similar to Sun (2013) and Mü ller (2014). They project the series on a different orthonormal basis, but the results are similar. Our preference for the periodogram and the WPE is due to the fact that it is more easily understood in the frequency domain, and (9) gives an intuitive interpretation also in relation with the WCE.

3   A Monte Carlo study

In this section we analyse the size and the 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.

3.1   Size Analysis

We simulate forecast errors as in Diebold and Mariano (1995) and Clark (1999). In particular, we simulate a bivariate, normally independently distributed vector, $(v_{1t}, v_{2t})'$, with covariance matrix $diag\{1, 1\}$, and

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

and

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

7

where $k = 1$, $\rho = 0.5$ and $\theta = 0.5$, and 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 forecast horizon q.

The summary of the experiment, with theoretical size set to 5% is in Tables 1–2. In all cases we use 10000 replications (entries in the tables are rounded to the third decimal digit) an 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.

In the first part of the experiment, we study the size properties when standard asymptotics and limit normality is used. In Table 1, we report results for various consistent estimates of σ, for $T = 40$ and $T = 120$, respectively, using limit normal asymptotics. In column WCE-DM we report the empirical size when the Diebold and Mariano WCE in (3) is used; in WCE-BART we use (5)–(6) and $M = \lfloor T^{1/3} \rfloor$; in WPE-DAN we use the WPE estimate (10) with $m = \lfloor T^{1/2} \rfloor$. In all these cases, as the estimates are consistent, we use standard normal limit distributions to compute the empirical size. Finally, in column $\widehat{\sigma}^2_{DM} < 0$ we report the frequency of negative estimates of the long run variance when using the Diebold and Mariano WCE estimate as in (3).

Consistently with results in Clarks (1999), we find that, as the forecast horizon q increases, the size of the test with the Diebold and Mariano WCE estimate deteriorates, although the size distortion is less serious in the larger sample. We also find that the risk of a negative estimate $\widehat{\sigma}^2_{DM}$ is not negligible, especially for small T and large q. In comparison, using the Bartlett kernel in the WCE gives better size properties even when asymptotic normality is used, especially for the longest forecast horizons. We also experimented with the automatic bandwidth selection procedure for the Bartlett kernel in Newey and West (1994) but, as we found that this resulted in worse size properties than setting $M = \lfloor T^{1/3} \rfloor$, in the interest of brevity we do not include it in this discussion. On balance, within the estimates that we considered, Table 1 shows that when asymptotic normality is maintained the best properties in size are for the WPE with Daniell kernel and $m = \lfloor T^{1/2} \rfloor$, although a certain size distortion still occurs, especially in the smallest sample.

In Table 2 we report results when the properties of the estimates of σ are derived assuming fixed-b or fixed-m asymptotics. In columns WCE-BART, $M = \lfloor T^{1/3} \rfloor$ or $M = \lfloor T^{1/2} \rfloor$, we use (5)–(6) and fixed-b asymptotics, with limit (7); in WPE-DAN we use the estimate (10) with $m = \lfloor T^{1/4} \rfloor$ and $m = \lfloor T^{1/3} \rfloor$ and asymtotics from (11). Note that in one case the same test statistic is used both in Table 1 and Table 2: this is the

Table 1: Size of tests with standard asymptotics

T=40

	$\widehat{\sigma}^2_{DM} < 0$	WCE-DM	WCE-BART	WPE-DAN
MA(1)	0.000	0.077	0.085	0.075
MA(2)	0.001	0.099	0.090	0.076
MA(3)	0.007	0.124	0.096	0.078
MA(4)	0.017	0.157	0.097	0.081
MA(5)	0.037	0.196	0.097	0.080

T=120

	$\widehat{\sigma}^2_{DM} < 0$	WCE-DM	WCE-BART	WPE-DAN
MA(1)	0.000	0.059	0.068	0.061
MA(2)	0.000	0.058	0.066	0.060
MA(3)	0.000	0.068	0.072	0.062
MA(4)	0.000	0.074	0.075	0.062
MA(5)	0.001	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 processes and alternative estimates of the long run variance: WCE-DM is for the WCE with the truncated kernel as in Diebold and Mariano (1995), WCE-BART is for the Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$, and WPE-DAN for the WPE with Daniell kernel and $m = \lfloor T^{1/2} \rfloor$. The column $\widehat{\sigma}^2_{DM} < 0$ reports the frequency of negative estimates of the long run variance.

