Comparing predictive accuracy in small samples using fixed-smoothing asymptotics

Online Appendix
"Monte Carlo with forecasts as primitive"

Laura Coroneo
University of York

Fabrizio Iacone
Università degli Studi di Milano
University of York

S.1 Simulation design

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

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

In all cases, we use 10,000 replications and a quadratic loss function, i.e. $d_t = e_{1t}^2 - e_{2t}^2$. The expected loss differential is zero for $k = 1$, so we use $k = 1$ to evaluate size and $k = 1 + c/\sqrt{T}$ for $c = 1, \ldots, 30$ to evaluate power.

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 samples in the first empirical analysis. The objective of this Monte Carlo exercise is to experiment with a variety of bandwidths with a design and sample size that is reasonable in the context of forecast evaluation.

Our choice of candidates is informed by recommendations and practice in the theoretical and empirical literature. One popular criterion for bandwidth choice is based on minimising the MSE for the estimates $\widehat{\sigma}$. For the WCE-B, Newey and West (1994) show that the optimal bandwidth, in minimal MSE sense, is proportional to $M = \lfloor T^{1/3} \rfloor$; for the WPE-D, Delgado and Robinson (1996), Phillips (2005) and Sun (2013) show that the optimal bandwidth, in MSE sense, is proportional to $\lfloor T^{4/5} \rfloor$. More recently, alternative criteria have been proposed to define optimal bandwidths. These are based on the fact that in the context of testing, our interest in estimates $\widehat{\sigma}$ depends on the fact that these affect the properties, in terms of type 1 or type 2 error, of the test. Thus, an estimate may well be a poor choice if it causes poor type 1 or type 2 error performances, and this regardless of the MSE of such estimate. This approach was pioneered by Sun, Phillips and Jin (2008), also see Sun (2013), Sun (2014), Lazarus, Lewis, Stock and Watson (2018) and Sun (2018) for further discussion. For example, Sun (2014) shows that taking as criterion the minimum type 2 error subject to control of the type 1 error, the resulting optimal WCE-B bandwidth has larger order of magnitude than the minimum MSE optimal bandwidth. Conversely, for the WPE-D Sun (2013) shows that the bandwidth that minimises the Coverage Probability Error is proportional to $\lfloor T^{2/3} \rfloor$. In all these cases, the optimal bandwidth requires the knowledge of a scaling factor that typically depends on the lower order bias generated by the steepness of the spectrum in the neighborhood of the origin. This may be estimated in a preliminary step, thus making automatic bandwidths feasible, however, this adds some complexity and makes the results dependent on this intermediate step. To avoid this inconvenience,

this scaling factor is sometimes approximated as a constant, and the performance of this naive choice is analysed in Monte Carlo simulations. For example, with standard critical values, Abadir, Distaso and Giraitis (2009) recommend $m = \lfloor T^{2/3} \rfloor$, whereas with critical values from fixed-smoothing asymptotics Lazarus et al. (2018) recommend the bandwidth rules $M = \lfloor 1.3T^{1/2} \rfloor$ for the WCE-B and $2m = \lfloor 0.4T^{2/3} \rfloor$, or $m = \lfloor 0.2T^{2/3} \rfloor$, for the WPE-D. We notice, however, that the Monte Carlo exercises used in these experiments are very different from ours both in the design and the sample sizes, and may not necessarily provide a good indication for the problem of forecast evaluation.

S.2 Size analysis

To evaluate size, we set $k = 1$. Results in Clark (1999) suggest only limited sensitivity of size to ρ and θ, so we fix $\rho = 0.5$ and $\theta = 0.75$, and investigate the effect of increasing the serial correlation with q. DM, Clark (1999) and Harvey, Leybourne and Whitehouse (2017) 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.[1] In Tables S2–S2 we report results of the Monte Carlo, with theoretical size set to 5%. In the first part of the experiment, we study the size properties treating the estimates of σ_T as consistent and using standard asymptotics, i.e. the limit normal distribution, to compute the empirical size. In Table S2 we report the empirical size of the tests when the WCE-DM, WCE-B and WPE-D are used to estimate σ_T using standard asymptotics. When using the WCE-DM estimate, negative estimates are possible. We treat these instances as rejections of the null hypothesis.[2][3]

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

[1] In S.5.1, we report a size study for $\theta = \{-0.5, 0, 0.5\}$.

[2] We discuss these occurrences in S.5.2.

[3] In S.5.3, we report a size study for the case in which the innovations are drawn from an asymmetric distribution.

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).[4] The second bandwidth, $M = \lfloor T^{1/2} \rfloor$, is chosen because existing Monte Carlo evidence for fixed-b asymptotics suggests that longer bandwidths are associated with better empirical size, it is therefore interesting to compare the performance of the same test statistic when standard and fixed-smoothing asymptotics are used. As for the bandwidths for the WPE-D, in samples as small as 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_T 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.[5]

In general, Table S2 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, at least for small values of q. 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 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_T at higher frequencies, which affects $m = \lfloor T^{2/3} \rfloor$ most, and the high variance of the estimate, which is more a problem

[4] In S.5.4 we also consider the automatic procedures from Newey and West (1994).

[5] Monte Carlo results for bandwidths proportional to $\lfloor T^{4/5} \rfloor$, reported in S.5.5, indicate large size distortions and, therefore, we do not recommend using these bandwidths.

Table S2: 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	0.089	0.109	0.093	0.075	0.081
2	0.095	0.106	0.116	0.095	0.082	0.106
3	0.115	0.120	0.121	0.090	0.089	0.137
4	0.141	0.135	0.128	0.096	0.102	0.163
5	0.173	0.153	0.133	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	0.068	0.076	0.084	0.062	0.064
2	0.057	0.074	0.079	0.079	0.058	0.070
3	0.064	0.086	0.084	0.082	0.063	0.089
4	0.073	0.097	0.090	0.082	0.069	0.108
5	0.085	0.111	0.097	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 $m = \lfloor T^{2/3} \rfloor$.

