Journal of Econometrics 186 (2015) 160–177

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Nested forecast model comparisons: A new approach to testing equal accuracy

Todd E. Clark[a,*], Michael W. McCracken[b]

[a] Economic Research Department, Federal Reserve Bank of Cleveland, P.O. Box 6387, Cleveland, OH 44101, United States

[b] Research Division, Federal Reserve Bank of St. Louis, P.O. Box 442, St. Louis, MO 63166, United States

ARTICLE INFO

Article history:
Received 4 January 2012
Received in revised form
19 December 2013
Accepted 27 June 2014
Available online 22 July 2014

JEL classification:
C53
C12
C52

Keywords:
Mean square error
Prediction

ABSTRACT

We develop methods for testing whether, in a finite sample, forecasts from nested models are equally accurate. Most prior work has focused on a null of equal accuracy in population — basically, whether the additional coefficients of the larger model are zero. Our asymptotic approximation instead treats the coefficients as non-zero but small, such that, in a finite sample, forecasts from the small and large models are expected to be equally accurate. We derive the limiting distributions of tests of equal mean square error, and develop a bootstrap for inference. Simulations show that our procedures have good size and power properties.

1. Introduction

In this paper we examine the asymptotic and finite-sample properties of bootstrap-based tests of equal accuracy of out-of-sample forecasts from a baseline nested model and an alternative nesting model. In our analysis, we address two forms of the null hypothesis of equal predictive ability. One hypothesis, considered in such studies as Clark and McCracken (2001, 2005a), Corradi and Swanson (2002), Inoue and Kilian (2004), and McCracken (2007), is that the models have equal population-level predictive ability. This situation arises when the coefficients associated with the additional predictors in the nesting model are zero and hence at the population level, the forecast errors are identical and thus the models have equal predictive ability.

However, this paper focuses on a different null hypothesis, one that arises when some of the additional predictors have non-zero coefficients associated with them, but the marginal predictive content is small. In this case, addressed in Trenkler and Toutenburg (1992), Giacomini and White (2006), Hjalmarsson (2009) and Clark and McCracken (2009), the two models can have equal predictive ability at a fixed forecast origin (say time R) due to a bias–variance trade-off between a more accurately estimated, but misspecified, nested model and a correctly specified, but imprecisely estimated, nesting model. Building upon this insight, we derive the asymptotic distributions associated with standard out-of-sample tests of equal predictive ability between estimated models with weak predictors. We then develop a bootstrap-based method for imposing the null of equal predictive ability upon these distributions and conducting asymptotically valid inference. In our results, the forecast models may be estimated either recursively or with a rolling sample. Giacomini and White (2006) use a different asymptotic approximation to testing equal forecast accuracy in a given sample, but their asymptotics apply only to models estimated with a rolling window of fixed and finite width.

Our approach to modeling weak predictors is identical to the standard Pitman drift used to analyze the power of in-sample tests against small deviations from the null of equal population-level predictive ability. It has also been used by Inoue and Kilian (2004) in the context of analyzing the power of out-of-sample tests. In that sense, some (though not all) of our analytical results are quite similar to those in Inoue and Kilian (2004).

We differ, though, in our focus. While Inoue and Kilian (2004) are interested in examining the power of out-of-sample tests against the null of equal population-level predictive ability, we are interested in using out-of-sample tests to test the null hypothesis of equal finite sample predictive ability. This distinction arises because the estimation error associated with estimating unknown

[*] Corresponding author.
E-mail addresses: todd.clark@clev.frb.org (T.E. Clark), michael.w.mccracken@stls.frb.org (M.W. McCracken).

http://dx.doi.org/10.1016/j.jeconom.2014.06.016
0304-4076/© 2014 Elsevier B.V. All rights reserved.

regression parameters can cause a misspecified, restricted model to be as accurate or more accurate than a correctly specified unrestricted model when the additional predictors are imprecisely estimated (or, in our terminology, are "weak"). We use Pitman drift simply as a tool for constructing an asymptotic approximation to the finite sample problem associated with estimating a regression coefficient when the marginal signal associated with it is small.

The lengthy literature evaluating direct, multi-step (DMS) forecasts from nested models indicates our results for these forecasts should be useful to many researchers. Applications considering DMS forecasts from nested linear models include, among others: many of the studies cited above; Diebold and Rudebusch (1991); Mark (1995); Kilian (1999); Lettau and Ludvigson (2001); Stock and Watson (2003); Bachmeier and Swanson (2005); Butler et al. (2005); Cooper and Gulen (2006); Giacomini and Rossi (2006); Guo (2006); Rapach and Wohar (2006); Bruneau et al. (2007); Bordo and Haubrich (2008); Inoue and Rossi (2008); Molodtsova and Papell (2009); Chen et al. (2010); and Ferreira and Santa-Clara (2011).

The remainder proceeds as follows. Section 2 uses a very simple illustrative data-generating process to flesh out the intuition behind our analysis and results — including the precise nature of the null hypothesis, the bootstrap algorithm, and the validity of the bootstrap. With that foundation, the paper then turns to the more general case. Section 3 introduces the notation, assumptions, and presents our theoretical results and bootstrap for testing the null of equal forecast accuracy in the finite sample. Proofs are provided in a supplementary online appendix (see Appendix A). Section 4 presents Monte Carlo results on the finite-sample performance of the asymptotics and the bootstrap. Section 5 applies our tests to evaluate the predictability of US stock returns and core PCE inflation. Section 6 concludes.

2. An illustrative example

We begin by using a simple example to first clarify how our results differ from those obtained in Giacomini and White (2006) and to then illustrate our essential ideas. This example uses a simple DGP: $y_{t+1} = \mu + u_{t+1}$, where μ is non-stochastic and u_{t+1} forms a homoskedastic martingale difference sequence with variance σ^2.

2.1. *Simple version of our test of equal forecast accuracy in the finite sample*

Consider comparing the finite sample forecast accuracy of two nested models, with accuracy measured under quadratic loss. In this simple example, Model 0 is a no-change model, such that $\hat{y}_{0,t+1} = 0$. Model 1 is an OLS-estimated location model, corresponding to the form of the DGP: $\hat{y}_{1,t+1} = \bar{y}_t$, where $\bar{y}_t$ equals $t^{-1}\sum_{s=1}^{t} y_s$ and $R^{-1}\sum_{s=t-R+1}^{t} y_s$ under the recursive (expanding window) or rolling window estimation schemes, respectively. From these models, we produce a total of P forecasts, take the difference in the squared forecast errors, and average across the forecast origins $t = R, \ldots, R+P-1$. The expectation of this difference in average squared errors, $P^{-1}\sum_{t=R}^{R+P-1} E(\hat{u}_{0,t+1}^2 - \hat{u}_{1,t+1}^2)$, equals $\mu^2 - P^{-1}\sum_{t=R}^{R+P-1}\frac{\sigma^2}{t}$ and $\mu^2 - \frac{\sigma^2}{R}$ for the recursive and rolling schemes, respectively. The μ^2 (bias) term of the difference in mean square errors (MSEs) arises due to misspecification in model 0, while the second term arises due to marginally greater estimation risk (variance) in model 1. We say the two models are expected to exhibit equal finite sample accuracy when the tradeoff between the bias and variance terms implies $P^{-1}\sum_{t=R}^{R+P-1} E(\hat{u}_{0,t+1}^2 - \hat{u}_{1,t+1}^2) = 0$.

The goal then becomes to develop the distribution of a statistic when this moment condition forms the null hypothesis. To avoid strong assumptions about the predictors and the model errors, we focus on asymptotic distributions. Unfortunately, as the number of forecasts P and initial sample size R diverge to infinity, $P^{-1}\sum_{t=R}^{R+P-1} E(\hat{u}_{0,t+1}^2 - \hat{u}_{1,t+1}^2)$ converges to μ^2 and hence in large samples the two models will be equally accurate only in the trivial case in which $\mu = 0$. Since the null of interest is one of equal finite sample accuracy, we cannot simply proceed with this formulation of the DGP and the null implication that $\mu = 0$. The reason is that with $\mu = 0$, the models will not be equally accurate in the finite sample; if $\mu = 0$, model 1 has to be less accurate than model 0 in the finite sample, because model 1 introduces estimation risk of a parameter that is 0 in population. Put another way, when $\mu = 0$ there can be no bias–variance tradeoff that makes the models equally accurate in the finite sample.

To develop a test of equal accuracy in the finite sample, Giacomini and White (2006) worked with a different null hypothesis. They departed from a null hypothesis formulated using $\lim_{R,P\to\infty} P^{-1}\sum_{t=R}^{R+P-1} E(\hat{u}_{0,t+1}^2 - \hat{u}_{1,t+1}^2) = 0$ because, if the asymptotics allow R to increase, the estimation risk component σ^2/R converges to zero, precluding the bias-tradeoff needed for the models to be equally accurate in forecasting in the finite sample. They instead assumed the estimation window size R to be fixed and finite, with model parameters and forecasts produced using a rolling window scheme. They then formulated the null hypothesis as $\lim_{P\to\infty} P^{-1}\sum_{t=R}^{R+P-1} E(\hat{u}_{0,t+1}^2 - \hat{u}_{1,t+1}^2) = 0$. With the model estimation sample size held fixed and this version of the null hypothesis of equal accuracy in the finite sample, the null-implied hypothesis $\mu^2 - \frac{\sigma^2}{R} = 0$ is viable even when P diverges to infinity. Note that in this asymptotic framework, parameter estimation error remains "large" because the parameter estimates do not converge in probability.

In this paper, to permit environments in which the parameter estimates are estimated using an expanding, or recursive, window as we proceed across forecast origins, we must take a different approach to the asymptotics and null hypothesis. The main problem is that, with recursive estimation and standard large R, P asymptotics, estimation error eventually becomes "small" in the sense that the parameter estimates converge in probability. To avoid this problem, and yet still allow a bias–variance tradeoff to exist, we model the bias as being equally small. In the context of our current example, consider modeling the unconditional mean μ as being local-to-zero such that $\mu = \mu_w/R^{1/2}$. Let $\lim_{P,R\to\infty} P/R = \lambda_P \in (0,\infty)$. If we then restate the null hypothesis as $\lim_{R,P\to\infty}\sum_{t=R}^{R+P-1} E(\hat{u}_{0,t+1}^2 - \hat{u}_{1,t+1}^2) = 0$, we find that this is equivalent to $\lambda_P\mu_w^2 - \ln(1+\lambda_P)\sigma^2 = 0$ and $\mu_w^2 - \sigma^2 = 0$ under the recursive and rolling estimation schemes, respectively.

Under this null hypothesis we derive the asymptotic distribution of two tests of equal mean square error. The simpler one we will focus on in this example is an F-type test of equal MSE, given by

$$\text{MSE-F} = P \times \frac{\text{MSE}_0 - \text{MSE}_1}{\text{MSE}_1}.$$

Because the asymptotic distribution of the statistic is nonstandard, we use a bootstrap to obtain asymptotic critical values. In this simple example, the bootstrap proceeds as follows.

1. (a) Estimate the model $y_s = m + u_{1,s}$, $s = 1, \ldots, T$, by OLS. Save the residuals $\hat{u}_{1,s}$ and residual variance $\hat{\sigma}_1^2$. (b) Estimate the ridge regression

$$\tilde{\mu}_{w,T} = \arg\min_{m}\sum_{s=1}^{T}(y_s - m)^2 \quad \text{s.t. } m^2 = \hat{d}/R,$$

where $\hat{\lambda}_P = P/R$ and $\widehat{d}$ equals $\frac{\ln(1+\widehat{\lambda}_P)}{\widehat{\lambda}_P}\hat{\sigma}_1^2$ and $\hat{\sigma}_1^2$ for the recursive and rolling schemes, respectively. Save $\tilde{\mu}_{w,T}$. This solution is not unique and can take the two values: $\tilde{\mu}_{w,T} = \pm \hat{d}^{1/2} R^{-1/2} sign(\bar{y}_T)$. Pick one solution and call it $\tilde{\mu}_{w,T}$.

2. Define the bootstrapped data $y^*_s = \tilde{\mu}_{w,T} + \hat{u}^*_{1,s}$, where $\hat{u}^*_{1,s} = \hat{u}_{1,s}\eta_s$ for a time series of *iid* $N(0, 1)$ increments η_s.
3. Using the time series of artificial data, construct forecasts and an estimate of the test statistic (MSE-F) as if this were the original data.
4. Repeat steps 2 and 3 a large number of times: $j = 1, \ldots, N$.
5. Reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100 - \alpha)$%-ile of the empirical distribution of the simulated test statistics.

2.2. Intuition for our approach and results

To establish the intuition behind our null hypothesis and the validity of the bootstrap, it is helpful to consider an alternative, familiar test of in-sample predictive ability. Continuing with the simple location model described above, the standard Wald statistic for testing in-sample predictive ability can be expressed as

$$GC(T) = T\frac{\hat{\sigma}_0^2 - \hat{\sigma}_1^2}{\hat{\sigma}_1^2} = \frac{\left(T^{-1/2}\sum_{s=1}^{T} u_s + T^{1/2}\mu\right)^2}{\hat{\sigma}_1^2},$$

where $\hat{\sigma}_i^2$, $i = 0, 1$, denote residual variances for the baseline and nesting models 0 and 1, respectively. Consider first the case in which the mean parameter μ is not local-to-zero — that is, μ is a constant (and not a function of T). In this case, the Wald test of in-sample predictive ability is asymptotically non-central chi-square with 1 degree of freedom and non-centrality parameter $T\mu^2/\sigma^2$, which, unless μ or σ^2 depends on T, will diverge to infinity and thereby provide a consistent test of H_0: $\mu = 0$ against the alternative H_A: $|\mu| = 0$.