Table 2: Size of tests with fixed-b and fixed-m asymptotics

T=40

| | WCE-BART | | WPE-DAN | |
	$M = \lfloor T^{1/3} \rfloor$	$M = \lfloor T^{1/2} \rfloor$	$m = \lfloor T^{1/4} \rfloor$	$m = \lfloor T^{1/3} \rfloor$
MA(1)	0.056	0.050	0.044	0.045
MA(2)	0.058	0.049	0.042	0.042
MA(3)	0.060	0.049	0.041	0.041
MA(4)	0.063	0.049	0.040	0.040
MA(5)	0.061	0.049	0.040	0.043

T=120

| | WCE-BART | | WPE-DAN | |
	$M = \lfloor T^{1/3} \rfloor$	$M = \lfloor T^{1/2} \rfloor$	$m = \lfloor T^{1/4} \rfloor$	$m = \lfloor T^{1/3} \rfloor$
MA(1)	0.053	0.050	0.049	0.048
MA(2)	0.055	0.045	0.044	0.044
MA(3)	0.058	0.049	0.046	0.048
MA(4)	0.060	0.047	0.043	0.044
MA(5)	0.062	0.052	0.049	0.049

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size using fixed-b or fixed-m asymptotics for various MA processes and alternative estimates of the long run variance: WCE-BART is for the WCE with Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$ or $M = \lfloor T^{1/2} \rfloor$; WPE-DAN for the WPE with Daniell kernel and $m = \lfloor T^{1/4} \rfloor$ or $m = \lfloor T^{1/3} \rfloor$.

WCE with Bartlett kernel and $M = \lfloor T^{1/3} \rfloor$. The difference in size in the two tables is then due only to the different critical values. Results in Table 2 show that predictive accuracy tests with fixed-b and fixed-m asymptotics are overall correctly sized, although a slight size distortion is still present when the smaller bandwidth, $M = \lfloor T^{1/3} \rfloor$, is used with fixed-$b$ asymptotics.

3.2   Power Analysis

In the second part of the Monte Carlo exercise, we study the power of the proposed tests of equal predictive accuracy. In this case, we test $H_0 : \{\mu = 0\}$ in processes with $\mu = cT^{-1/2}$, for c ranging between 0 and 7. As 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 10000 repetitions and $T = 40$ and $T = 120$. We compare the tests with fixed-b or fixed-m 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 with $b \to 0$ or a WPE 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 lower 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 noting, however, that the power loss due to estimating σ is minimal, especially when the WCE is used.

4   Predictive Accuracy of the SPF

To illustrate the usefulness of alternative asymptotics for equal predictive accuracy tests, we evaluate the predictive accuracy of the 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