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 S2 we report results when the properties of the estimates of σ_T and of the test statistic are derived assuming fixed-smoothing asymptotics. In columns WCE, we use (6)–(7), with $M = \lfloor T^{1/3} \rfloor$, $M = \lfloor T^{1/2} \rfloor$, and $M = T$, and fixed-b asymptotics, with limit (8); in columns WPE, we use the estimate (11) with $m = \lfloor T^{1/4} \rfloor$, $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/2} \rfloor$ and asymptotics from (12). 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 S2 and Table S2, and the difference in the empirical size in the two tables is due only to the different

Table S2: 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	0.059	0.051	0.054	0.044	0.044	0.051
2	0.068	0.055	0.054	0.043	0.043	0.052
3	0.082	0.056	0.054	0.037	0.041	0.057
4	0.095	0.061	0.057	0.039	0.039	0.065
5	0.105	0.064	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	0.057	0.050	0.049	0.050	0.047	0.047
2	0.059	0.047	0.048	0.044	0.046	0.044
3	0.071	0.051	0.048	0.045	0.045	0.049
4	0.081	0.055	0.048	0.043	0.043	0.052
5	0.093	0.061	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 $m = \lfloor T^{1/2} \rfloor$.

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 S2. The size distortion for $m = \lfloor T^{2/3} \rfloor$ documented in Table S2 is due to the bias in the estimation of the long run variance and therefore it 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 S2 and S2, 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 asymptotics 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, with fixed-b asymptotics it seems desirable to choose bandwidths M longer than what we would consider when standard asymptotics is used; this result is mirrored for fixed-m asymptotics, in this 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 to limit this size distortion, but not to 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.

S.3 Power analysis

The previous exercise shows 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, we now study the power of the tests.

In this experiment, we only consider test statistics in which σ_T is estimated as the WCE-B or as WPE-D, and only use critical values from fixed-smoothing asymptotics.[6] Notice that we also include two cases in which even the non-standard asymptotics does not completely eliminate the size distortion: when σ_T is estimated with $M = \lfloor T^{1/3} \rfloor$

[6] A power comparison between standard and fixed-smoothing asymptotics is reported in S.5.6.

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.

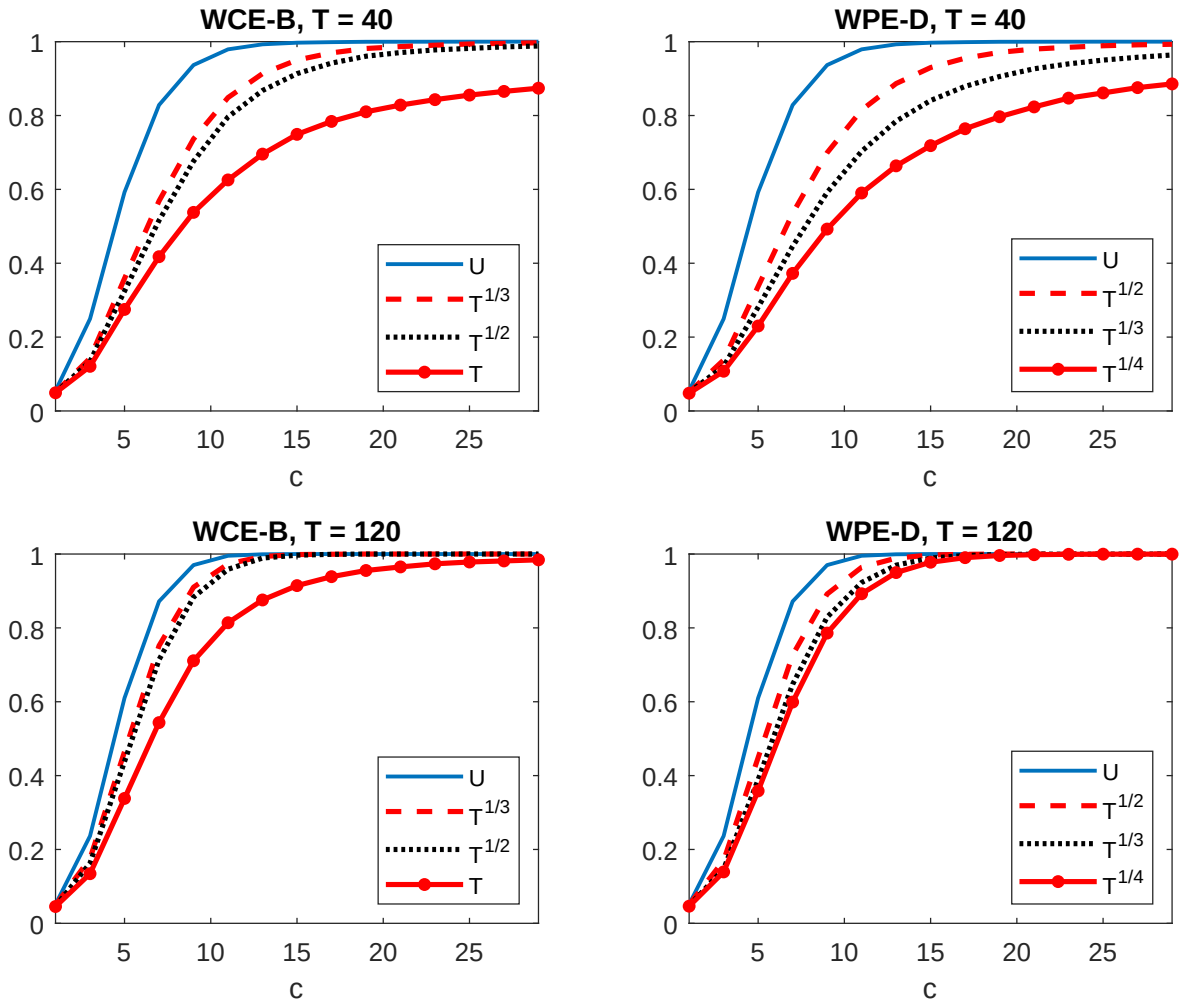
To evaluate power we set $k = 1 + c/\sqrt{T}$ for $c = 1, \ldots, 30$. As in the size exercise, we fix $\rho = 0.5$, however since in this part of the exercise we are interested in power, rather than in size distortion, we fix $\theta = 0$.[7] We also compare the tests with fixed-smoothing asymptotics against a benchmark case in which σ_T 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_T^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 S1. Previous simulations in Kiefer and Vogelsang (2005), in Hualde and Iacone (2015) and in Lazarus et al. (2018) 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_T^2 has the highest power, as expected. It is worth noticing, however, that the power loss due to estimating σ_T 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.