However, if we are interested in testing equal forecasting accuracy in a finite sample we might consider instead a null hypothesis of the form H_0: $E(\hat{u}^2_{0,T+1} - \hat{u}^2_{1,T+1}) = 0$ versus an alternative H_A: $E(\hat{u}^2_{0,T+1}) > E(\hat{u}^2_{1,T+1})$.[1] However, to do so under weak assumptions on the observables, we first modify the null hypothesis a bit, along the lines described in the previous subsection, to H_0: $\lim_{T\to\infty} TE(\hat{u}^2_{0,T+1} - \hat{u}^2_{1,T+1}) = 0$ and then model the mean as local-to-zero such that $\mu = \mu_w/T^{1/2}$. Together we find that the null implies the restriction $\mu_w^2 = \sigma^2$.

With this change in the DGP to make μ local-to-zero instead of a fixed constant, the standard Wald statistic becomes

$$GC(T) = T\frac{\hat{\sigma}_0^2 - \hat{\sigma}_1^2}{\hat{\sigma}_1^2} = \frac{\left(T^{-1/2}\sum_{s=1}^{T} u_s + \mu_w\right)^2}{\hat{\sigma}_1^2}$$
$$= \left(T^{-1/2}\sum_{s=1}^{T} u_s/\hat{\sigma}_1 + \mu_w/\hat{\sigma}_1\right)^2.$$

Under the null of interest, $\mu_w^2 = \sigma^2$, so, asymptotically, the term $\mu_w/\hat{\sigma}_1$ simplifies to just $sign(\mu_w)$. In turn, the Wald statistic can be expressed as

$$GC(T) = \left(T^{-1/2}\sum_{s=1}^{T} u_s/\sigma + sign(\mu_w)\right)^2 + o_p(1),$$

which is asymptotically non-central chi-square with one degree of freedom and non-centrality parameter equal to 1. If we are most interested in identifying the alternative H_A: $\lim_{T\to\infty} TE(\hat{u}^2_{0,T+1} - \hat{u}^2_{1,T+1}) > 0$, then inference can be conducted using the upper tail of the $\chi^2(1, 1)$ distribution.

With the illustrative DGP and the GC-based test of equal forecast accuracy, inference is trivial using tables of non-central chi-square critical values. However, as we establish in the general results below, the distribution of the out-of-sample version of this test is not so trivially tabulated, because of the presence of unknown nuisance parameters. So suppose one uses the bootstrap detailed above, modified as needed for the in-sample GC-based test instead of the out-of-sample MSE $-$ F-based test, as follows.

1. (a) Estimate the model $y_s = m + u_{1,s}$, $s = 1, \ldots, T$ by OLS. Save the residuals $\hat{u}_{1,s}$ and residual variance $\hat{\sigma}_1^2$. (b) Estimate the ridge regression
$$\tilde{\mu}_{w,T} = \arg\min_m \sum_{s=1}^{T}(y_s - m)^2 \quad \text{s.t. } m^2 = \hat{\sigma}_1^2/T$$
and save $\tilde{\mu}_{w,T}$. This solution is not unique and can take the two values: $\tilde{\mu}_{w,T} = \pm\hat{\sigma}_1 T^{-1/2} sign(\bar{y}_T)$. Pick one solution and call it $\tilde{\mu}_{w,T}$.
2. Define the bootstrapped data $y^*_s = \tilde{\mu}_{w,T} + \hat{u}^*_{1,s}$ where $\hat{u}^*_{1,s} = \hat{u}_{1,s}\eta_s$ for a time series of *iid* $N(0, 1)$ increments η_s.
3. Using the time series of artificial data, form the GC statistic as if this were the original data.
4. Repeat steps 2 and 3 a large number of times: $j = 1, \ldots, N$.
5. Reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100 - \alpha)$%-ile of the empirical distribution of the simulated test statistics.

The validity of the bootstrap obtains because the bootstrapped critical value is consistent for that from the $\chi^2(1, 1)$ distribution. To see how the bootstrap works, assume that we choose $\tilde{\mu}_{w,T} = \hat{\sigma}_1 sign(\bar{y}_T)T^{-1/2}$ in step 1. Straightforward algebra gives us

$$GC^*(T) = T\frac{\hat{\sigma}_0^{*2} - \hat{\sigma}_1^{*2}}{\hat{\sigma}_1^{*2}} = \frac{\left(T^{-1/2}\sum_{s=1}^{T} \hat{u}^*_{1,s} + T^{1/2}\tilde{\mu}_{w,T}\right)^2}{\hat{\sigma}_1^{*2}}$$
$$= \frac{\left(T^{-1/2}\sum_{s=1}^{T} \hat{u}^*_{1,s} + \hat{\sigma}_1 sign(\bar{y}_T)\right)^2}{\hat{\sigma}_1^{*2}}$$
$$= \left(T^{-1/2}\sum_{s=1}^{T} \hat{u}^*_{1,s}/\hat{\sigma}^*_1 + \hat{\sigma}_1 sign(\bar{y}_T)/\hat{\sigma}^*_1\right)^2.$$

A bit of algebra shows that both $\hat{\sigma}^*_1$ and $\hat{\sigma}_1^2$ converge in probability to σ^2. Note also that $T^{-1/2}\sum_{s=1}^{T}\hat{u}^*_{1,s}/\sigma = T^{-1/2}\sum_{s=1}^{T}\hat{u}_{1,s}\eta_s/\sigma$ is asymptotically standard normal (see Goncalves and Kilian, 2007). If (a) μ is a non-zero constant that does not depend on T, then $sign(\bar{y}_T) = sign(\mu) + o_p(1)$. Moreover, if (b) $\mu = \mu_w/T^{1/2}$, then $sign(\bar{y}_T) = sign(\mu_w + T^{-1/2}\sum_{s=1}^{T} u_s)$, which is 1 with probability $1 - \Phi(-\mu_w/\sigma)$ in large samples and -1 otherwise. Note also that regardless of case (a) or (b), $sign(\bar{y}_T)$ is independent of $T^{-1/2}\sum_{s=1}^{T}\hat{u}_{1,s}\eta_s/\sigma$ due to the *i.i.d.* nature of the η_s.

Together, these results imply that

$$GC^*(T) = \left(T^{-1/2}\sum_{s=1}^{T} \hat{u}^*_s/\sigma + sign(\bar{y}_T)\right)^2 + o_{p^*}(1).$$

If μ is a non-zero constant, then $GC^*(T)$ is asymptotically $\chi^2(1, 1)$ irrelevant of the sign of $sign(\bar{y}_T)$ or, equivalently, $sign(\mu)$. If $\mu = \mu_w/T^{1/2}$, $sign(\bar{y}_T)$ is random in large samples, but because it is independent of $T^{-1/2}\sum_{s=1}^{T}\hat{u}^*_s/\sigma$, we still find that $GC^*(T)$ is

[1] We could also test the null H_0:$E(\hat{u}^2_{0,T+1}) < E(\hat{u}^2_{1,T+1})$ but want to keep this discussion focused on the leading alternative of interest.

asymptotically $\chi^2(1, 1)$. Note that if we had started this derivation using $\tilde{\mu}_{w,T} = -\hat{\sigma}_1 sign(\bar{y}_T)T^{-1/2}$, the same arguments would hold.

While we are able to establish that the bootstrap is valid, at no point do we suggest that $\tilde{\mu}_{w,T}$ is consistent for any "true" value of μ or μ_w. The bootstrap works because there exist DGPs such that the distribution of $GC(T)$ does not depend on the "true" value of μ or μ_w so much as it depends on a known, consistently estimable function of it. In this example, that function is $\mu_w^2 = \sigma^2$, which allows us to consistently estimate μ_w^2 despite not being able to identify the sign of μ_w. But that is sufficient because $GC(T)$ is asymptotically $\chi^2(1, 1)$ if $sign(\mu_w) = 1$ or $sign(\mu_w) = -1$.

Building on the results for the illustrative example above, in subsequent sections we develop a test of equal finite sample forecast accuracy in an out-of-sample context with a null hypothesis akin to that considered in Giacomini and White (2006) but which permits recursive estimation windows. As in the simple case above, we first derive various asymptotic distributions using a local-to-zero framework for the DGP. We then establish conditions under which these distributions are invariant to knowledge of, for example, μ_w, but instead depend on a known function of μ_w that is consistently estimable. Finally, since the distributions are nonstandard we provide a bootstrap approach to inference that works in those special cases in which the distribution of the statistic is invariant to knowledge of μ_w.

In all instances we restrict attention to linear models estimated by OLS. As we noted in the introduction, the linear-OLS framework covers a large fraction of the literature in applied forecasting and forecast evaluation.

3. Theoretical results

In this section, after detailing the necessary notation and test statistics, we present asymptotic results for tests of equal accuracy applied to forecasts from two nested models in the presence of weak predictive ability. The section then describes our proposed bootstrap and proves its validity. We focus all of this presentation on results for the recursive estimation and forecasting scheme. The last subsection summarizes the changes in moments and functions that apply under a rolling estimation and forecasting scheme.

3.1. Environment

The sample of observations $\{y_t, x'_{1,t}\}_{t=1}^T$ includes a scalar random variable y_t to forecast at a horizon of τ periods ahead and a $(k \times 1)$ vector of predictors $x_{1,t}$.[2] The vector of predictors contains one set of variables (denoted $x_{0,t}$, with k_0 elements) included in the null model and another distinct set of variables (denoted $x_{w,t}$, with k_w elements) with weak predictive content: $x_{1,t} = (x'_{0,t}, x'_{w,t})'$. The vector of predictors $(x_{1,t})$ may include lags of the dependent variable. At each origin of forecasting $t = R, \ldots T = R+P-\tau$, forecasts of $y_{t+\tau}$, $\tau \geq 1$, are generated from OLS-estimated linear parametric models of the form $x'_{i,t}\beta_i$, $i = 0, 1$, where the restricted model 0 includes only $x_{0,t}$ as predictors and the unrestricted model 1 includes both $x_{0,t}$ and $x_{w,t}$ as predictors. The forecast sample size is P.

The possibility of weak predictors is modeled using a sequence of linear data-generating processes (DGPs) of the form[3]

$$y_{t+\tau} = x'_{1,t}\beta_1 + u_{t+\tau} = x'_{0,t}\beta_0 + x'_{w,t}(R^{-1/2}\beta_w) + u_{t+\tau}, \qquad (1)$$
$$Ex_{1,t}u_{t+\tau} \equiv Eh_{1,t+\tau} = 0 \quad \text{for all } t = 1, \ldots, R, \ldots R + P - \tau.$$

Turning to notation needed below, we denote the loss associated with the τ-step ahead forecast errors as $u^2_{i,t+\tau} = (y_{t+\tau} - x'_{i,t}\beta_i)^2$, $i = 0, 1$, for the restricted and unrestricted models, respectively. Let $H_i(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}u_{s+\tau}) = (t^{-1}\sum_{s=1}^{t-\tau} h_{i,s+\tau})$, $B_i(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{i,s}x'_{i,s})^{-1}$, and $B_i = \lim_{R\to\infty}(Ex_{R,i,s}x'_{R,i,s})^{-1}$, for $i = 0, 1$. For $U_{R,t} = (h'_{R,1,t+\tau}, vec(x_{R,1,t}x'_{R,1,t})')'$, $\Omega_j = \lim_{R\to\infty} R^{-1}\sum_{t=1}^{R} E(U_{R,t}U'_{R,t-j})$ for all $j \geq 0$, $\Omega_{11,j}$ is the upper block-diagonal element of Ω_j, and $V = \sum_{j=-\tau+1}^{\tau-1}\Omega_{11,j}$. Define selection matrices that pull out the elements of x_1 associated with x_0 and x_w, respectively: $J = (I_{k_0\times k_0}, 0_{k_0\times k_w})'$ and $J_w = (0_{k_w\times k_0}, I_{k_w\times k_w})'$. The population residual variance is $\sigma^2 = \lim_{R\to\infty} Eu^2_{R,t+\tau}$. For a $(k_w \times k)$ matrix $\tilde{A}$ satisfying $\tilde{A}'\tilde{A} = B_1^{-1/2}(-J'B_0J + B_1)B_1^{-1/2}$, let $\tilde{h}_{1,t+\tau} = \sigma^{-1}\tilde{A}B_1^{1/2}h_{1,t+\tau}$ and $\tilde{H}_1(t) = \sigma^{-1}\tilde{A}B_1^{1/2}H_1(t)$. Let $F_1 = J'_wB_1J_w$ and $F_1(t) = J'_wB_1(t)J_w$. If we define $\gamma_{\tilde{h}\tilde{h}}(i) = \lim_{R\to\infty} E\tilde{h}_{R,1,t+\tau}\tilde{h}'_{R,1,t+\tau-i}$, then $S_{\tilde{h}\tilde{h}} = \gamma_{\tilde{h}\tilde{h}}(0) + \sum_{i=1}^{\tau-1}(\gamma_{\tilde{h}\tilde{h}}(i) + \gamma'_{\tilde{h}\tilde{h}}(i))$. For notational convenience in the presentation of the results, we define a vector containing 0's (for the $x_{0,t}$ variables) and the weak predictor coefficients: $\delta = (0_{1\times k_0}, \beta'_w)'$. Finally, for any $(m \times n)$ matrix A, let $|A|$ denote the max norm and $tr(A)$ denote the trace.

3.2. Test statistics

In the context of non-nested models, Diebold and Mariano (1995) propose a test for equal MSE based upon the sequence of loss differentials $\hat{L}_{t+\tau} = \hat{u}^2_{0,t+\tau} - \hat{u}^2_{1,t+\tau}$, where $\hat{u}^2_{i,t+\tau} = (y_{t+\tau} - x'_{i,t}\hat{\beta}_{i,t})^2$. If we define $\text{MSE}_i = (P - \tau + 1)^{-1}\sum_{t=R}^{R+P-\tau}\hat{u}^2_{i,t+\tau}$ $(i = 0, 1)$, $\bar{L} = (P - \tau + 1)^{-1}\sum_{t=R}^{R+P-\tau}\hat{L}_{t+\tau} = \text{MSE}_0 - \text{MSE}_1$, $\widehat{\gamma}_{LL}(j) = (P - \tau + 1)^{-1}\sum_{t=R+j}^{R+P-\tau}(\hat{L}_{t+\tau} - \bar{L})(\hat{L}_{t+\tau-j} - \bar{L})$, $\widehat{\gamma}_{LL}(-j) = \widehat{\gamma}_{LL}(j)$, and $\hat{S}_{LL} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\widehat{\gamma}_{LL}(j)$ for some kernel $K(\cdot)$ defined below, the statistic takes the form

$$\text{MSE-}t = (P - \tau + 1)^{1/2} \times \frac{\bar{L}}{\sqrt{\hat{S}_{LL}}}. \qquad (2)$$

Under the null that $x_{w,t}$ has no population-level predictive power for $y_{t+\tau}$, the population difference in MSEs, $E(u^2_{0,t+\tau} - u^2_{1,t+\tau})$, will equal 0 for all t. When $x_{w,t}$ has predictive power, the population difference in MSEs will be positive. Even so, the finite sample difference need not be positive and in fact, for a given sample size (say, $t = R$) the difference in finite sample MSEs, $E(\hat{u}^2_{0,R+\tau} - \hat{u}^2_{1,R+\tau})$, may be zero, thus motivating a distinct null hypothesis of equal finite-sample predictive ability.