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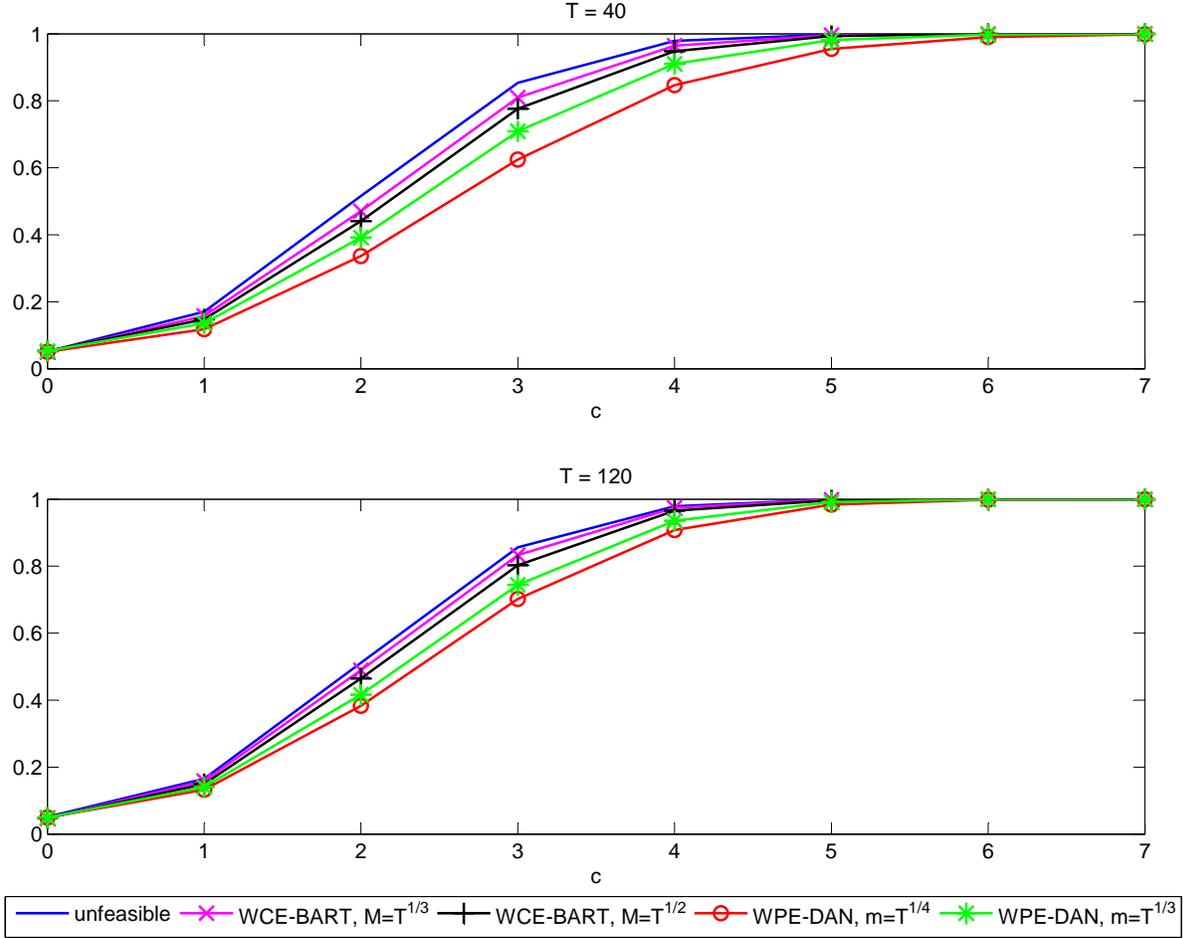
The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $cT^{-1/2}$ and independent innovations. Unfeasible refers to the case in which the unknown variance is used and the test statistic has standard normal limit distribution. For the feasible tests, fixed-b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

forecasts is the middle of the quarter. We focus on median responses for the period from 1985:Q1 until 2014:Q4 and use a quadratic loss function to evaluate the predictive accuracy of the SPF on the full sample and also on 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 the surveys from 1992:Q1 to 1995:Q4, and the chain-weighted price index in the surveys thereafter. Accordingly, real output is defined as 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. 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 similarly, see Stark (2010). Finally, the three-month Treasury bill rate and the unemployment rate are expressed in levels.

We compute benchmark forecasts and realised values using the vintages of data that were available to the public before the survey's mid-quarter deadline. For all the variables considered, we use as benchmark a naive random walk, i.e. a no change benchmark. 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, we use as benchmark a random walk with drift on real GNP/GDP levels: this is a more appropriate benchmark for real GNP/GDP growth, see Stark (2010). We estimate the drift parameter using a rolling average of real GNP/GDP growth with a window of 60 observations.