Recommended bandwidth. *Considering size control and power loss in our Monte*

[7] Results when autocorrelation is preserved under the alternative are reported in S.5.7

Figure S1: Finite sample local power



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $c/\sqrt{T}$. 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. 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 $m = \lfloor T^{1/4} \rfloor$.

Carlo experiment, we recommend $M = \lfloor T^{1/2} \rfloor$ *for the WCE-B and* $m = \lfloor T^{1/3} \rfloor$ *for the WPE-D.*

Our bandwidths recommendations are consistent with Lazarus et al. (2018), as their rule $m = \lfloor 0.2T^{2/3} \rfloor$ with $T = 40$ and $T = 120$ generates bandwidths $m = 2$ and $m = 4$ respectively, which are very close to $m = 3$ and $m = 4$ that we use with the rule $m = \lfloor T^{1/3} \rfloor$ (indeed, for $T = 120$ both rules give the same bandwidth).

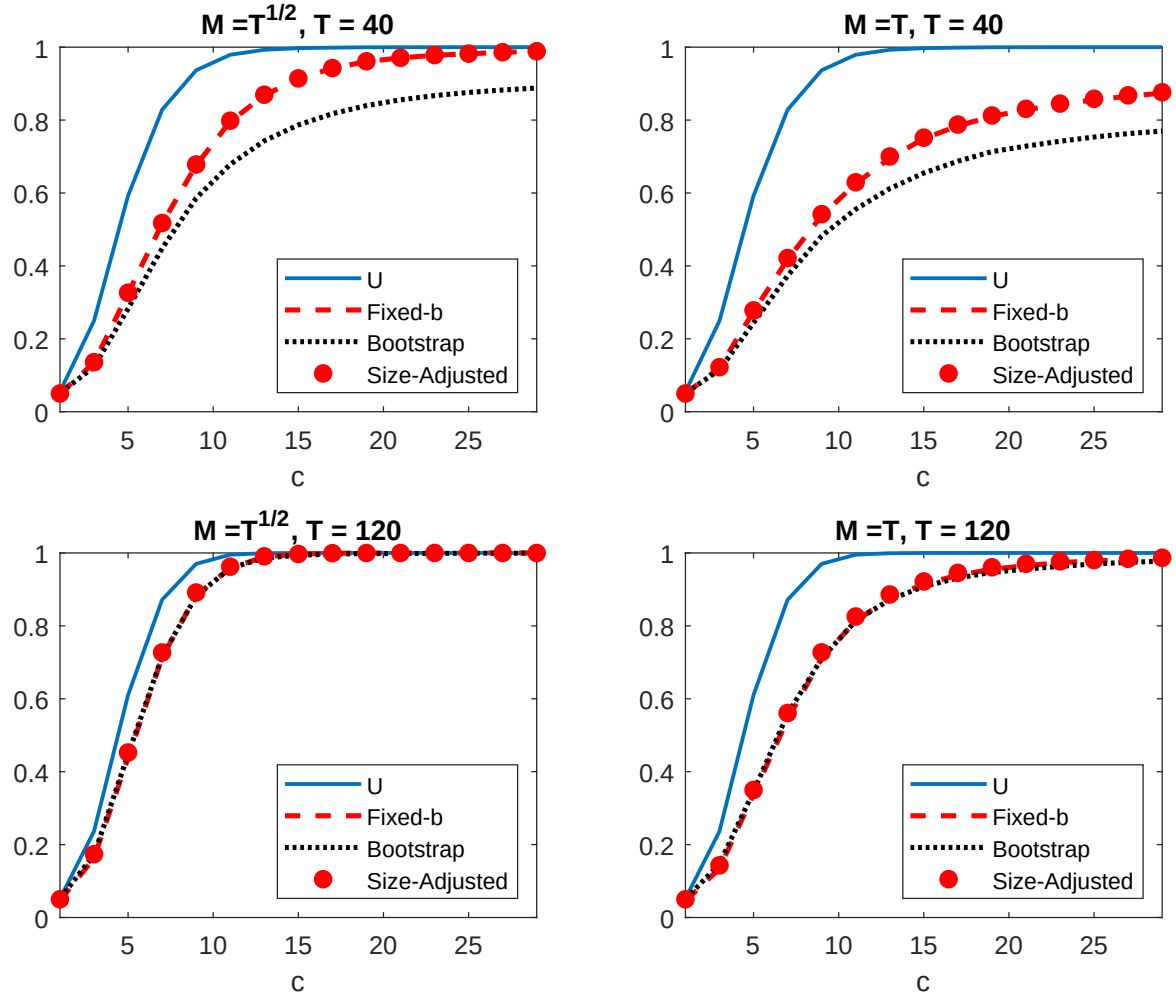
S.4 Comparison with the bootstrap

The 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 analysis of the size and power of the tests for equal predictive ability using bootstrap critical values, and make a comparison with the results obtained using fixed-smoothing asymptotics.

Bootstrap critical values are computed using the overlapping stationary block-bootstrap of Politis and Romano (1994) with a circular scheme, as described in Appendix C. In Table S4, we report the size of tests of equal predictive ability using bootstrap critical values for various MA processes with $\rho = 0.5$, $\theta = 0.75$ 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 $m = \lfloor T^{1/2} \rfloor$. Results in Table S4 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).