In our applications testing the null hypothesis of equal finite-sample predictive ability, we treat the MSE-t test and the other equal MSE test described below as one-sided to the right. We do so because, under the forecasting principle of parsimony, most practitioners and researchers seem to be inclined toward smaller models; that is, they are inclined to use a smaller model rather than a larger unless the larger model is significantly more accurate than the smaller, implying a preference for one-sided testing.[4]

[2] For simplicity, we suppress the triangular array notation implied by our use of local-to-zero asymptotics. The triangular array structure of the data implies underlying notation of a predictand $y_{R,t+\tau}$, the predictors $x_{R,1,t}$ and the error term $u_{R,t+\tau}$. We omit the R subscript for readability unless necessary to convey concepts.

[3] The coefficient vector β_1 does not vary with the forecast horizon τ because, in our analysis, τ is treated as fixed.

[4] However, there exist alternatives for which the left tail might be relevant. For example, the MSE-t and MSE-F statistics can diverge to negative infinity if $x_{w,t}$ contains no predictive content late in the sample but does so early in the sample. See Clark and McCracken (2005b) for a discussion. In addition, the null of equal finite-sample predictive ability can be rejected in the lower tail if $x_{w,t}$ has very weak predictive content in finite samples. Our bootstrap procedure can also be used to estimate critical values associated with the lower tail of the distribution. For brevity we focus attention on the upper tail of the distribution because we feel that is the alternative of greatest importance among practitioners.

While West (1996) proves directly that the MSE-t statistic can be asymptotically standard normal when applied to non-nested forecasts, this is not the case when the models are nested. In particular, the results in West (1996) require that under the null, the population-level long run variance of $\hat{L}_{t+\tau}$ is positive. This requirement is violated with nested models regardless of the presence of weak predictors. Intuitively, with nested models (and for the moment ignoring the weak predictors), the null hypothesis that the restrictions imposed in the benchmark model are true implies the population errors of the competing forecasting models are exactly the same. As a result, in population $L_{t+\tau} = 0$ for all t, which makes the corresponding variance also equal to 0. Because the sample analogues (for example, $\bar{L}$ and its variance) converge to zero at the same rate, the test statistics have non-degenerate null distributions, but they are non-standard.

Motivated by (i) the degeneracy of the long-run variance of $L_{t+\tau}$ and (ii) the functional form of the standard in-sample F-test, McCracken (2007) develops an out-of-sample F-type test of equal MSE, given by

$$\text{MSE-}F = (P-\tau+1) \times \frac{\text{MSE}_0 - \text{MSE}_1}{\text{MSE}_1}$$
$$= (P-\tau+1) \times \frac{\bar{L}}{\text{MSE}_1}. \qquad (3)$$

Like the MSE-t test, the limiting distribution of the MSE-F test is non-standard when the forecasts are nested under the null. Clark and McCracken (2005a) and McCracken (2007) show that, for τ-step ahead forecasts, the MSE-F statistic converges in distribution to functions of stochastic integrals of quadratics of Brownian motion, with limiting distributions that depend on the sample split (denoted λ_P below), the number of exclusion restrictions k_w, and the unknown nuisance parameter $S_{\tilde{h}\tilde{h}}$. While this continues to hold in the presence of weak predictors, the asymptotic distributions will depend not only upon the unknown coefficients associated with the weak predictors but also upon other unknown second moments of the data.

3.3. Asymptotic results

We use the following assumptions to derive the asymptotic distributions of the test statistics considered, for the case of weak predictors.

Assumption 1. The parameters of the forecasting models are estimated using OLS, yielding $\hat{\beta}_{i,t} = \arg\min_{\beta_i} t^{-1} \sum_{s=1}^{t-\tau} (y_{R,s+\tau} - x'_{R,i,s}\beta_i)^2$, $i = 0, 1$, for the restricted and unrestricted models, respectively.

Assumption 2. (a) $R^{-1} \sum_{t=1}^{[rR]} U_{R,t}U'_{R,t-j} \Rightarrow r\Omega_j$. (b) $\Omega_{11,j} = 0$ all $j \geq \tau$. (c) For some $q > 2$, $\sup_{R\geq 1, t\leq R+P} E|U_{R,t}|^{2q} < \infty$. (d) The zero mean array $U_{R,t} - EU_{R,t} = (h'_{R,1,t+\tau}, vec(x_{R,1,t}x'_{R,1,t} - Ex_{R,1,t}x'_{R,1,t})')'$ satisfies Theorem 3.2 of de Jong and Davidson (2000).

Assumption 3. $\lim_{P,R\to\infty} P/R = \lambda_P \in (0, \infty)$.

Assumption 4. (a) Let $K(x)$ be a continuous kernel such that for all real scalars x, $|K(x)| \leq 1$, $K(x) = K(-x)$ and $K(0) = 1$. (b) For some bandwidth M and constant $i \in (0, 0.5)$, $M = O(P^i)$. (c) The number of covariance terms $\bar{j}$, used to estimate the long-run covariance S_{LL} defined in Section 3.2, satisfies $\tau - 1 \leq \bar{j} < \infty$.

Assumption 2 imposes three types of conditions. First, in (a) and (c) we require that the observables, while not necessarily covariance stationary, are asymptotically mean square stationary with finite second moments. We do so in order to allow the observables to have marginal distributions that vary as the weak predictive ability strengthens along with the sample size but are 'well-behaved' enough that, for example, sample averages converge in probability to the appropriate population means. Second, in (b) we impose the restriction that the forecast errors form an MA($\tau - 1$) process and hence the model has sufficient lags to pick up all the autocorrelation in the errors other than that associated with the τ-step ahead nature of the forecasts. We do so in order to emphasize the role that weak predictors have on forecasting without also introducing other forms of model misspecification. Finally, in (d) we impose the high level assumption that, in particular, $h_{1,t+\tau}$ satisfies Theorem 3.2 of de Jong and Davidson (2000). By doing so we not only insure that certain weighted partial sums converge weakly to standard Brownian motion, but also allow ourselves to take advantage of various results pertaining to convergence in distribution to stochastic integrals.

Assumption 3's requirement on limiting sample sizes implies that the duration of forecasting is finite but non-trivial. This assumption, while standard in our previous work, differs importantly from that in Giacomini and White (2006). In their approach to predictive inference for nested models, they assume that a rolling window of fixed and *finite* width is used for estimation of the model parameters (hence $\lim_{P\to\infty} P/R = \infty$). While we allow rolling windows, our asymptotics assume that the window width is a non-trivial magnitude of the out-of-sample period and hence $\lim_{P,R\to\infty} P/R \in (0, \infty)$. This difference in the assumed window width, along with our assumption that the additional predictors in the nesting model are weak, is fundamentally what drives the difference in our results from theirs and, in particular, allows us to derive results that permit the use of the recursive scheme.

Finally, Assumption 4 is necessitated by the serial correlation in the multi-step (τ-step) forecast errors — errors from even well-specified models exhibit serial correlation, of an MA($\tau - 1$) form. Typically, researchers constructing a t-statistic utilizing the squares of these errors account for serial correlation of at least order $\tau - 1$ in forming the necessary standard error estimates. Meese and Rogoff (1988), Groen (1999), and Kilian and Taylor (2003), among other applications to forecasts from nested models, use kernel-based methods to estimate the relevant long-run covariance.[5] We therefore impose conditions sufficient to cover applied practices. Parts (a) and (b) are not particularly controversial. Part (c), however, imposes the restriction that since the orthogonality conditions used to identify the parameters form a moving average of finite order $\tau - 1$, this fact is taken into account (in the sense of assuming a finite bandwidth) when constructing the MSE-t statistic.

Under these assumptions, the asymptotic distributions will depend on the following functions of a $k_w \times 1$ vector standard Brownian motion, denoted $W(s)$, with (as defined above) $\delta = (0_{1\times k_0}, \beta'_w)'$: $\Gamma_1 = \int_1^{1+\lambda_P} s^{-1}W'(s)S_{\tilde{h}\tilde{h}}dW(s)$, $\Gamma_2 = \int_1^{1+\lambda_P} s^{-2}W'(s)S_{\tilde{h}\tilde{h}}W(s)ds$, $\Gamma_3 = \int_1^{1+\lambda_P} (\delta' B_1^{-1/2}\tilde{A}'/\sigma)S_{\tilde{h}\tilde{h}}^{1/2}dW(s)$, $\Gamma_4 = \int_1^{1+\lambda_P} \beta'_w F_1^{-1}\beta_w/\sigma^2 ds = \lambda_P\beta'_w F_1^{-1}\beta_w/\sigma^2$, $\Gamma_5 = \int_1^{1+\lambda_P} s^{-2}W'(s)S_{\tilde{h}\tilde{h}}^2W(s)ds$, $\Gamma_6 = \int_1^{1+\lambda_P} s^{-1}(\delta' B_1^{-1/2}\tilde{A}'/\sigma)S_{\tilde{h}\tilde{h}}^{3/2}W(s)ds$, and $\Gamma_7 = \lambda_P(\delta' B_1^{-1/2}\tilde{A}'/\sigma)S_{\tilde{h}\tilde{h}}(\tilde{A}B_1^{-1/2}\delta/\sigma)$.

The following two Theorems provide the asymptotic distributions of the MSE-F and MSE-t statistics in the presence of weak predictors.

Theorem 3.1. *Maintain Assumptions 1–3. MSE-F* $\to^d \{2\Gamma_1 - \Gamma_2\} + 2\{\Gamma_3\} + \{\Gamma_4\}$.

[5] For similar uses of kernel-based methods in analyses of non-nested forecasts, see, for example, Diebold and Mariano (1995) and West (1996).

Theorem 3.2. *Maintain Assumptions 1–4. MSE-t* $\rightarrow^d (\{\Gamma_1 - .5\Gamma_2\} + \{\Gamma_3\} + \{.5\Gamma_4\})/(\Gamma_5 + \Gamma_6 + \Gamma_7)^{.5}$.

Theorems 3.1 and 3.2 show that the limiting distributions of the MSE-*t* and MSE-*F* tests are neither normal nor chi-square when the forecasts are nested, regardless of the presence of weak predictors. Theorem 3.1 is very similar to Proposition 2 in Inoue and Kilian (2004) while Theorem 3.2 is unique. And again, the limiting distributions are free of nuisance parameters in only very special cases. In particular, the distributions here are free of nuisance parameters only if there are no weak predictors and if $S_{\tilde{h}\tilde{h}} = I$. If this is the case – if, for example, $\tau = 1$ and the forecast errors are conditionally homoskedastic – both representations simplify to those in McCracken (2007) and hence his critical values can be used for testing for equal population-level predictive ability. In the absence of weak predictors alone, the representation simplifies to that in Clark and McCracken (2005a) and hence the asymptotic distributions still depend upon $S_{\tilde{h}\tilde{h}}$. In this case, and in the most general case where weak predictors are present, we will use bootstrap methods to estimate asymptotic critical values. Before describing our bootstrap approach, however, it will be helpful to clarify the null hypothesis of interest.

3.4. A null hypothesis with weak predictors

The non-centrality terms, especially those associated with the asymptotic distribution of the MSE-*F* statistic (Γ_4), give some indication of the power that the test statistics have against deviations from the null hypothesis of equal population-level predictive ability $H_0{:}E(u^2_{0,t+\tau} - u^2_{1,t+\tau}) = 0$ for all t — for which it must be the case that $\beta_w = 0$. As noted earlier, it is in that sense that our analytical results are closely related to those in Inoue and Kilian (2004). Closer inspection, however, shows that the results provide opportunities for testing another form of the null hypothesis of equal predictive ability when weak predictors are present.

For example, under the assumptions made earlier in this section it is straightforward to show that the mean of the asymptotic distribution of the MSE-*F* statistic can be used to approximate the mean difference in the average out-of-sample predictive ability of the two models, as[6]:

$$E\sum_{t=R}^{R+P-\tau}(\hat{u}^2_{0,t+\tau} - \hat{u}^2_{1,t+\tau})$$

$$\approx \int_1^{1+\lambda_P} [-s^{-1}tr((-JB_0J' + B_1)V) + \beta'_w F_1^{-1}\beta_w]ds.$$

Intuitively, one might consider using these expressions as a means of characterizing when the two models have equal average finite-sample predictive ability over the out-of-sample period. For example, having set these two expressions to zero, integrating and solving for the marginal signal-to-noise ratio implies $\frac{\beta'_w F_1^{-1}\beta_w}{tr((-JB_0J'+B_1)V)}$ equals $\frac{\ln(1+\lambda_P)}{\lambda_P}$. This ratio simplifies further when $\tau = 1$ and the forecast errors are conditionally homoskedastic, in which case $tr((-JB_0J' + B_1)V) = \sigma^2 k_w$.