Tables 3–6 report the test statistics presented in Section 2 for the null hypothesis of equal predictive accuracy of the SPF' forecasts for real GNP/GDP growth, GNP/GDP inflation, the unemployment rate and the three-month T-Bill rates with respect to the random walk. Results in Tables 3–6 show that the predictive ability of the SPF is stronger for the three months Treasury bill rate and GNP/GDP inflation than for real GNP/GDP growth and the unemployment rate. 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 3 shows that the SPF's forecasts for real GNP/GDP growth do not in general outperform the random walk. For the current quarter, the Diebold and Mariano (1995) test 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 4 shows a much stronger

Table 3: 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-BART, $M = \lfloor T^{1/3} \rfloor$	1.83*	1.32	0.90	0.06	-0.06
WCE-BART, $M = \lceil T^{1/2} \rceil$	1.83*	1.33	0.88	0.06	-0.05
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.66	1.16	0.77	0.05	-0.05
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	1.54	1.20	0.89	0.77	0.60
WCE-BART, $M = \lceil T^{1/2} \rceil$	1.56	1.20	0.90	0.71	0.58
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.46	1.14	0.91	1.10	0.57
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	1.19	-0.72	-1.80	-1.64	-1.13
WCE-BART, $M = \lceil T^{1/2} \rceil$	1.24	-0.82	-1.71	-1.53	-1.13
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.11	-0.97	-1.45	-1.30	-1.04
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	1.32	1.17	1.12	0.52	0.47
WCE-BART, $M = \lceil T^{1/2} \rceil$	1.29	1.18	1.16	0.59	0.50
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.09	1.01	1.01	0.54	0.44
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-side significance at the 5% and 10% level.

Table 4: 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-BART, $M = \lfloor T^{1/3} \rfloor$	3.28**	3.65**	2.39**	0.72	1.59
WCE-BART, $M = \lceil T^{1/2} \rceil$	2.89**	3.71**	2.38**	0.62	1.57
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	2.59**	3.23**	1.81	0.50	1.24
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	2.73**	3.09**	1.50	-0.52	0.35
WCE-BART, $M = \lceil T^{1/2} \rceil$	3.08**	3.11**	1.54	-0.50	0.35
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	3.59**	2.61*	1.82	-0.42	0.28
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	1.04	1.19	0.59	0.33	0.38
WCE-BART, $M = \lceil T^{1/2} \rceil$	0.94	1.16	0.56	0.31	0.36
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.01	1.16	0.50	0.28	0.33
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	2.42**	2.37**	2.04*	1.42	2.25**
WCE-BART, $M = \lceil T^{1/2} \rceil$	2.35*	2.63**	2.05*	1.27	2.89**
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.85	2.59*	1.91	0.99	3.07**
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-side significance at the 5% and 10% level.

Table 5: 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-BART, $M = \lfloor T^{1/3} \rfloor$	2.42**	1.98*	2.06*	2.22**	2.54**
WCE-BART, $M = \lceil T^{1/2} \rceil$	2.31**	1.91*	1.95*	2.07*	2.30**
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	2.12*	1.80	1.82	1.87	2.03*
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	2.69**	1.75	1.85*	2.18*	2.58**
WCE-BART, $M = \lceil T^{1/2} \rceil$	2.82**	1.82	1.96	2.42**	2.89**
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	3.42**	1.73	2.09	2.92**	3.74**
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	2.00*	1.73	1.43	1.32	1.28
WCE-BART, $M = \lceil T^{1/2} \rceil$	2.10*	1.71	1.35	1.25	1.18
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.80	1.47	1.16	1.14	1.04
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	1.86*	1.66	1.78	1.97*	2.27**
WCE-BART, $M = \lceil T^{1/2} \rceil$	1.81	1.61	1.71	1.88	2.12*
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.50	1.32	1.39	1.54	1.72
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-side significance at the 5% and 10% level.