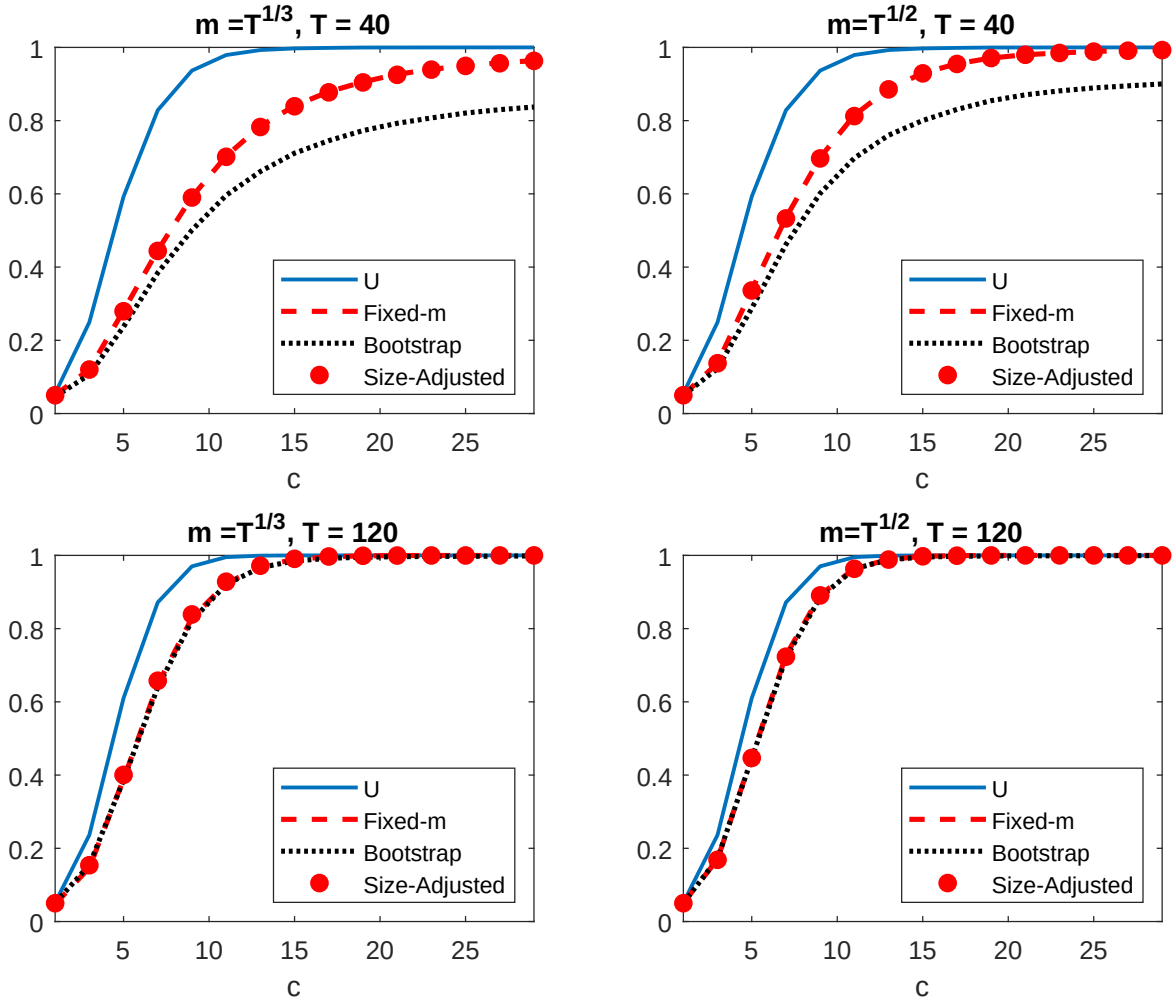
In Figures S2-S3, we report the finite sample local power comparison of fixed-b and fixed-m asymptotics with the bootstrap. As for the power exercise in S.3, we fix $\rho = 0.5$ and $\theta = 0$. For additional comparison, we also plot the size-adjusted power for the statistics of interest: this is the local power that we would obtain using the test statistic

Figure S2: Finite sample local power: fixed-b vs bootstrap



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $c/\sqrt{T}$. U refers to the unfeasible case in which the unknown variance is used and the test statistic has standard normal limit distribution. Size-Adjusted refers to the case in which simulated size-adjusted critical values are used. 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 $M = \lfloor T^{1/2} \rfloor$ (left panel) or $M = T$ (right panel).

Figure S3: Finite sample local power: fixed-m vs bootstrap



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $c/\sqrt{T}$. U refers to the unfeasible case in which the unknown variance is used and the test statistic has standard normal limit distribution. Size-Adjusted refers to the case in which simulated size-adjusted critical values are used. 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 $m = \lfloor T^{1/3} \rfloor$ (left panel) or $m = \lfloor T^{1/2} \rfloor$ (right panel).

Table S4: 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 $m = \lfloor T^{1/2} \rfloor$.

of interest if we could eliminate any size distortion. Both figures indicate that the bootstrap local power mimics size-adjusted local power quite well, especially when the largest sample is used. However, we also find that using fixed-smoothing asymptotics in testing we replicate the infeasible size-adjusted local power even better, especially in the smallest sample.

Size 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. 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. Gonçalves and Vo-

gelsang (2011) also find that the power of the naive block-bootstrap closely follows the power when using the fixed-b critical value. Our results, on the other hand, find some divergence, at least in the smallest sample, thus suggesting that the consideration of Kiefer and Vogelsang (2005) might apply here too. Overall, we conclude that in the set up of our experiment both using fixed-smoothing asymptotics and bootstrapping deliver good size and approximate well the correct (size-adjusted) local power function under a local alternative. However, as the bootstrap is much more computationally intensive, fixed-smoothing asymptotics may be preferred in forecast evaluations.

S.5 Additional results

In this section, we report additional Monte Carlo results that include the size of standard asymptotics for $\theta = \{-0.5, 0, 0.5\}$ and the frequency of negative estimates for the long run variance using the WCE-DM. We also include the size of tests with asymmetric data generating process, the size for the WCE-B with automatic bandwidth selection and size for the WPE-D with feasible minimum MSE bandwidth. Finally, we present power comparisons with standard asymptotics, and power comparisons with autocorrelation.