The marginal signal-to-noise ratio $\frac{\beta'_w F_1^{-1}\beta_w}{tr((-JB_0J'+B_1)V)}$ forms the basis of our new approach to testing for equal predictive ability. Rather than testing for equal population-level predictive ability $H_0{:}\,E(u^2_{0,t+\tau} - u^2_{1,t+\tau}) = 0$ for all t – for which it must be the case that $\beta_w = 0$ – we test for expected equal average out-of-sample predictive ability $H_0 : \lim_{P,R\rightarrow\infty} E(\sum_{t=R}^{R+P-\tau}(\hat{u}^2_{0,t+\tau} - \hat{u}^2_{1,t+\tau})) = 0$ — for which it is the case that $\beta'_w F_1^{-1}\beta_w = d$, where d equals $\frac{\ln(1+\lambda_P)}{\lambda_P}tr((-JB_0J' + B_1)V)$.[7] In the context of our simple example from Section 2.1 this is equivalent to testing the null that $\mu^2_w = \frac{\ln(1+\lambda_P)}{\lambda_P}\sigma^2$.[8]

While we believe the result is intuitive, it is not immediately clear how such a restriction on the regression parameters can be used to achieve asymptotically valid inference. If we look back at the asymptotic distribution of the MSE-*F* statistic, we see that in general it not only depends upon the unknown value of β_w, but also the asymptotic distribution is non-standard, thus requiring either extensive tables of critical values or simulation-based methods for constructing the critical values.

In the following section we develop a new bootstrap-based method for constructing asymptotically valid critical values that can be used to test the null of equal average finite-sample predictive ability. Conceptually, a viable alternative would be to use Monte Carlo methods to simulate the asymptotic distributions, after using the data to compute the moments that enter the distributions. However, most researchers and practitioners seem likely to find our bootstrap method easier to implement.

3.5. Bootstrap-based critical values with weak predictors

Our new, bootstrap-based method of approximating the asymptotically valid critical values for comparisons between nested models is different from that previously used in studies such as Kilian (1999) and Clark and McCracken (2005a). In those applications, an appropriately dimensioned VAR was initially estimated by OLS imposing the restriction that β_w was set to zero and the residuals saved for resampling. The recursive structure of the VAR was then used to generate a large number of artificial samples, each of which was used to construct one of the test statistics discussed above. The relevant sample percentile from this large collection of artificial statistics was then used as the critical value. Simulations show that this approach provides accurate inference for the null of equal population-level predictive ability not only for one-step ahead forecasts but also for longer horizons (in our direct multistep framework).

However, there are two reasons we should not expect this bootstrap approach to provide accurate inference in the presence of weak predictors. First, imposing the restriction that β_w is set to zero implies a null of equal population – not finite-sample – predictive ability. Second, by creating the artificial samples using the recursive structure of the VAR we are imposing the restriction that equal one-step ahead predictive ability implies equal predictive ability at longer horizons. Our present framework in no way imposes that restriction. We therefore take an entirely different approach to imposing the relevant null hypothesis and generating the artificial samples.

To test whether the two models have equal average predictive ability over the out-of-sample period, we need to determine whether $\beta'_w F_1^{-1}\beta_w$ equals $\frac{\ln(1+\lambda_P)}{\lambda_P}tr((-JB_0J' + B_1)V)$. While this

[6] By taking this approach we are using the fact that under our assumptions, notably the L^2-boundedness portion of Assumption 2, $\sum_{t=R}^{R+P-\tau}(\hat{u}^2_{0,t+\tau} - \hat{u}^2_{1,t+\tau})$ is uniformly integrable and hence the expectation of its limit is equal to the limit of its expectation.

[7] Note that this condition for equal forecast accuracy and, in turn, the null hypothesis, depend on population-level parameters or moments, not random variables.

[8] One could also derive a test for equal forecast accuracy at the end of the out-of-sample period. Using similar arguments, this hypothesis implies that $\beta'_w F_1^{-1}\beta_w = d$, where d equals $\frac{1}{1+\lambda_P}tr((-JB_0J'+B_1)V)$. Under this null hypothesis, our proposed bootstrap is valid so long as $\hat{d}$ (defined below) is modified appropriately.

restriction is infeasible due to the various unknown moments and parameters, it suggests a closely related, feasible restriction quite similar to that used in ridge regression. However, instead of imposing the restriction that $\beta_w'\beta_w = c$ for some finite constant – as one would in a ridge regression – we instead impose the restriction that $\delta' J_w F_1^{-1}(T) J_w' \delta$ equals $\frac{\ln(1+\widehat{\lambda}_P)}{\widehat{\lambda}_P} tr((-JB_0(T)J' + B_1(T))V(T))$, where the relevant unknowns are estimated using the obvious sample moments: $\widehat{\lambda}_P = P/R$, $B_i(T) = (T^{-1}\sum_{s=1}^{T-\tau} x_{i,s}x_{i,s}')^{-1}$ for $i = 0, 1$, $F_1(T) = J_w' B_1(T) J_w$, and $V(T) =$ an estimate of the long-run variance of $h_{1,t+\tau}$.[9] In addition, we estimate δ using the approximation $\widehat{\delta} = (0_{1\times k_0}, R^{1/2}\widetilde{\beta}_{w,T}')'$ where $\widetilde{\beta}_{w,T}$ denotes the restricted least squares estimator of the parameters associated with the weak predictors satisfying

$$\begin{aligned}\widetilde{\beta}_{1,T} &\equiv (\widetilde{\beta}_{0,T}', \widetilde{\beta}_{w,T}')' \\ &= \arg\min_{b_1} \sum_{s=1}^{R+P-\tau} (y_{s+\tau} - x_{1,s}' b_1)^2 \quad \text{s.t. } b_1' J_w F_1^{-1}(T) J_w' b_1 = \widehat{d}/R\end{aligned} \tag{4}$$

where $\widehat{d}$ equals $\frac{\ln(1+\widehat{\lambda}_P)}{\widehat{\lambda}_P} tr((-JB_0(T)J' + B_1(T))V(T))$. For a given sample size, this estimator is equivalent to a ridge regression if the weak predictors are orthonormal. More generally, though, it lies in the class of asymptotic shrinkage estimators discussed in Hansen (2008). While this estimator falls within the shrinkage class, it bears emphasizing that the key to the estimator is the imposition of restrictions on the coefficients of the weak regressors such that, in the finite sample, the competing forecasting models are equally accurate (in expectation). Our estimator provides a convenient method for imposing the needed restrictions on the larger forecasting model. With simple models like the one used as an example in Section 2.1 or our second Monte Carlo DGP, the general conditions needed to impose equal accuracy given in Section 3.3 simplify considerably. In this case, the coefficient values needed to make the competing forecasting models equally accurate can be easily and directly computed (see, for example, the discussion in Sections 2.1 and 3.1); the solution to the constrained estimation problem of Eq. (4) delivers exactly the same result.

Note that this approach to imposing the null hypothesis is consistent with the direct multi-step forecasting approach we assume is used to construct the forecasts and hence the restriction can vary with the forecast horizon τ. This approach therefore precludes using a VAR and its recursive structure to generate the artificial samples. Instead we use a variant of the wild fixed regressor bootstrap developed in Goncalves and Kilian (2007) that accounts for the direct multi-step nature of the forecasts. Specifically, in our framework the x's are held fixed across the artificial samples and the dependent variable is generated using the direct multi-step equation $y_{s+\tau} = x_{1,s}'\widetilde{\beta}_{1,T} + \widehat{v}_{s+\tau}^*$, $s = 1, \ldots, T$, for a suitably chosen artificial error term $\widehat{v}_{s+\tau}^*$ designed to capture both the presence of conditional heteroskedasticity and an assumed $MA(\tau - 1)$ serial correlation structure in the τ-step ahead forecasts. Specifically, we construct the artificial samples and bootstrap critical values using the following algorithm.[10]

1. (a) Estimate the unrestricted model by OLS without imposing the restriction in (4) and save the residuals $\widehat{v}_{1,s+\tau}$, $s = 1, \ldots, T-\tau$. (b) Set $\widehat{d}$ to $\frac{\ln(1+\widehat{\lambda}_P)}{\widehat{\lambda}_P} tr((-JB_0(T)J' + B_1(T))V(T))$. Estimate the unrestricted model using the weighted ridge regression from Eq. (4) above and save the fitted values $x_{1,s}'\widetilde{\beta}_{1,T}$ $s = 1, \ldots, T$. Note that the resulting parameter estimate will vary with the forecast horizon.
2. Using NLLS, estimate an $MA(\tau - 1)$ model for the OLS residuals $\widehat{v}_{1,s+\tau}$ (from the unrestricted model without the restriction in (4)) such that $v_{1,s+\tau} = \varepsilon_{1,s+\tau} + \theta_1\varepsilon_{1,s+\tau-1} + \cdots + \theta_{\tau-1}\varepsilon_{1,s+1}$. Let $\eta_{s+\tau}$, $s = 1, \ldots, T$, denote an *i.i.d* $N(0, 1)$ sequence of simulated random variables.[11] Define $\widehat{v}_{1,s+\tau}^* = (\eta_{s+\tau}\widehat{\varepsilon}_{1,s+\tau} + \widehat{\theta}_1\eta_{s-1+\tau}\widehat{\varepsilon}_{1,s+\tau-1} + \cdots + \widehat{\theta}_{\tau-1}\eta_{s+1}\widehat{\varepsilon}_{1,s+1})$, $s = 1, \ldots, T$. Form artificial samples of $y_{s+\tau}^*$ using the fixed regressor structure, $y_{s+\tau}^* = x_{1,s}'\widetilde{\beta}_{1,T} + \widehat{v}_{1,s+\tau}^*$.
3. Using the artificial data, construct forecasts and an estimate of the test statistics (MSE-F, MSE-t) as if this were the original data.
4. Repeat steps 2 and 3 a large number of times: $j = 1, \ldots, N$.
5. Reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100 - \alpha)$%-ile of the empirical distribution of the simulated test statistics.

By using the weighted ridge regression to estimate the model parameters we are able, in large samples, to impose the restriction that the implied estimates ($R^{1/2}\widetilde{\beta}_{w,T}$) of the local-to-zero parameters β_w satisfy our approximation to the null hypothesis. This is despite the fact that the estimates of β_w are not consistent. While this estimator, along with the fixed regressor structure of the bootstrap, imposes the null hypothesis upon the artificial samples, it is not necessarily the case that the bootstrap is asymptotically valid in the sense that the estimated critical values are consistent for their population values. To see how this might happen, note that the asymptotic distributions from Theorem 3.1 depend explicitly upon the local-to-zero parameters β_w through the terms Γ_3 and Γ_4. In the case of Γ_4, this is not an issue because the null hypothesis imposes a restriction on the value of this term that does not depend upon β_w explicitly, just an appropriately chosen weighted quadratic that is known under the null. Γ_3 is a different story. This term is asymptotically normal with a zero mean and variance $\lambda_P\beta_w' V\beta_w$ that, in general, need not have any relationship to the restriction $\beta_w' F_1^{-1}\beta_w = d$ implied by the null hypothesis. Hence, in general, the asymptotic distribution is an explicit function of the value of β_w, implying that the null hypothesis itself does not imply a unique asymptotic distribution for either the MSE-F or MSE-t statistics.

Even so, we can show that the bootstrap is asymptotically valid in two empirically relevant special cases. In both cases the bootstrap works despite the fact that we cannot consistently estimate β_w. The trick is to note that (i) while $R^{1/2}\widetilde{\beta}_{w,T}$ is not a consistent estimate of β_w, the first stage of the bootstrap insures that its probability limit $\tilde{\beta}_w$ is on the sphere defined by $\beta_w' F_1^{-1}\beta_w = d$ and (ii) in the two special cases discussed below, the null asymptotic distribution is invariant to the actual value of β_w so long as the relationship $\beta_w' F_1^{-1}\beta_w = d$ holds. To prove the result, however, we require a modest strengthening of the moment conditions on the model residuals.

Assumption 2′. (a) $R^{-1}\sum_{t=1}^{[rR]} U_{R,t}U_{R,t-j}' \Rightarrow r\Omega_j$. (b) $\mathrm{E}(\varepsilon_{1,s+\tau}| \varepsilon_{1,s+\tau-j}, x_{R,1,s-j}\ j \geq 0) = 0$. (c) Let $\gamma_R = (\beta_{1,R}', \theta_1, \ldots, \theta_{\tau-1})'$, $\widehat{\gamma}_R = (\hat{\beta}_{1,T}', \widehat{\theta}_1, \ldots, \widehat{\theta}_{\tau-1})'$, and define the function $\widehat{\varepsilon}_{1,s+\tau} = \widehat{\varepsilon}_{1,s+\tau}(\widehat{\gamma}_R)$ such that $\widehat{\varepsilon}_{1,s+\tau}(\gamma_R) = \varepsilon_{1,s+\tau}$. In an open neighborhood N_R around γ_R, there exists a finite constant c such that $\sup_{1\leq s\leq R, R\geq 1} \| \sup_{\gamma\in N_R}(\widehat{\varepsilon}_{1,s+\tau}(\gamma), \nabla\widehat{\varepsilon}_{1,s+\tau}'(\gamma), x_{R,1,s})'\|_4 \leq c$. (d) The zero mean array $U_{R,t} - EU_{R,t} = (h_{R,1,t+\tau}', vec(x_{R,1,t}x_{R,1,t}' - Ex_{R,1,t}x_{R,1,t}')')'$ satisfies Theorem 3.2 of de Jong and Davidson (2000).

[9] In our Monte Carlo simulations and empirical work we use a Newey–West kernel with bandwidth 0 for horizon = 1 and bandwidth 1.5*horizon otherwise.

[10] Our approach to generating artificial samples of multi-step forecast errors builds on a sampling approach proposed in Hansen (1996).

[11] The assumption of Normality is not necessary for our results. Instead it is sufficient if η_{s+t} denotes an *i.i.d.* zero-mean process with unit variance and is uniformly L^r-bounded for some $r > 2$.