Table 6: 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-BART, $M = \lfloor T^{1/3} \rfloor$	4.29**	4.31**	3.72**	2.48**	1.48
WCE-BART, $M = \lceil T^{1/2} \rceil$	4.46**	4.46**	4.01**	2.61**	1.50
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	4.89**	5.08**	3.66**	2.21*	1.38
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	5.02**	4.12**	3.28**	1.33	0.75
WCE-BART, $M = \lceil T^{1/2} \rceil$	5.87**	4.75**	3.96**	1.32	0.72
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	9.34**	5.70**	6.21**	1.35	0.71
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	1.94*	1.88*	1.58	1.20	0.40
WCE-BART, $M = \lceil T^{1/2} \rceil$	1.83	1.78	1.59	1.35	0.49
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.55	1.53	1.44	1.50	0.63
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-BART, $M = \lfloor T^{1/3} \rfloor$	1.93*	2.21**	2.11*	1.92*	1.46
WCE-BART, $M = \lceil T^{1/2} \rceil$	1.81	2.08*	1.87	1.68	1.23
WPE-DAN, $m = \lfloor T^{1/4} \rfloor$	1.65	2.17*	1.82	1.56	1.06
WPE-DAN, $m = \lceil T^{1/3} \rceil$	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-side significance at the 5% and 10% level.

predictive ability of the SPF, especially for short horizons and in the first and the last subsamples. Results in Table 5 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-b and fixed-m asymptotics. Finally, Table 6 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 the three tests described in Section 2, we can see that the Diebold and Mariano (1995) test in (4) rejects the null hypothesis of equal predictive ability more often that the WCE and the WPE tests, especially in the subsamples. This is due to the fact that in the subsamples the tests are performed only on 40 observations, exacerbating the size distortions of the Diebold and Mariano (1995) test, see Table 1. For example, Table 5 shows that for inflation the Diebold and Mariano (1995) test rejects the null of equal predictive ability of the SPF and the random walk on the last subsample and 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 WPE based tests fail to reject the null of equal predictive ability for all forecasting horizons, indicating that the SPF did not have any predictive ability for the unemployment rate in this period.

5   Conclusion

We reassess the predictive accuracy of the Survey of Professional Forecasters (SPF) against a simple random walk for four core macroeconomic indicators: output growth, inflation, the unemployment rate and the three month Treasury bill rate. As the Diebold and Mariano (1995) test of predictive accuracy is subject to relevant size distortion when the standard normal asymptotics are used, we consider fixed-b and fixed-m asymptotics instead. We find that the predictive ability of the SPF was partially spurious, especially in the last decade.

6   Appendix

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 $E\left(\varepsilon_t\right) = 0$, $E\left(\varepsilon_t^2\right) = 1$, $E\left(\varepsilon_t^4\right) < \infty$, and $\sum_{l=0}^{\infty} j^{1/2}\left|A_l\right| < \infty$. Define the Fourier frequencies $\lambda_j = 0, \pm 1, ..., \lfloor T/2 \rfloor$ and the Fourier transform $w_x\left(\lambda\right) = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} x_t e^{i\lambda t}$, the periodogram $I_x\left(\lambda\right) = \left|w_x\left(\lambda\right)\right|^2$, 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 $T \to \infty$,

$$\tau \to_d t_{2m} \tag{12}$$

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 $w_x\left(\lambda_j\right) = w_u\left(\lambda_j\right)$. Moreover, following Hannan (1970), page 247,

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

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

$$w_u\left(\lambda_j\right) = \left(\sum_{l=0}^{\infty} A_l e^{i\lambda_j l}\right) w_\varepsilon\left(\lambda_j\right) + o_p\left(1\right) \tag{13}$$

Now let

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

then sufficient conditions for the central limit theorem are that

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

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.,

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

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

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 conclude

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

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 $\sum_{t=1}^{T}\frac{1}{\sqrt{2\pi T}}\varepsilon_t\cos\left(\frac{2\pi jt}{T}\right)$ and $\sum_{t=1}^{T}\frac{1}{\sqrt{2\pi T}}\varepsilon_t\sin\left(\frac{2\pi jt}{T}\right)$ to a bivariate vector of independently distributed random variables with diagonal covariance matrix follows from an application of the Cramer-Wold device.

Therefore,

$$2\left(2\pi\right)I_\varepsilon\left(\lambda_j\right) \to_d \chi_2^2.$$

Moreover, as in Hannan (1970), page 249,

$$\text{for } j \neq k, \; Cov\left(I_\varepsilon\left(\lambda_j\right)I_\varepsilon\left(\lambda_k\right)\right) \to 0$$

and therefore

$$2\pi\frac{1}{m}\sum\nolimits_{j=1}^{m}I_{\varepsilon}\left(\lambda_{j}\right)\to_{d}C_{2m}^{2}/\left(2m\right)$$

where $C_{2m}^{2}/\left(2m\right)$ is a χ_{2m}^{2} distributed random variable divided by the number of degrees of freedom.