S.5.1 Sensitivity to θ

In Table S5, we study the size properties of the DM test for various estimates of σ_T when $\theta = \{-0.5, 0, 0.5\}$ assuming standard asymptotics. This exercise allows a comparison with Table S2 in which $\theta = 0.75$ s 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 change of the value of θ to $\theta = -0.5$ or $\theta = 0.5$; on the other hand, the reduction in the dependence is associated with an improvement in the size properties when the WCE-B or the WPE-D is used. In the cases of $\theta = -0.5$ and $\theta = 0.5$, the evidence that the test with statistic standardized by the WPE-D estimate

Table S5: Size of tests with standard asymptotics for $\theta = \{-0.5, 0, 0.5\}$

$\theta = -0.5$

	T=40			T=120		
q	WCE-DM	WCE-B	WPE-D	WCE-DM	WCE-B	WPE-D
1	0.075	0.083	0.074	0.056	0.067	0.065
2	0.102	0.093	0.079	0.065	0.075	0.065
3	0.126	0.092	0.076	0.075	0.079	0.071
4	0.166	0.100	0.083	0.078	0.076	0.067
5	0.194	0.097	0.079	0.092	0.081	0.069

$\theta = 0$

	T=40			T=120		
q	WCE-DM	WCE-B	WPE-D	WCE-DM	WCE-B	WPE-D
1	0.081	0.073	0.072	0.060	0.059	0.063
2	0.112	0.073	0.072	0.070	0.059	0.063
3	0.140	0.073	0.072	0.074	0.059	0.063
4	0.174	0.073	0.072	0.085	0.059	0.063
5	0.210	0.073	0.072	0.090	0.059	0.063

$\theta = 0.5$

	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 WCE with the Bartlett kernel and $M = \lfloor T^{1/3} \rfloor$, and WPE-D for the WPE with Daniell kernel and $m = \lfloor T^{1/2} \rfloor$. Results in the top panel are simulated using $\theta = -0.5$, results in the middle panel refer to $\theta = 0$ and results in the bottom panel are simulated using $\theta = 0.5$.

with $m = \lfloor T^{1/2} \rfloor$ gives best size is even more compelling. In the case of $\theta = 0$, the simulations show that the test with WCE-DM is heavily oversized. This is due to the fact that one does not know that θ is 0 and, therefore, adds more lagged autocovariances, increasing the risk of a negative estimate and of a larger bias.

S.5.2 Negative estimates of the long run variance

In Table S5, we study the frequency of negative estimates for $\widehat{\sigma}^2_{DM}$, the WCE estimate with the rectangular kernel (WCE-DM) defined in (4).

Table S5: Frequency of negative estimates for the long run variance

	T=40				T=120			
θ / q	−0.50	0.00	0.50	0.75	−0.50	0.00	0.50	0.75
1	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
2	0.002	0.005	0.001	0.000	0.000	0.000	0.000	0.000
3	0.006	0.014	0.007	0.003	0.000	0.000	0.000	0.000
4	0.019	0.033	0.017	0.007	0.000	0.000	0.000	0.000
5	0.038	0.060	0.037	0.019	0.000	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.

The table 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 $|\theta|$. 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. This is due to the fact that one does not know that θ is 0 and therefore adds more lagged autocovariances, increasing the risk of a negative estimate.

S.5.3 Asymmetric distribution

In Table S5, we study the size properties of the DM test for various estimates of σ_T assuming standard and fixed-smoothing asymptotics for the case in which the innovations $(v_{1t}, v_{2t})'$ are drawn from a standardized χ^2_5.

Table S5: Size of tests with asymmetric data generating process

Standard asymptotics

	T=40					T=120				
		WCE		WPE			WCE		WPE	
q	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	DM	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$
1	0.063	0.082	0.096	0.083	0.066	0.052	0.068	0.075	0.078	0.060
2	0.084	0.101	0.109	0.084	0.074	0.052	0.068	0.085	0.081	0.060
3	0.105	0.111	0.110	0.083	0.081	0.065	0.090	0.084	0.079	0.066
4	0.126	0.125	0.115	0.081	0.090	0.071	0.100	0.090	0.081	0.067
5	0.168	0.146	0.129	0.093	0.108	0.085	0.119	0.097	0.084	0.080

Fixed-smoothing asymptotics

	T=40				T=120			
	WCE		WPE		WCE		WPE	
q	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$	$\lfloor T^{1/3} \rfloor$	$\lfloor T^{1/2} \rfloor$
1	0.048	0.042	0.038	0.041	0.055	0.046	0.040	0.045
2	0.063	0.044	0.034	0.045	0.055	0.050	0.045	0.045
3	0.072	0.049	0.037	0.053	0.074	0.053	0.045	0.049
4	0.084	0.050	0.034	0.057	0.083	0.054	0.043	0.053
5	0.100	0.060	0.038	0.069	0.099	0.060	0.046	0.059

Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size for various MA(q) processes with $\theta = 0.75$ and innovations $(v_{1t}, v_{2t})'$ drawn from a standardized χ^2_5. The top panel uses standard normal asymptotics and the bottom panel uses fixed-smoothing asymptotics for alternative estimates of the long run variance. For the WCE, we use the truncated rectangular kernel as in DM and the Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$ and $M = \lfloor T^{1/2} \rfloor$. For the WPE, we use the Daniell kernel with $m = \lfloor T^{1/3} \rfloor$ and $m = \lfloor T^{1/2} \rfloor$.

Results are in line with the main results in Tables S2–S2. In particular, we can observe severe size distortions for the tests with standard asymptotics for both sample sizes. Fixed-smoothing asymptotics always improves the empirical size, yielding results close to the prescribed 5%. As in Table S2, we find that the bandwidth $M = \lfloor T^{1/3} \rfloor$ in the WCE-B still yields some size distortion, even when fixed-b asymptotics 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. As for the case with Gaussian innovations, with fixed-b asymptotics it seems desirable to choose the bandwidth $M = \lfloor T^{1/2} \rfloor$, while with fixed-$m$ asymptotics it seems desirable to choose the bandwidth $m = \lfloor T^{1/3} \rfloor$.

S.5.4 Automatic bandwidth selection

In Table S5, we study the application of the automatic bandwidth selection of Newey and West (1994), when $\theta = 0.75$. We compare the performance for the naive $M = \lfloor T^{1/3} \rfloor$ bandwidth (already available in Table S2) against the Newey and West (1994) estimate with prewhitening as in Newey and West (1994), and against a third estimate in which the same procedure is applied, but without prewhitening.