Assumption 2′ differs from Assumption 2 in two ways. First, in (b) it emphasizes the point that the forecast errors, and by implication $h_{1,t+\tau}$, form an $MA(\tau-1)$ process. Second, in (c) it bounds the second moments not only of $h_{1,t+\tau} = (\varepsilon_{1,s+\tau} + \theta_1\varepsilon_{1,s+\tau-1} + \cdots + \theta_{\tau-1}\varepsilon_{1,s+1})x_{1,s}$ (as in Assumption 2) but also the functions $\widehat{\varepsilon}_{1,s+\tau}(\gamma)x_{1,s}$ and $\nabla\widehat{\varepsilon}_{1,s+\tau}(\gamma)x_{1,s}$ for all γ in an open neighborhood of γ_R. These assumptions are primarily used to show that the bootstrap-based artificial samples, which are a function of the estimated errors $\widehat{\varepsilon}_{1,s+\tau}$, adequately replicate the time series properties of the original data in large samples. Specifically we must insure that the bootstrap analog of $h_{1,s+\tau}$ is not only zero mean but has the same long-run variance V. Such an assumption is not needed for our earlier results since the model forecast errors $\widehat{u}_{i,s+\tau}, i = 0, 1$, are linear functions of $\hat{\beta}_i$ and Assumption 2 already imposes moment conditions on $\widehat{u}_{1,s+\tau}$ via moment conditions on $h_{1,s+\tau}$.

In the following let MSE-F^* and MSE-t^* denote statistics generated using the artificial samples from our bootstrap and let $=^{d^*}$ and $\to^{d^*}$ denote equality and convergence in distribution (more specifically, convergence occurs in a set with probability limiting to 1) with respect to the bootstrap-induced probability measure P^*. Similarly let $\Gamma_i^*, i = 1, \ldots, 7$, denote random variables generated using the artificial samples satisfying $\Gamma_i^* =^{d^*} \Gamma_i, i = 1, \ldots, 7$, for Γ_i defined in the discussion preceding the assumptions.

Theorem 3.3. *Let* $\beta_w' F_1^{-1}\beta_w = d$ *and assume either* (i) $\tau = 1$ *and the forecast errors from the unrestricted model are conditionally homoskedastic, or* (ii) $\dim(\beta_w) = 1$. (a) *Given Assumptions* 1, 2′ *and* 3, *MSE-F*$^* \to^{d^*} \{2\Gamma_1^* - \Gamma_2^*\} + 2\{\Gamma_3^*\} + \{\Gamma_4^*\}$. (b) *Given Assumptions* 1, 2′, 3, *and* 4, *MSE-t*$^* \to^{d^*} (\{2\Gamma_1^* - \Gamma_2^*\} + 2\{\Gamma_3^*\} + \{\Gamma_4^*\})/(\Gamma_5^* + \Gamma_6^* + \Gamma_7^*)^{.5}$.

In Theorem 3.3 we show that our fixed-regressor bootstrap provides an asymptotically valid method of estimating the critical values associated with the null of equal average finite sample forecast accuracy. The result, however, is applicable in only two special cases. In the first, we require that the forecast errors be one-step ahead and conditionally homoskedastic. In the second, we allow serial correlation and conditional heteroskedasticity but require that β_w is scalar. While neither case covers the broadest situation in which β_w is not scalar and the forecast errors exhibit either serial correlation or conditional heteroskedasticity, these two special cases cover a wide range of empirically relevant applications. Kilian (1999) argues that conditional homoskedasticity is a reasonable assumption for one-step ahead forecasts of quarterly macroeconomic variables. Moreover, in many applications in which a nested model comparison is made (examples include, among others Chen et al., 2010; Goyal and Welch, 2008; Stock and Watson, 2003), the unrestricted forecasts are made by simply adding one lag of a single predictor to the baseline restricted model.

By itself, Theorem 3.3 is insufficient for recommending the use of the bootstrap: it does not tell us whether the proposed bootstrap is adequate for constructing asymptotically valid critical values under the alternative that the unrestricted model forecasts more accurately than the restricted model. Unfortunately, there are any number of ways to model the case in which $\beta_w' F_1^{-1}\beta_w > d$. For example, rather than modeling the weak predictive ability as $R^{-1/2}\beta_w$ with $\beta_w' F_1^{-1}\beta_w = d$, one could model the predictive content as $R^{-a}C\beta_w$ for constants $C < \infty$ and $a \in (0, 1/2]$ satisfying $\beta_w' F_1^{-1}\beta_w > d$. While mathematically elegant, this alternative is not particularly tractable. Accordingly, we instead focus on a more tractable alternative in which the predictive ability is no longer weak; the predictive content is driven by a simple, constant non-local-to-zero coefficient vector β_w (obtained by setting $a = 0$ in the general alternative just described).[12] In this case, forecasts from the unrestricted model are more accurate than forecasts from the restricted model, and $J_w'\hat{\beta}_{1,T}$ is a consistent estimator of $\beta_w \neq 0$. Under this alternative hypothesis, we address the validity of the bootstrap in the following Theorem.

Theorem 3.4. *Let* $J_w'\hat{\beta}_{1,T} \to^p \beta_w \neq 0$ *and assume either* (i) $\tau = 1$ *and the forecast errors from the unrestricted model are conditionally homoskedastic, or* (ii) $\dim(\beta_w) = 1$. (a) *Given Assumptions* 1, 2′ *and* 3, *MSE-F*$^* \to^{d^*} \{2\Gamma_1^* - \Gamma_2^*\} + 2\{\Gamma_3^*\} + \{\Gamma_4^*\}$. (b) *Given Assumptions* 1, 2′, 3 *and* 4, *MSE-t*$^* \to^{d^*} (\{2\Gamma_1^* - \Gamma_2^*\} + 2\{\Gamma_3^*\} + \{\Gamma_4^*\})/(\Gamma_5^* + \Gamma_6^* + \Gamma_7^*)^{.5}$.

In Theorem 3.4 we see that indeed, the bootstrap-based test is consistent for testing the null hypothesis of equal finite sample predictive accuracy (that $\beta_w' F_1^{-1}\beta_w = d$) against the alternative that the unrestricted model is more accurate (that $J_w'\hat{\beta}_{1,T} \to^p \beta_w \neq 0$). This follows since under this alternative, the data-based statistics MSE-F and MSE-t each diverge to $+\infty$ while the bootstrap-based statistics MSE-F^* and MSE-t^* each retain the same asymptotic distribution as they did under the null.

As we will show in Section 4, our fixed regressor bootstrap provides reasonably sized tests in our Monte Carlo simulations, outperforming other bootstrap-based methods for estimating the asymptotically valid critical values necessary to test the null of equal average finite sample predictive ability.

3.6. Differences under the rolling scheme

The results presented above for the recursive estimation and forecasting scheme also apply under a rolling scheme, under which the number of observations used for estimation is held constant as we proceed forward across forecast origins. This subsection lists the changes that apply under a rolling scheme.

First, under the rolling scheme, the parameter estimates and associated moments are defined as $\hat{\beta}_{i,t} = \arg\min_{\beta_i} R^{-1}\sum_{s=t-\tau-R+1}^{t-\tau}(y_{s+\tau} - x_{i,s}'\beta_i)^2$ for models $i = 0, 1$.

Second, some of the functions entering the asymptotic distributions are slightly different under the rolling scheme: $\Gamma_1 = \int_1^{1+\lambda_P}(W(s) - W(s-1))'S_{\tilde{h}\tilde{h}}dW(s)$, $\Gamma_2 = \int_1^{1+\lambda_P}(W(s) - W(s-1))'S_{\tilde{h}\tilde{h}}(W(s) - W(s-1))ds$, $\Gamma_5 = \int_1^{1+\lambda_P}(W(s) - W(s-1))' S_{\tilde{h}\tilde{h}}^2(W(s) - W(s-1))ds$, and $\Gamma_6 = \int_1^{1+\lambda_P} s^{-1}(\delta' B_1^{-1/2}\tilde{A}'/\sigma) S_{\tilde{h}\tilde{h}}^{3/2}(W(s) - W(s-1))ds$.

Third, under the rolling scheme, the approximation of the mean difference in average forecast accuracy is

$$E\sum_{t=R}^{R+P+\tau}(\hat{u}_{0,t+\tau}^2 - \hat{u}_{1,t+\tau}^2) \approx \int_1^{1+\lambda_P}[-tr((-JB_0J' + B_1)V) + \beta_w' F_1^{-1}\beta_w]ds.$$

Solving for the marginal signal-to-noise ratio that makes the models equally accurate implies $\frac{\beta_w' F_1^{-1}\beta_w}{tr((-JB_0J'+B_1)V)} = 1$. In turn, equal average out-of-sample predictive ability implies a condition $\beta_w' F_1^{-1}\beta_w = d$, where $d = tr((-JB_0J' + B_1)V)$.

Finally, under the rolling scheme, the first step of the bootstrap consists of estimating the parameter vector β_1 associated with the unrestricted model (as detailed in Eq. (4)) subject to the restriction $\widehat{d} = tr((-JB_0(T)J' + B_1(T))V(T))$.

[12] When $a = 0$ the test is consistent. For $a = 1/2$ the test will have power so long as $\beta_w' F_1^{-1}\beta_w > d$, but will not be consistent.

4. Monte Carlo evidence

We use simulations of DGPs that are in most cases based on common macroeconomic applications to evaluate the finite sample properties of the above approaches to testing for equal forecast accuracy. In these simulations, the benchmark forecasting model is a univariate model of the predictand y; the alternative models add lags of various other variables of interest. We also consider results from a simpler DGP that takes the form used as an example in Section 2.1. The null hypothesis is that the forecast from the alternative model is no more accurate than the benchmark forecast, in the sense that the additional variables in the alternative model have non-zero coefficients, but the coefficients are small enough that the benchmark and alternative models are expected to be equally accurate over the forecast sample. We focus our presentation on recursive forecasts, but include some results for rolling forecasts. We report empirical rejection rates using a nominal size of 10%. Size results using nominal sizes of 5% and 1% are qualitatively the same and available upon request.

While our focus is on evaluating the size and power of MSE-*F* and MSE-*t* tests based on the asymptotic distributions and bootstrap developed in the last section, we also consider two other approaches to inference — that is, sources of critical values and tests. First, we include results for the MSE-*t* test compared against standard normal critical values. Second, we provide results based on a non-parametric bootstrap patterned on White's (2000) method: we create bootstrap samples of forecast errors by sampling (with replacement) from the time series of sample forecast errors, and construct test statistics for each sample draw. However, as noted above and in White (2000), this procedure is not, in general, asymptotically valid when applied to nested models. We include the method in part for its computational simplicity and in part to examine the potential pitfalls of using the approach.

In our non-parametric bootstrap implementation, we follow the approach of White (2000) in using the stationary bootstrap of Politis and Romano (1994) and centering the bootstrap distributions around the sample values of the test statistics. The stationary bootstrap is parameterized to make the average block length equal to twice the forecast horizon. As to centering of test statistics, under the non-parametric approach, the relevant null hypothesis is that the MSE difference (benchmark MSE less alternative model MSE) is at most 0, and the MSE ratio (benchmark MSE/alternative model MSE) is at most 1. Following White (2000), each bootstrap draw of a given test statistic is re-centered around the corresponding sample test statistic. Bootstrapped critical values are computed as percentiles of the resulting distributions of re-centered test statistics.

4.1. Monte Carlo design

For all DGPs, we generate data using independent draws of innovations from the normal distribution and (as appropriate) the autoregressive structure of the DGP.[13] We consider forecast horizons of one and four steps. With quarterly data in mind, we also consider a range of sample sizes (R, P), reflecting those commonly available in practice: 40, 80; 40, 120; 80, 40; 80, 80; 80, 120; 120, 40; and 120, 80.

The DGPs are in most cases generally based on empirical relationships among US inflation and a range of predictors, estimated with 1968–2008 data.[14] We focus on results for five DGPs, all of which satisfy the assumptions necessary to prove the asymptotic validity of our proposed bootstrap. But to assess the reliability of our proposed approach in more general settings, we also include some results for two additional DGPs that feature conditional heteroskedasticity or serial correlation in the forecast error, with model 1 having three more variables than model 0. We also provide results for a DGP in which the forecasting models are misspecified. In all cases, our reported results are based on 5000 Monte Carlo draws and 499 bootstrap replications.

4.1.1. DGPs

DGP 1 takes the very simple form of the DGP used in Section 2 to provide intuition for our basic approach and theory:

$$y_{t+1} = \mu + u_{t+1}, \qquad u_{t+1} \sim iid\, N(0, 1). \tag{5}$$

In the DGP 1 experiments, which focus on a forecast horizon of 1 step, the null forecast is a no-change forecast, and the alternative (unrestricted) forecasting model takes the form of the DGP equation for y_{t+1}:

$$\text{null: } y_{t+1} = u_{0,t+1} \tag{6}$$

$$\text{alternative: } y_{t+1} = \beta_0 + u_{1,t+1}. \tag{7}$$

Note that Section 2.1 spells out the simple version of the bootstrap used for inference with this DGP.

In size experiments, the intercept μ is set to a value that makes the models equally accurate (in expectation) on average over the forecast sample. Based on the theoretical solution given in Section 2.1 and the DGP parameterization $\mathrm{var}(u_{t+1}) = 1$, in each recursive forecasting experiment we set the DGP value as $\mu = \sqrt{\frac{\log\left(1+\hat{\lambda}_P\right)}{\hat{\lambda}_P}}/\sqrt{R}$, where $\hat{\lambda}_P = P/R$ as implied by the P, R setting for the experiment. Note, however, that our bootstrap results do not exploit this population-level value of the coefficient; the bootstrap involves estimating the coefficient in each data set using the data available. In each rolling experiment, in the DGP we set $\mu = 1/\sqrt{R}$. In power experiments, the coefficient μ is set to 0.3, such that the alternative model is expected to be more accurate than the null.