Using (13) and $\sum_{l=0}^{\infty}A_{l}e^{i\lambda_{j}l}\to\sum_{l=0}^{\infty}A_{l}=\sigma$, it also follows that

$$2\pi\frac{1}{m}\sum\nolimits_{j=1}^{m}I_{x}\left(\lambda_{j}\right)\to_{d}\sigma^{2}C_{2m}^{2}/\left(2m\right).$$

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

$$u_{t}=\left(\sum\nolimits_{l=0}^{\infty}A_{l}\right)\varepsilon_{t}+\widetilde{\varepsilon}_{t-1}-\widetilde{\varepsilon}_{t}$$

where

$$\widetilde{\varepsilon}_{t}=\sum\nolimits_{l=0}^{\infty}\widetilde{A}_{l}\varepsilon_{t-l},\ \widetilde{A}_{l}=\sum\nolimits_{k=l+1}^{\infty}A_{k}$$

and

$$\frac{1}{\sqrt{T}}\sum\nolimits_{t=1}^{T}u_{t}=\left(\sum\nolimits_{l=0}^{\infty}A_{l}\right)\frac{1}{\sqrt{T}}\sum\nolimits_{t=1}^{T}\varepsilon_{t}+\frac{1}{\sqrt{T}}\left(\widetilde{\varepsilon}_{0}-\widetilde{\varepsilon}_{T}\right) \tag{14}$$

where $\frac{1}{\sqrt{T}}\left(\widetilde{\varepsilon}_{0}-\widetilde{\varepsilon}_{T}\right)=o_{p}\left(1\right)$ as in 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 $\sum_{l=0}^{\infty}l^{1/2}\left|A_{l}\right|<\infty$, Phillips and Solo (1992) page 973, so

$$\frac{1}{\sqrt{T}}\sum\nolimits_{t=1}^{T}u_{t}=\sigma\frac{1}{\sqrt{T}}\sum\nolimits_{t=1}^{T}\varepsilon_{t}+o_{p}\left(1\right)\to_{d}N\left(0,\sigma^{2}\right).$$

Finally, it is easy to verify that

$$\begin{aligned}E\left(\frac{1}{\sqrt{T}}\sigma\sum\nolimits_{t=1}^{T}\varepsilon_{t}\left(\sum\nolimits_{l=0}^{\infty}A_{l}e^{i\lambda_{j}l}\right)w_{\varepsilon}\left(\lambda_{j}\right)\right)&=\frac{\sigma\left(\sum_{l=0}^{\infty}A_{l}e^{i\lambda_{j}l}\right)}{\sqrt{2\pi T}}E\left(\sum\nolimits_{t=1}^{T}\varepsilon_{t}\sum\nolimits_{s=1}^{T}\varepsilon_{s}e^{i\lambda_{j}s}\right)\\&=\frac{\sigma\left(\sum_{l=0}^{\infty}A_{l}e^{i\lambda_{j}l}\right)}{\sqrt{2\pi T}}\sum\nolimits_{s=1}^{T}e^{i\lambda_{j}s}=0\end{aligned}$$

and thus conclude that (12) holds.

Remark. Condition $\sum_{l=0}^{\infty}l^{1/2}\left|A_{l}\right|<0$ 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 $I_{\varepsilon}\left(\lambda_{j}\right)$ at different frequencies is also discussed. The main reason

of interest for the proof here is then in the fact that, using the decompositions (13) and (14) we see that we 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.

7 References

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. W., 2001. Tests of Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Forecasting* , 105, 85-110.

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

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

Giacomini, R., and Rossi, B., 2010. Forecast Comparisons in Unstable Environments, *Journal of Applied Econometrics*, 25, 595-620.

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

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

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

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, 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*, 33:1, 22-24.

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

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.