Table S5: 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	0.089	0.129	0.125	0.068	0.074	0.079
2	0.106	0.122	0.130	0.074	0.066	0.082
3	0.120	0.110	0.129	0.086	0.067	0.085
4	0.130	0.107	0.135	0.097	0.065	0.092
5	0.153	0.108	0.140	0.111	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).

In general, using the Newey and West (1994) 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 Newey and West (1994) procedure would not fare well when the dependence is relatively weak. Table S5, 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.

S.5.5 Minimum MSE bandwidth for the WPE

In this section, 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 approximate 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.

Notice that the minimum MSE bandwidth trades off bias and variance, but this may not be the best criterion for application in tests, as in testing we are looking at different properties, namely, minimum size distortion and maximum power. With standard asymptotics, both the bias and variance of the estimate of the long run variance cause size distortion in the test, whereas with fixed-smoothing asymptotics the effect of the variance of the estimate of the long run variance is accounted for by the change in the distribution of the test statistic, and we are only concerned about the effect due

Table S5: 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

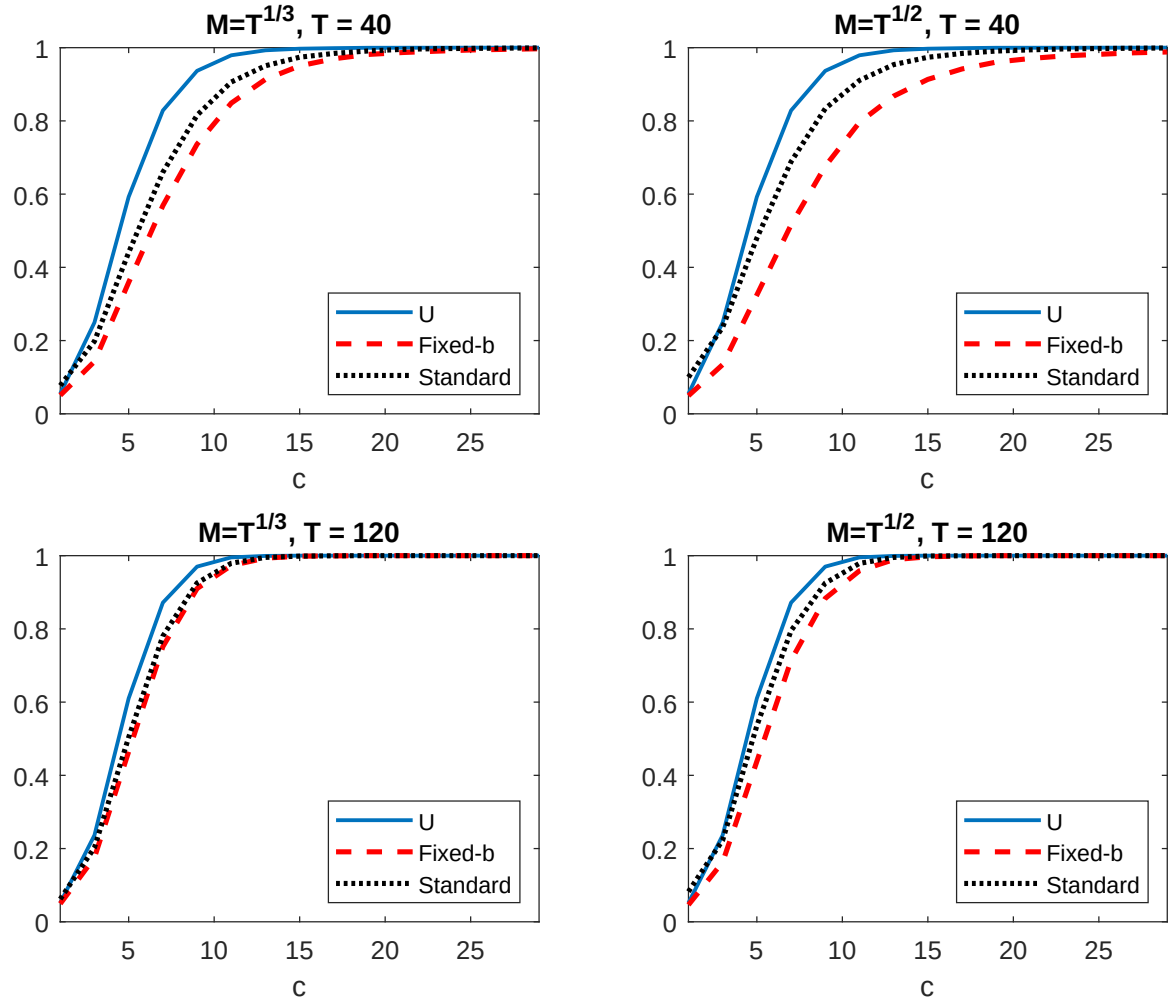
Note: empirical rejection frequencies for tests of equal predictive ability at 5% nominal size for various MA(q) processes with $\theta = 0.75$ using the WPE estimator of the long run variance and the feasible minimum MSE bandwidth. The test with standard asymptotics uses standard normal critical values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

to the (lower order) bias. This bias is stronger the larger is the bandwidth m, as with larger bandwidths periodograms that are more distant from frequency zero are used to estimate the spectral density at frequency zero. Bandwidths proportional to $T^{4/5}$ are therefore more prone to causing size distortion in testing. Indeed, results in Table S5 indicate that, as for the Newey and West (1994) 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 larger than $\lfloor T^{1/4} \rfloor$, $\lfloor T^{1/3} \rfloor$ or $\lfloor T^{1/2} \rfloor$ used in Table S2, resulting in a larger bias. For this reason, we do not recommend using this bandwidth.