DGP 2 also takes a simple form, but involves estimation of the restricted forecasting model:

$$\begin{aligned} y_{t+1} &= 1.0 + b_{11}x_{1,t} + u_{t+1} \\ x_{1,t+1} &= v_{1,t+1} \\ \mathrm{var}\begin{pmatrix} u_{t+1} \\ v_{1,t+1} \end{pmatrix} &= \begin{pmatrix} 1.0 & \\ 0.0 & 0.25 \end{pmatrix}. \end{aligned} \tag{8}$$

In the DGP 2 experiments, which focus on a forecast horizon of 1 step, the alternative (unrestricted) forecasting model takes the form of the DGP equation for y_{t+1}, while the null or benchmark (restricted) model includes just a constant:

$$\text{null: } y_{t+1} = \beta_0 + u_{0,t+1} \tag{9}$$

$$\text{alternative: } y_{t+1} = \beta_0 + \beta_1 x_{1,t} + u_{1,t+1}. \tag{10}$$

In size experiments, the coefficient b_{11} on $x_{1,t}$, which corresponds to the elements of our theoretical construct $\beta_w/\sqrt{R}$, is set to a value that makes the models equally accurate (in expectation) on average over the forecast sample. We determined the appropriate value on the basis of the population moments implied by the model and our asymptotic approximations given in Section 3.4. In the case of this simple DGP, under the recursive forecasting scheme (the solution is the same in the rolling case, except that the terms involving $\hat{\lambda}_P$ are not present), the general condition for equal accuracy detailed in Section 3.4 simplifies to yield the following solution for the DGP coefficient b_{11}:

$$b_{11} = \sqrt{\frac{1}{R}\frac{\log\left(1+\hat{\lambda}_P\right)}{\hat{\lambda}_P}\frac{\mathrm{var}(u_{t+1})}{\mathrm{var}(x_{1,t})}}.$$

[13] In our baseline size experiments, using innovations drawn from a t distribution with 5 degrees of freedom yields very similar results that are available upon request.

[14] While the empirical models underlying the DGP parameterizations and the forecasting models used in our Monte Carlos include constants, for simplicity we leave constants out of most of our data-generating processes. Including non-zero intercepts in the DGPs yields results exactly the same as those reported, because all of the forecasting models include intercepts.

For example, given the variance parameter values indicated in Eq. (8), recursive forecasting, and R and P both equal to 80, this value is 0.1862.[15] Note, however, that our bootstrap results do not exploit this population-level value of the coefficient (the same is true for experiments with all of the DGPs described below); the bootstrap involves estimating the coefficient in each data set using the data available and the algorithm detailed in Section 3.5. In power experiments, the coefficient is set to 0.5, such that the alternative model is expected to be more accurate than the null.

DGP 3 is based on the empirical relationship between the change in core PCE inflation (y_t) and the Chicago Fed's index of the business cycle ($x_{1,t}$, the CFNAI):

$$\begin{aligned} y_{t+1} &= -0.4y_t - 0.1y_{t-1} + b_{11}x_{1,t} + u_{t+1} \\ x_{1,t+1} &= 0.7x_{1,t} + v_{1,t+1} \\ \text{var}\begin{pmatrix} u_{t+1} \\ v_{1,t+1} \end{pmatrix} &= \begin{pmatrix} 0.8 & \\ 0.0 & 0.3 \end{pmatrix}. \end{aligned} \tag{11}$$

In the DGP 3 experiments, which also focus on a forecast horizon of 1 step, the alternative (unrestricted) forecasting model takes the AR(2) form of the DGP equation for y_{t+1} (with constant added); the null or benchmark (restricted) model drops $x_{1,t}$:

$$\text{null:}\ y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + u_{0,t+1} \tag{12}$$

$$\text{alternative:}\ y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 x_{1,t} + u_{1,t+1}. \tag{13}$$

In size experiments, the coefficient b_{11} on $x_{1,t}$ is set to a value that makes the forecasting models equally accurate (in expectation) on average over the forecast sample. As described above, we used our asymptotic approximations to determine the appropriate value. For example, with recursive forecasts and R and P both equal to 80, this value is 0.1086, about 1/2 of the empirical estimate. In power experiments, the coefficient is set to 0.3, such that the alternative model is expected to be more accurate than the null.

DGP 4 is based on the empirical relationship of the change in core PCE inflation (y_t) to the CFNAI ($x_{1,t}$), PCE food price inflation less core inflation ($x_{2,t}$), and import price inflation less core inflation ($x_{3,t}$). To simplify the lag structure necessary for reasonable forecasting models, the inflation rates used in forming variables $x_{2,t}$ and $x_{3,t}$ are computed as two-quarter averages. Based on these data, DGP 4 takes the form

$$\begin{aligned} y_{t+1} &= -0.4y_t - 0.1y_{t-1} + b_{11}x_{1,t} + b_{21}x_{2,t} + b_{31}x_{3,t} + u_{t+1} \\ x_{1,t+1} &= 0.7x_{1,t} + v_{1,t+1} \\ x_{2,t+1} &= 0.9x_{2,t} - 0.2x_{2,t-1} + v_{2,t+1} \\ x_{3,t+1} &= 1.1x_{3,t} - 0.3x_{3,t-1} + v_{3,t+1} \\ \text{var}\begin{pmatrix} u_t \\ v_{1,t+1} \\ v_{2,t+1} \\ v_{3,t+1} \end{pmatrix} &= \begin{pmatrix} 0.8 & & & \\ 0.0 & 0.3 & & \\ -0.1 & 0.0 & 2.2 & \\ 0.5 & 0.1 & 0.8 & 9.0 \end{pmatrix}. \end{aligned} \tag{14}$$

In DGP 4 experiments, which also focus on a forecast horizon of 1 step, the null (restricted) and alternative (unrestricted) forecasting models take the following forms, respectively:

$$y_{t+1} = \beta_0 + \beta_1 y_t + \beta_1 y_{t-1} + u_{0,t+1} \tag{15}$$

$$\begin{aligned} y_{t+1} = \beta_0 &+ \beta_1 y_t + \beta_1 y_{t-1} + \beta_3 x_{1,t} \\ &+ \beta_4 x_{2,t} + \beta_5 x_{3,t} + u_{1,t+1}. \end{aligned} \tag{16}$$

In power experiments, the b_{ij} coefficients are set at $b_{11} = 0.3$, $b_w = 0.1$, and $b_{13} = .015$ (roughly their empirical values). With these values, the alternative model is expected to be more accurate than the null. In size experiments, these values of the b_{ij} coefficients are multiplied by a constant less than one, such that, in population, the null and alternative models are expected to be equally accurate, on average, over the forecast sample (we computed the scaling factor using the population moments implied by the model and Section 3.4's asymptotic approximations). For example, with R and P at 80, this multiplying constant is 0.4118.

DGP 5, which incorporates a forecast horizon of four periods, is also based on the empirical relationship between the change in core PCE inflation (y_t) and the Chicago Fed's index of the business cycle. In this case, though, the model is based on empirical estimates using (changes in) the four-quarter rate of inflation[16]:

$$\begin{aligned} y_{t+4} &= b_{11}x_{1,t} + e_{t+4} \\ e_{t+4} &= u_{t+4} + .95u_{t+3} + .9u_{t+2} + .8u_{t+1} \\ x_{1,t+4} &= 0.7x_{1,t+3} + v_{1,t+4} \\ \text{var}\begin{pmatrix} u_{t+4} \\ v_{1,t+4} \end{pmatrix} &= \begin{pmatrix} 0.2 & \\ 0.0 & 0.3 \end{pmatrix}. \end{aligned} \tag{17}$$

In these experiments, the forecasting models are:

$$\text{null:}\ y_{t+4} = \beta_0 + u_{0,t+4} \tag{18}$$

$$\text{alternative:}\ y_{t+4} = \beta_0 + \beta_1 x_{1,t} + u_{1,t+4}. \tag{19}$$

Again, in size experiments, the coefficient b_{11} on $x_{1,t}$ is set to a value that makes the models equally accurate (in expectation) on average over the forecast sample (on the basis of the model-implied population moments and Section 3.4's asymptotic approximations). For example, with recursive forecasts and R and P both equal to 80, this value is 0.1634. In power simulations, the coefficient is set to its empirical value of 0.4, such that the alternative model is expected to be more accurate than the null.

DGP 6 takes the same form as DGP 4 (Eq. (14)), except that it incorporates multiplicative conditional heteroskedasticity in the error term of the equation for y:

$$u_{t+1} = \frac{|x_{1,t}|}{\sigma_{x,1}} e_{t+1},$$

where $\sigma_{x,1}$ denotes the standard deviation of $x_{1,t}$ implied by the DGP, and the variance–covariance matrix of e_{t+1} and the innovations $v_{i,t}$, $i = 1, \ldots, 3$, is set to match the covariance matrix given in Eq. (14). The forecasting models in DGP 6 experiments are the same as in DGP 4 (Eqs. (15) and (16)).

With DGP 6, in the interest of brevity we consider only size experiments. Because conditional heteroskedasticity makes it very difficult to compute the population moments needed to determine the coefficient settings that imply equal accuracy, we rely instead on preliminary rounds of Monte Carlo simulations to set the b_{ij} coefficients to yield equal accuracy of the competing models. We begin with (empirically-based) coefficients of $b_{11} = 0.3$, $b_{21} = 0.1$, $b_{31} = 0.015$. For each R, P combination, we use our asymptotic theory assuming conditional homoskedasticity to determine a preliminary re-scaling of the coefficient vector to yield equal accuracy. For each R, P combination, we then conduct three sets of Monte Carlo experiments (with a large number of draws), searching across grids of the re-scaling of the coefficient vector to select a scaling of the set of b_{ij} coefficients that minimizes the average

[15] This value is obtained by plugging $R = 80$, $\hat{\lambda}_P = 1$, $\text{var}(u_{t+1}) = 1$, and $\text{var}(x_{1,t}) = 0.25$ into the formula above.

[16] Specifically, in the empirical estimates underlying the DGP settings, we defined $y_{t+4} = 100\ln(p_{t+4}/p_t) - 100\ln(p_t/p_{t-4})$, where p denotes the core PCE price index.

(across Monte Carlo draws) difference in MSEs from the competing forecasting models.[17]

DGP 7 extends DGP 5 to include more predictands for y:

$$
\begin{aligned}
&y_{t+4} = b_{11}x_{1,t} + b_{21}x_{2,t} + b_{31}x_{3,t} + e_{t+4}\\
&e_{t+4} = u_{t+4} + .95u_{t+3} + .9u_{t+2} + .8u_{t+1}\\
&x_{1,t+4} = 0.7x_{1,t+3} + v_{1,t+4}\\
&x_{2,t+4} = 0.8x_{2,t+3} + v_{2,t+4}\\
&x_{3,t+4} = 0.8x_{3,t+3} + v_{3,t+4}\\
&\mathrm{var}\begin{pmatrix} u_{t+4}\\ v_{1,t+4}\\ v_{2,t+4}\\ v_{3,t+4}\end{pmatrix} = \begin{pmatrix} 0.2 & & & \\ -0.01 & 0.3 & & \\ 0.03 & 0.03 & 2.2 & \\ -0.2 & 0.02 & 0.8 & 9.0\end{pmatrix}.
\end{aligned}
$$

In the DGP 7 experiments, the forecasting models are:

$$\text{null: } y_{t+4} = \beta_0 + u_{0,t+4} \tag{20}$$

$$\text{alternative: } y_{t+4} = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_{1,t+4}. \tag{21}$$

In the interest of brevity, we consider only size results for DGP 7. For these experiments, we use our asymptotic theory to determine an initial scaling of the set of b_{ij} coefficients that implies equal accuracy in the finite sample given R and P, with the coefficients initially set (before scaling) to $b_{11} = 0.4$, $b_{21} = 0.2$, $b_{31} = 0.05$ (based roughly on empirical estimates). Using the approach described for DGP 6, we conduct three preliminary rounds of Monte Carlo simulations to refine the coefficient settings to make the competing forecasts equally accurate in the finite sample.

Finally, we also consider size experiments with a **DGP 8** in which the competing forecasting models under consideration are both misspecified. The data-generating portion of this DGP takes the same form as DGP 4:

$$
\begin{aligned}
&y_{t+1} = -0.4y_t - 0.1y_{t-1} + b_{11}x_{1,t} + b_{21}x_{2,t} + b_{31}x_{3,t} + u_{t+1}\\
&x_{1,t+1} = 0.7x_{1,t} + v_{1,t+1}\\
&x_{2,t+1} = 0.9x_{2,t} - 0.2x_{2,t-1} + v_{2,t+1}\\
&x_{3,t+1} = 1.1x_{3,t} - 0.3x_{3,t-1} + v_{3,t+1}\\
&\mathrm{var}\begin{pmatrix} u_{t}\\ v_{1,t+1}\\ v_{2,t+1}\\ v_{3,t+1}\end{pmatrix} = \begin{pmatrix} 0.8 & & & \\ 0.0 & 0.3 & & \\ -0.1 & 0.0 & 2.2 & \\ 0.5 & 0.1 & 0.8 & 9.0\end{pmatrix}.
\end{aligned}
\tag{22}
$$

To make the 1-step ahead forecasting models misspecified, we use the forecasting equations of DGP 3. These forecasting equations both exclude the $x_{2,t}$ and $x_{3,t}$ that are in the data-generating process for the predictand y_t:

$$\text{null: } y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + u_{0,t+1} \tag{23}$$

$$\text{alternative: } y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 x_{1,t} + u_{1,t+1}. \tag{24}$$

In setting the b_{ij} coefficients to yield equal accuracy of the competing, misspecified forecasting models, we fixed the coefficients on $x_{2,t-1}$ and $x_{3,t-1}$ at $b_{21} = 0.1$ and $b_{31} = 0.015$. To set the value of the coefficient b_{11} on $x_{1,t-1}$, we relied on preliminary rounds of Monte Carlo simulations. We conducted three sets of Monte Carlo experiments (with a large number of draws), searching across a grid of values for b_{11} and choosing the value that minimized the average (across Monte Carlo draws) difference in MSEs from the competing forecasting models. This value varied across experiments with different settings of R and P. For example, with R and P both equal to 80, this value is 0.1185.

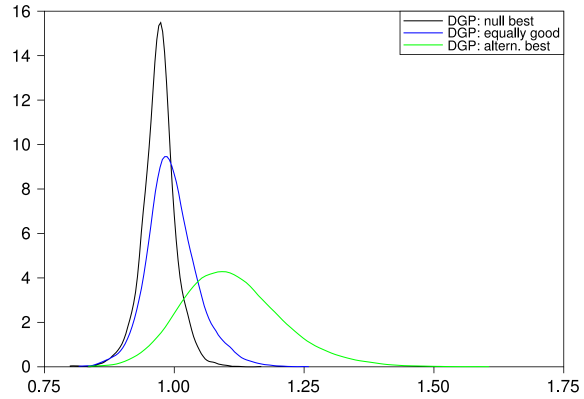

Fig. 1. Densities of MSE (null model)/MSE (alt. model), $R = 80$, $P = 80$, DGP 4 experiments.

4.2. Results

Our interest lies in identifying those testing approaches that yield reasonably accurate inferences on the forecast performance of models. At the outset, then, it may be useful to broadly summarize the forecast performance of competing models under various alternatives. Accordingly, Fig. 1 shows estimated densities of the MSE ratio statistic (the ratio of the null model's MSE to the alternative model's MSE), based on experiments with DGP 4, using $R = P = 80$. We provide three densities, for the cases in which the b_{ij} coefficients of the DGP (14) are: (i) set to 0, such that the null model should be more accurate; (ii) set to non-zero values so as to make the null and alternative models (15) and (16) equally accurate over the forecast sample, according to our local-to-zero asymptotic results; and (iii) set at larger values, such that the alternative model is expected to be more accurate.

As the figure shows, for the DGP which implies the null model should be best, the MSE ratio distribution mostly lies below 1.0. For the DGP that implies the models can be expected to be equally accurate, the distribution is centered at about 1.0. Finally, for the DGP that implies the alternative model can be expected to be best, the distribution mostly lies above 1.0. Our proposed fixed regressor bootstrap is intended to estimate a null distribution like that shown for the equally good models DGP. In most of our results, the null will be rejected when the sample MSE ratio lies in the right tail of the bootstrapped distribution.

4.2.1. Size results: recursive forecasts

Table 1 presents results for DGPs in which the b_{ij} coefficients on some x variables are non-zero but small enough that, under our asymptotic approximation, the null and alternative forecasting models are expected to be equally accurate over the sample considered. These size results show that, for testing the null of equal forecast accuracy, our proposed fixed regressor procedure is quite reliable, in the sense of yielding correctly-sized tests.

Tests based on the fixed regressor bootstrap generally have rejection rates of about 10% (the nominal size). For example, in the case of the MSE-F test applied to 1-step ahead forecasts, rejection rates range from 8.3% to 10.7%. Admittedly, rejection rates for 4-step ahead forecast tests are modestly higher, ranging from

[17] Specifically, we first consider 11 different experiments, each using 20,000 draws and a modestly different set of coefficient values obtained by scaling the baseline values, using a grid of scaling factors. We then pick the coefficient scaling that yields the lowest (in absolute value) average (across draws) difference in MSEs. We then repeat the 11-experiment exercise. Finally, we consider a third set of 21 experiments, with a more refined grid of coefficient scaling values and 200,000 draws. The coefficient scaling value that yields the smallest (absolute) difference in MSEs in this third set of experiments is then used to set the coefficients in the DGP simulated for the purpose of evaluating test properties.

Table 1
Monte Carlo results on size (nominal size = 10%).

DGP 1, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.058	0.053	0.078	0.070	0.059	0.077	0.070
MSE-*F*	Fixed regressor	0.097	0.096	0.091	0.106	0.102	0.087	0.093
MSE-*t*	Non-parametric	0.062	0.058	0.077	0.073	0.062	0.075	0.073
MSE-*t*	Fixed regressor	0.094	0.099	0.091	0.102	0.098	0.087	0.094
MSE-*t*	Normal	0.062	0.057	0.077	0.075	0.060	0.075	0.071
MSE-*t*, 2-sided	Normal	0.105	0.096	0.122	0.103	0.104	0.120	0.113

DGP 2, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.062	0.058	0.084	0.067	0.063	0.083	0.068
MSE-*F*	Fixed regressor	0.107	0.107	0.106	0.099	0.103	0.096	0.095
MSE-*t*	Non-parametric	0.070	0.068	0.101	0.073	0.069	0.095	0.078
MSE-*t*	Fixed regressor	0.089	0.093	0.096	0.088	0.094	0.086	0.085
MSE-*t*	Normal	0.066	0.064	0.094	0.069	0.067	0.084	0.070
MSE-*t*, 2-sided	Normal	0.097	0.105	0.113	0.106	0.100	0.109	0.106

DGP 3, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.054	0.048	0.080	0.062	0.057	0.083	0.070
MSE-*F*	Fixed regressor	0.101	0.096	0.101	0.102	0.096	0.099	0.103
MSE-*t*	Non-parametric	0.065	0.055	0.094	0.074	0.064	0.097	0.079
MSE-*t*	Fixed regressor	0.088	0.088	0.092	0.089	0.085	0.091	0.093
MSE-*t*	Normal	0.059	0.053	0.085	0.068	0.058	0.086	0.076
MSE-*t*, 2-sided	Normal	0.098	0.100	0.113	0.114	0.099	0.115	0.112

DGP 4, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.041	0.044	0.068	0.060	0.055	0.080	0.072
MSE-*F*	Fixed regressor	0.083	0.094	0.089	0.097	0.090	0.084	0.093
MSE-*t*	Non-parametric	0.055	0.050	0.092	0.075	0.064	0.100	0.084
MSE-*t*	Fixed regressor	0.077	0.087	0.086	0.089	0.082	0.088	0.088
MSE-*t*	Normal	0.047	0.049	0.081	0.070	0.061	0.085	0.078
MSE-*t*, 2-sided	Normal	0.093	0.098	0.108	0.094	0.093	0.102	0.099

DGP 5, 4-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.102	0.091	0.156	0.111	0.094	0.162	0.114
MSE-*F*	Fixed regressor	0.148	0.142	0.131	0.132	0.131	0.128	0.124
MSE-*t*	Non-parametric	0.110	0.094	0.152	0.114	0.097	0.152	0.115
MSE-*t*	Fixed regressor	0.133	0.136	0.122	0.117	0.123	0.117	0.113
MSE-*t*	Normal	0.115	0.103	0.158	0.115	0.105	0.162	0.119
MSE-*t*, 2-sided	Normal	0.154	0.150	0.209	0.161	0.152	0.209	0.165

Notes:
1. The data generating processes are defined in Eqs. (5), (8), (11), (14) and (17). In these experiments, the coefficients μ or b_{ij} are scaled such that the null and alternative models are expected to be equally accurate (on average) over the forecast sample.
2. For each artificial data set, forecasts of $y_{t+\tau}$ (where τ denotes the forecast horizon) are formed recursively using estimates of Eqs. (6) and (7) in the DGP 1 experiments, Eqs. (9) and (10) in DGP 2 experiments, Eqs. (12) and (13) in DGP 3 experiments, Eqs. (15) and (16) in DGP 4 experiments, and Eqs. (18) and (19) in DGP 5 experiments. These forecasts are then used to form the indicated test statistics, defined in Section 3.2. R and P refer to the number of in-sample observations and 1-step ahead forecasts, respectively.
3. In each Monte Carlo replication, the simulated test statistics are compared against bootstrapped critical values, using a significance level of 10%. Section 3.5 describes the bootstrap procedures.
4. The number of Monte Carlo simulations is 5000; the number of bootstrap draws is 499.

12.4% to 14.8% percent.[18] For multi-step horizons, using the fixed regressor bootstrap works better (yielding rates closer to nominal size) when R is relatively large than when R is relatively small. Rejection rates for the MSE-t test compared against critical values from the fixed regressor bootstrap are similar, although a bit lower, ranging from 7.7% to 10.2% at the 1-step horizon and from 11.3% to 13.6% at the 4-step horizon.

Tests based on the other bootstrap intended to test the null of equal accuracy, the non-parametric bootstrap, are somewhat – although not entirely – less reliable indicators of equal accuracy. With critical values from the non-parametric bootstrap, the MSE-F test is somewhat undersized at the 1-step horizon but correctly sized or somewhat oversized at the 4-step horizon. As shown in Table 1, the MSE-F test's rejection rate ranges from 4.1% to 8.4% at the 1-step horizon and from 9.1% to 16.2% at the 4-step horizon. With the non-parametric approach, empirical rejection rates generally rise as P/R falls. For example, with 4-step ahead forecasts (for DGP 5) and $R = 80$, the MSE-F rejection rate is 9.4% when $P = 120$ and 15.6% when $P = 40$. Rejection rates for the MSE-t test compared

[18] The over-sizing of the fixed regressor bootstrap at the 4-step horizon most likely has to do with the HAC estimation of the variance matrix V that determines the coefficient rescaling factor.

Table 2
Monte Carlo results on size, additional DGPs (nominal size = 10%).

DGP 6, 1-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.092	0.078	0.111	0.095	0.081	0.112	0.099
MSE-*F*	Fixed regressor	0.120	0.110	0.107	0.104	0.104	0.097	0.103
MSE-*t*	Non-parametric	0.111	0.094	0.143	0.110	0.095	0.137	0.117
MSE-*t*	Fixed regressor	0.103	0.102	0.099	0.096	0.098	0.098	0.095
MSE-*t*	Normal	0.096	0.080	0.109	0.092	0.082	0.114	0.098
MSE-*t*, 2-sided	Normal	0.090	0.084	0.106	0.098	0.093	0.112	0.092

DGP 7, 4-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.102	0.089	0.157	0.112	0.101	0.168	0.113
MSE-*F*	Fixed regressor	0.193	0.188	0.159	0.158	0.169	0.137	0.134
MSE-*t*	Non-parametric	0.120	0.106	0.169	0.123	0.112	0.169	0.118
MSE-*t*	Fixed regressor	0.168	0.170	0.139	0.140	0.149	0.125	0.120
MSE-*t*	Normal	0.123	0.110	0.171	0.128	0.117	0.169	0.122
MSE-*t*, 2-sided	Normal	0.141	0.115	0.218	0.148	0.132	0.213	0.150

DGP 8, 1-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.065	0.061	0.088	0.071	0.067	0.088	0.070
MSE-*F*	Fixed regressor	0.120	0.120	0.109	0.113	0.111	0.106	0.108
MSE-*t*	Non-parametric	0.077	0.068	0.103	0.081	0.075	0.105	0.080
MSE-*t*	Fixed regressor	0.104	0.104	0.100	0.099	0.103	0.098	0.092
MSE-*t*	Normal	0.072	0.067	0.094	0.077	0.072	0.095	0.073
MSE-*t*, 2-sided	Normal	0.098	0.095	0.123	0.111	0.109	0.116	0.118

Notes:
1. See the notes to Table 1.
2. These experiments address the properties of our proposed testing procedures in DGP and forecasting settings in which the assumptions necessary to prove the validity of the bootstrap are not satisfied. As detailed in Section 4.1, in DGP 6 the number of variables included in model 1 and not model 0 is 3 and the forecast errors are conditionally heteroskedastic. In DGP 7, the number of variables included in model 1 and not model 0 is 3 and the forecast errors are serially correlated. In DGP 8, the forecasting models are misspecified, in the sense that neither model includes two variables that appear in the data-generating process for the predictand. In each DGP, the coefficients b_{ij} are scaled such that the null and alternative models are equally accurate (on average) over the forecast sample.

against critical values from the non-parametric bootstrap are similar, although typically a bit higher, ranging from 5.0% to 10.1% at the 1-step horizon and from 9.4% to 15.2% at the 4-step horizon.

In addition, comparing the MSE-*t* test against standard normal critical values (with a one-sided testing approach) yields results similar to those obtained by comparing the test statistic against critical values from the non-parametric bootstrap. For instance, at the 1-step horizon, MSE-*t* rejection rates range from 4.7% to 9.4% under standard normal critical values, compared to a range of 5.0% to 10.1% under the non-parametric bootstrap. Accordingly, the MSE-*t* test compared against standard normal critical values is somewhat undersized at the 1-step horizon but correctly or somewhat oversized at the 4-step horizon.

Table 2 presents results from some additional experiments, with DGPs 6–8, that address the effectiveness of our proposed bootstrap in forecasting applications for which we are unable to prove the asymptotic validity of the bootstrap or the forecasting models are misspecified. In the DGP 6 and 7 experiments, the larger forecasting model has three more variables than the smaller model (so β_w is not scalar), and the forecast errors are either conditionally heteroskedastic (DGP 6, which extends DGP 4 to include conditional heteroskedasticity) or serially correlated (DGP 7, which extends DGP 5 to include more variables). The DGP 8 experiments address the case of misspecification of the forecasting models, using a data-generating process taken from DGP 4 but (misspecified) forecasting models taken from DGP 3.

The results of the additional experiments with DGPs 6 and 7 provided in Table 2 indicate that our fixed regressor bootstrap continues to perform well, in line with the baseline results of Table 1. In the case of the MSE-*F* test applied to 1-step ahead forecasts (DGP 6), rejection rates range from 9.7% to 12.0%. For the MSE-*t* test applied to 1-step ahead forecasts (DGP 6), rejection rates range from 9.5% to 10.3%. Again, rejection rates for 4-step ahead forecast tests are somewhat higher than for 1-step ahead forecasts. With DGP 7, the sizes of the MSE-*F* and MSE-*t* tests range from 13.4% to 19.3% and from 12.0% to 17.0%, respectively. The oversizing is modestly greater with DGP 7 (for which $k_w = 3$) than DGP 5 (for which $k_w = 1$). With both of these DGPs, the size of the fixed regressor-based tests improves as R increases (in some unreported experiments with DGP 7, we have verified that size improves further with even larger sample sizes). The modest difference in results between DGP 5 and DGP 7 is most likely due to the additional regressors in DGP 7 (relative to DGP 4) further reducing the finite-sample precision of the HAC estimation of the variance matrix V that determines the coefficient rescaling factor used in the bootstrap.