S.5.6 Power comparison with standard asymptotics

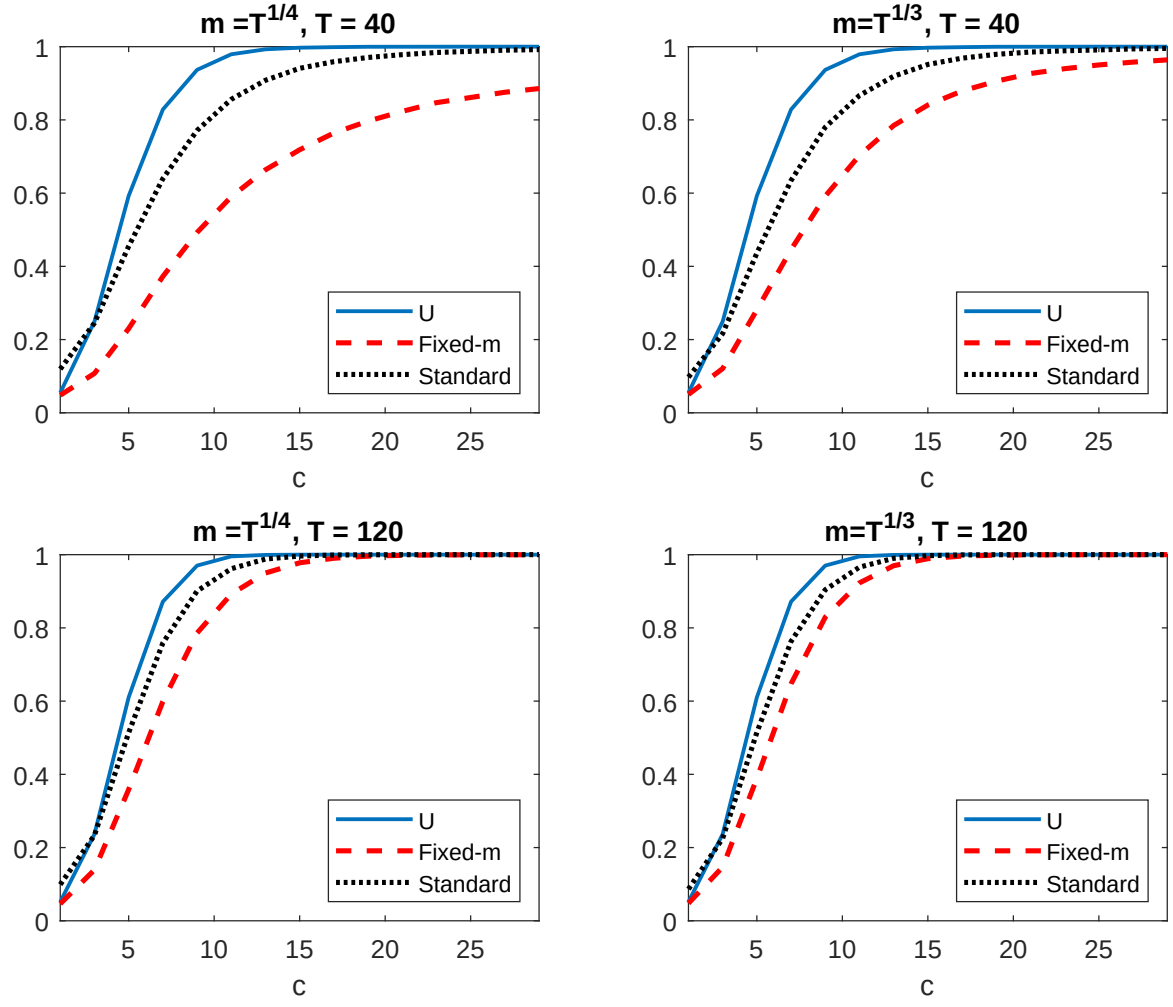
In this section, we analyze the power of the tests under standard asymptotics when the long run variance is estimated. Results in Figure S4 refer to the case in which a WCE estimate of the long run variance with Bartlett kernel is used, with $M = \lfloor T^{1/3} \rfloor$ (left panels) and $M = \lfloor T^{1/2} \rfloor$ (right panels). We take again as benchmark the limit

Figure S4: Finite sample local power using WCE



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $c/\sqrt{T}$. 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, the test statistic uses the WCE estimate of the long run variance with Bartlett kernel with $M = \lfloor T^{1/3} \rfloor$ (left panels) and $M = \lfloor T^{1/2} \rfloor$ (right panels). We use standard critical values (Standard) and fixed-b critical values (Fixed-b).

Figure S5: Finite sample local power using WPE



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $c/\sqrt{T}$. 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, the test statistic uses the WPE estimate of the long run variance with Daniell kernel with $m = \lfloor T^{1/4} \rfloor$ (left panels) and $m = \lfloor T^{1/3} \rfloor$ (right panels). We use standard critical values (Standard) and fixed-m critical values (Fixed-m).

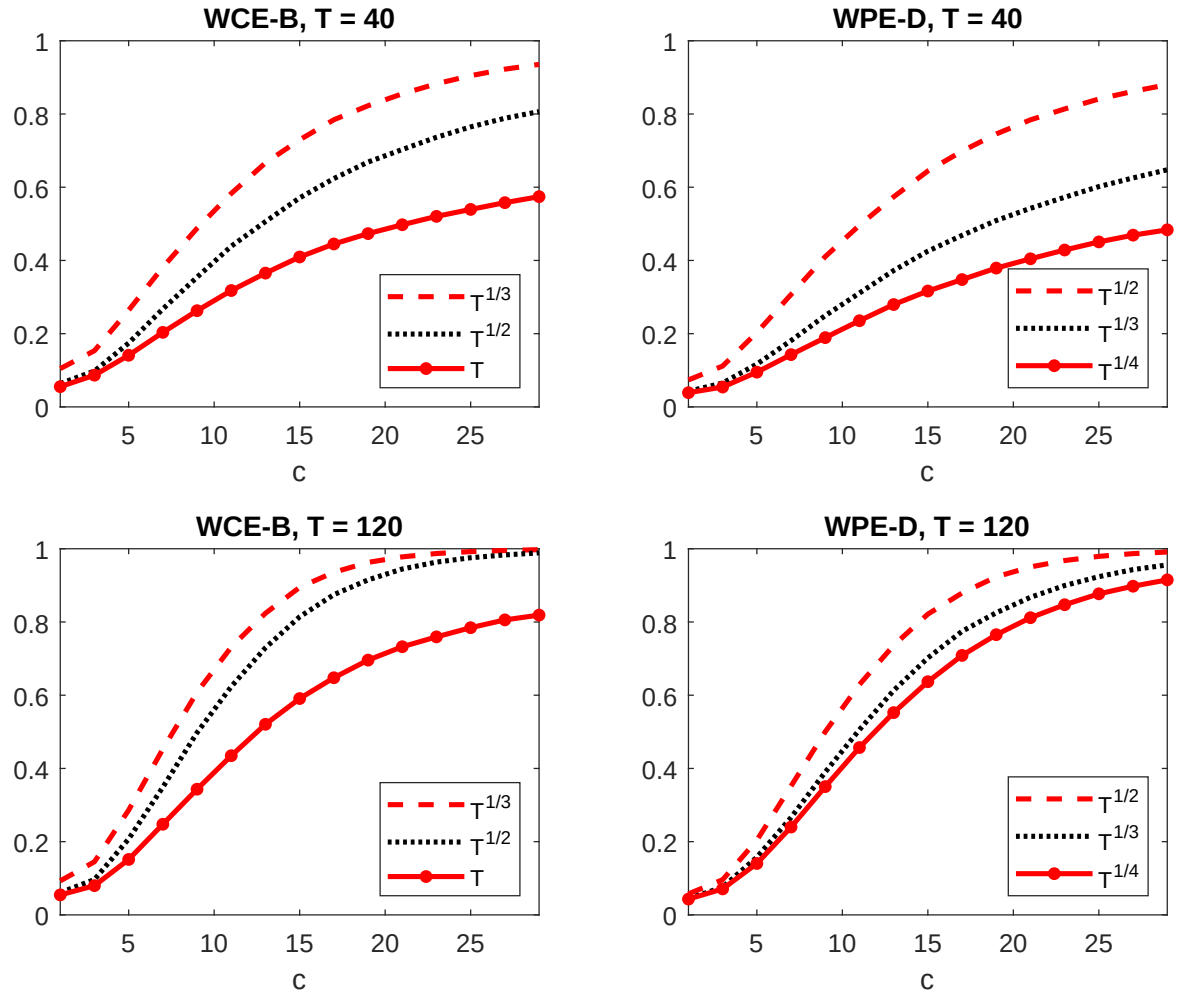
local power function obtained from the normal distribution, and we compare it to the simulated local power functions both when standard asymptotics and when fixed-b are used. Using standard asymptotics we would be misled into thinking that we attain power comparable to the limit benchmark. However, this is spurious, as we can see from the size distortion. Results are similar when using a WPE estimate of the long run variance, as shown in Figure S5 where we use a WPE estimate of the long run variance with Daniell kernel with $m = \lfloor T^{1/4} \rfloor$ (left panels) and $m = \lfloor T^{1/3} \rfloor$ (right panels).

S.5.7 Power with autocorrelation

In this section, we analyze the power of the test of equal forecast accuracy when autocorrelation is preserved under the alternative. To this end we set $k = 1 + c/\sqrt{T}$ for $c = 1 \dots 30$ as in the power study in S.3 but we set $q = 5$ and $\theta = 0.75$ instead.

In Figure S6, we report the simulated power of the tests, computed using critical values from fixed-smoothing asymptotics. The power ranking of the proposed methods when the processes are autocorrelated is the same as in Figure S1, with smaller bandwidths M in WCE estimates (or, larger bandwidths m in WPE estimates) associated with higher power, but notice again that some power may be spurious when $M = \lfloor T^{1/3} \rfloor$, as we already discussed commenting on the size study.

Figure S6: Finite sample local power with dependent innovations



The figure displays empirical rejection frequencies at 5% nominal size for deviations from the null by $c/\sqrt{T}$ and MA(5) dependent innovations. 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 $m = \lfloor T^{1/4} \rfloor$.

References

Abadir, Karim M, Walter Distaso, and Liudas Giraitis (2009) 'Two estimators of the long-run variance: beyond short memory.' *Journal of Econometrics* 150(1), 56–70

Clark, Todd E (1999) 'Finite-sample properties of tests for equal forecast accuracy.' *Journal of Forecasting* 18(7), 489–504

Delgado, Miguel A, and Peter M Robinson (1996) 'Optimal spectral bandwidth for long memory.' *Statistica Sinica* pp. 97–112

Gonçalves, Sílvia, and Timothy J Vogelsang (2011) 'Block bootstrap hac robust tests: The sophistication of the naive bootstrap.' *Econometric Theory* 27(4), 745–791

Harvey, David I, Stephen J Leybourne, and Emily J Whitehouse (2017) 'Forecast evaluation tests and negative long-run variance estimates in small samples.' *International Journal of Forecasting* 33(4), 833–847

Hualde, Javier, and Fabrizio Iacone (2015) 'Autocorrelation robust inference using the daniell kernel with fixed bandwidth.' Technical Report, Department of Economics, University of York

Kiefer, Nicholas M, and Timothy J Vogelsang (2002) 'Heteroskedasticity–autocorrelation robust standard errors using the bartlett kernel without truncation.' *Econometrica* 70(5), 2093–2095

—— (2005) 'A new asymptotic theory for heteroskedasticity-autocorrelation robust tests.' *Econometric Theory* 21(6), 1130–1164

Lazarus, Eben, Daniel J Lewis, James H Stock, and Mark W Watson (2018) 'Har inference: Recommendations for practice.' *Journal of Business & Economic Statistics* 36(4), 541–559

Newey, Whitney K, and Kenneth D West (1994) 'Automatic lag selection in covariance matrix estimation.' *The Review of Economic Studies* 61(4), 631–653

Phillips, Peter CB (2005) 'HAC estimation by automated regression.' *Econometric Theory* 21(1), 116–142

Politis, Dimitris N, and Joseph P Romano (1994) 'The stationary bootstrap.' *Journal of the American Statistical Association* 89(428), 1303–1313

Sun, Yixiao (2013) 'A heteroskedasticity and autocorrelation robust f test using an orthonormal series variance estimator.' *The Econometrics Journal* 16(1), 1–26

___ (2014) 'Fixed-smoothing asymptotics in a two-step generalized method of moments framework.' *Econometrica* 82(6), 2327–2370

___ (2018) 'Comment.' *Journal of Business & Economic Statistics* 36(4), 565–568

Sun, Yixiao, Peter CB Phillips, and Sainan Jin (2008) 'Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing.' *Econometrica* 76(1), 175–194