The performance of tests based on other sources of critical values is qualitatively the same in DGPs 6 and 7 (Table 2) as in DGPs 1–5 (Table 1). The key difference across the sets of experiments is that conditional heteroskedasticity raises and, in most cases, improves the sizes of MSE-*F* and MSE-*t* tests based on the non-parametric bootstrap and MSE-*t* tests based on standard normal critical values.[19] The rejection rate for the MSE-*F* test compared against critical values from the non-parametric bootstrap ranges from 7.8% to 11.2% in DGP 6, compared to 4.1% to 8.0% in DGP 4. The rejection rate for the MSE-*t* test compared against critical values from the non-parametric bootstrap ranges from 9.4% to 14.3% in DGP 6, compared to 5.0% to 10.0% in DGP 4 (the pattern is similar with standard normal critical values).

[19] This finding does not appear to be dependent on having k_w exceed 1. In unreported experiments, we obtained a similar result in a version of DGP 3 with conditional heteroskedasticity.

Table 3
Monte Carlo results on size, rolling forecasts (nominal size = 10%).

DGP 1, 1-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.039	0.037	0.073	0.057	0.041	0.076	0.063
MSE-*F*	Fixed regressor	0.099	0.104	0.093	0.107	0.097	0.086	0.094
MSE-*t*	Non-parametric	0.048	0.044	0.076	0.062	0.045	0.075	0.064
MSE-*t*	Fixed regressor	0.093	0.101	0.093	0.099	0.096	0.087	0.092
MSE-*t*	Normal	0.046	0.043	0.077	0.062	0.046	0.075	0.062
MSE-*t*, 2-sided	Normal	0.086	0.080	0.115	0.094	0.084	0.114	0.099

DGP 2, 1-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.047	0.038	0.078	0.055	0.049	0.081	0.063
MSE-*F*	Fixed regressor	0.106	0.107	0.105	0.100	0.105	0.096	0.097
MSE-*t*	Non-parametric	0.059	0.052	0.100	0.065	0.059	0.094	0.072
MSE-*t*	Fixed regressor	0.088	0.091	0.094	0.088	0.089	0.085	0.085
MSE-*t*	Normal	0.053	0.047	0.088	0.057	0.052	0.084	0.067
MSE-*t*, 2-sided	Normal	0.082	0.084	0.114	0.098	0.092	0.106	0.094

DGP 3, 1-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.036	0.032	0.078	0.052	0.039	0.080	0.065
MSE-*F*	Fixed regressor	0.097	0.099	0.103	0.097	0.098	0.102	0.103
MSE-*t*	Non-parametric	0.049	0.041	0.092	0.063	0.049	0.096	0.076
MSE-*t*	Fixed regressor	0.086	0.088	0.092	0.089	0.088	0.092	0.093
MSE-*t*	Normal	0.044	0.036	0.083	0.060	0.043	0.086	0.067
MSE-*t*, 2-sided	Normal	0.100	0.105	0.112	0.108	0.091	0.123	0.110

DGP 4, 1-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.020	0.018	0.062	0.044	0.034	0.080	0.060
MSE-*F*	Fixed regressor	0.074	0.080	0.087	0.090	0.088	0.084	0.094
MSE-*t*	Non-parametric	0.030	0.027	0.086	0.058	0.044	0.098	0.076
MSE-*t*	Fixed regressor	0.068	0.076	0.084	0.087	0.080	0.086	0.091
MSE-*t*	Normal	0.028	0.023	0.076	0.053	0.039	0.085	0.070
MSE-*t*, 2-sided	Normal	0.124	0.141	0.103	0.093	0.095	0.107	0.099

DGP 5, 4-step forecasts

Statistic	Source of critical values	$R = 40$ $P = 80$	$R = 40$ $P = 120$	$R = 80$ $P = 40$	$R = 80$ $P = 80$	$R = 80$ $P = 120$	$R = 120$ $P = 40$	$R = 120$ $P = 80$
MSE-*F*	Non-parametric	0.112	0.103	0.146	0.104	0.091	0.165	0.110
MSE-*F*	Fixed regressor	0.160	0.162	0.132	0.137	0.140	0.129	0.126
MSE-*t*	Non-parametric	0.132	0.127	0.151	0.114	0.101	0.162	0.118
MSE-*t*	Fixed regressor	0.142	0.148	0.120	0.126	0.132	0.116	0.114
MSE-*t*	Normal	0.128	0.123	0.156	0.115	0.102	0.165	0.115
MSE-*t*, 2-sided	Normal	0.158	0.147	0.198	0.166	0.143	0.208	0.153

Notes:
1. See the notes to Table 1.
2. In these experiments, the forecasting scheme is rolling, rather than recursive.

Introducing misspecification of the forecast models, as in the experiments with DGP 8, does not significantly affect the performance of our proposed testing approach. The size of the MSE-*F* and MSE-*t* tests based on our bootstrapped critical values (obtained with our fixed regressor approach) is very similar to their performance in correctly specified forecasting models. The tests are roughly correctly sized, with rejection rates for the MSE-*F* test across *R*, *P* combinations ranging from 10.6% to 12.0%, and rates for the MSE-*t* test ranging from 9.2% to 10.4%. Similarly, using critical values obtained with the non-parametric bootstrap or from the normal distribution yields results qualitatively very similar to those in the baseline experiments.

4.2.2. *Size results: rolling forecasts*

Table 3 provides size results for experiments using a rolling forecast scheme instead of the baseline recursive scheme, based on models parameterized to make the null and alternative forecasting models equally accurate (the necessary scaling factor in the DGP is a bit different in the rolling case than the recursive). In general, the results for the rolling scheme are very similar to those for the recursive. Tests based on our fixed regressor bootstrap have size of about 10% (the nominal size), although with some slight to modest oversizing at the 4-step horizon. Tests based on the non-parametric bootstrap or standard normal critical values continue to be undersized at the 1-step horizon, although the problem is a bit worse under the rolling scheme than the recursive.[20] For example, with DGP 3, $R = 40$, and $P = 80$, comparing the MSE-*t* test against

[20] The rise in rejection rates that occurs as P/R falls is a bit sharper in the rolling case than the recursive. As a consequence, the differences in rejection rates (based on the non-parametric bootstrap or standard normal critical values) across the recursive and rolling forecasting schemes are larger when P/R is relatively big than when it is relatively small.

critical values estimated with the non-parametric bootstrap yields a rejection rate of 6.5% for recursive forecasts (Table 1) and 4.9% for rolling forecasts (Table 3); comparing the test against fixed regressor bootstrap critical values yields corresponding rejection rates of 8.8 (recursive) and 8.6% (rolling). At the 4-step horizon, tests based on the non-parametric bootstrap or standard normal critical values continue to range from correctly sized to oversized, with oversizing that is sharpest when P is small.

Our rolling scheme results on the behavior of the MSE-t test compared against non-parametric bootstrap and standard normal critical values are somewhat at odds with the behavior of the test in Giacomini and White (2006). Giacomini and White (2006) compare the MSE-t test against standard normal critical values, and find a two-sided test to be roughly correctly sized at the one-step forecast horizon, with small-to-modest undersizing for some sample sizes and comparable oversizing for others. One source of differences in results is our treatment of the test as one-sided rather than two-sided. Giacomini and White (2006) permit rejections of the alternative model in favor of the null and conduct two-sided tests; we prefer to take the small model as the null and only consider rejections of the null in favor of the alternative, or one-sided tests. When we use a two-sided MSE-t test and standard normal critical values (while not shown in the interest of brevity, the same applies with critical values from the non-parametric bootstrap), the test is roughly correctly sized at the 1-step horizon and correctly sized to somewhat oversized at the 4-step horizon (the same applies in the recursive forecast results of Table 1). The increase in rejection rates that occurs with the move from a one-sided to two-sided test likely reflects an empirical distribution that is shifted to the left relative to the standard normal.

Admittedly, though, other aspects of our Monte Carlo results seem to be at odds with the asymptotic results of Giacomini and White (2006), if not their Monte Carlo results. Their asymptotics imply the MSE-t test has an asymptotic distribution that is standard normal for rolling forecasts but not recursive forecasts, suggesting the test should have better size properties in the rolling case than the recursive. But in our Monte Carlo results, the standard normal approximation for MSE-t seems to work better with recursive forecasts than rolling, yielding 1-step ahead rejection rates closer to nominal in the former case than the latter. In addition, their theory rests on asymptotics that treat R as fixed and P as limiting to infinity, which suggests the test should behave better when P is large relative to R than when P is relatively small. In fact, in our Monte Carlo results, rejection rates based on the non-parametric bootstrap and standard normal critical values tend to be farther from nominal size when P is large than when it is small. In the case of the second issue, the Monte Carlo results in Giacomini and White (2006) seem to yield a similar pattern, with rejection rates falling as the forecast sample increases relative to the estimation sample, often to levels consistent with the undersizing we have reported.

4.2.3. *Power results: recursive forecasts*

Table 4 provides results for DGPs in which the b_{ij} coefficients on some x variables (and the μ intercept in DGP 1) are large enough that, under our finite-sample asymptotics, the alternative model is expected to be more accurate than the null model in the finite sample.

Comparing the test statistics against critical values estimated with the fixed regressor bootstrap seems to yield relatively good power. In the case of the MSE-F test, rejection rates range from 42.8% to 85.3%. Comparing tests against distributions estimated with the non-parametric bootstrap yields materially lower power. In Table 4's results, using the non-parametric bootstrap for the MSE-F test yields a rejection rate between 25.0% and 58.0%.

Rejection rates for the MSE-t test are broadly similar to those for the MSE-F test, although with some noticeable differences. In most cases in Table 4's results, the MSE-t test is less powerful than the MSE-F test (as with the fixed regressor bootstrap), but in some cases (as with the non-parametric bootstrap), the MSE-t test is more powerful.

4.2.4. *Results summary*

Overall, the Monte Carlo results show that, for testing equal forecast accuracy over a given sample, our proposed fixed regressor bootstrap works well. When the null of equal accuracy in the finite sample is true, the testing procedures yield approximately correctly sized tests. When an alternative model is, in truth, more accurate than the null, the testing procedures seem to have reasonable power. The non-parametric bootstrap procedure, which just re-samples the data without imposing the equal accuracy null in the data generation, tends to be less reliable when applied to nested forecasting models.

5. Applications

In this section we use the tests and inference approaches described above in forecasting excess stock returns and core inflation, both for the US. Some recent examples from the long literature on stock return forecasting include Rapach and Wohar (2006), Goyal and Welch (2008), and Campbell and Thompson (2008). Some recent inflation examples include Atkeson and Ohanian (2001) and Stock and Watson (2003).

More specifically, in the stock return application, we use the data of Goyal and Welch (2008) and examine forecasts of monthly excess stock returns (CRSP excess returns measured on a log basis) from a total of 17 models. The null model includes just a constant. The alternative models add in one lag of a common predictor, taken from the set of variables in the Goyal–Welch data set available over all of our sample.[21] These include, among others, the dividend–price ratio, the earnings–price ratio, and the cross-sectional premium. The full set of 16 predictive variables is listed in Table 5, with details provided in Goyal and Welch (2008). Following studies such as Pesaran and Timmermann (1995), we focus on the post-war period. Our model estimation sample begins with January 1954, and we examine recursive 1-month ahead forecasts (that is, our estimation sample expands as forecasting moves forward in time) for 1970 through 2002.

In the inflation application, we examine 1-quarter ahead and 4-quarter ahead forecasts of core PCE inflation obtained from a few models, over a sample of 1985:Q1+horizon-1 to 2008:Q2. The null model includes a constant and lags of the change in inflation. One alternative model adds one lag of the CFNAI to the baseline model. Another includes one lag of the CFNAI, PCE food price inflation less core inflation, and import price inflation less core inflation.[22] We specify the models in terms of the change in inflation, following, among others, Stock and Watson (1999, 2003) and Clark and McCracken (2006). In one application, we consider 1-quarter ahead forecasts of inflation defined as $\pi_t = 400\ln(P_t/P_{t-1})$, using models relating $\Delta\pi_{t+1}$ to a constant, $\Delta\pi_t$, $\Delta\pi_{t-1}$, and the period t values of the CFNAI, relative food price inflation, and relative import price inflation. In another, we consider 4-quarter ahead forecasts of inflation defined as $\pi_t^{(4)} = 100\ln(P_t/P_{t-4})$, using models relating $\pi_{t+4}^{(4)} - \pi_t^{(4)}$ to a constant, $\pi_t^{(4)} - \pi_{t-4}^{(4)}$, and the period t values of the CFNAI, relative food price inflation, and relative import price inflation. To simplify the lag structure necessary for reasonable forecasting models, the (relative) food and import price

[21] We obtained the data from Amit Goyal's website.

[22] We obtained the CFNAI data from the Chicago Fed's website and the rest of the data from the FAME database of the Federal Reserve Board of Governors.

Table 4
Monte Carlo results on power (nominal size = 10%).

DGP 1, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.380	0.516	0.302	0.443	0.580	0.309	0.458
MSE-*F*	Fixed regressor	0.642	0.781	0.570	0.749	0.853	0.629	0.787
MSE-*t*	Non-parametric	0.414	0.543	0.324	0.474	0.610	0.335	0.490
MSE-*t*	Fixed regressor	0.505	0.657	0.358	0.547	0.696	0.364	0.556
MSE-*t*	Normal	0.418	0.548	0.323	0.477	0.612	0.339	0.491
MSE-*t*, 2-sided	Normal	0.265	0.388	0.218	0.330	0.442	0.224	0.343

DGP 2, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.272	0.354	0.248	0.330	0.423	0.264	0.352
MSE-*F*	Fixed regressor	0.486	0.599	0.451	0.589	0.705	0.523	0.649
MSE-*t*	Non-parametric	0.313	0.385	0.299	0.365	0.459	0.308	0.388
MSE-*t*	Fixed regressor	0.361	0.464	0.288	0.404	0.529	0.293	0.419
MSE-*t*	Normal	0.304	0.384	0.277	0.354	0.450	0.287	0.379
MSE-*t*, 2-sided	Normal	0.185	0.247	0.175	0.222	0.300	0.182	0.239

DGP 3, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.263	0.351	0.250	0.335	0.422	0.269	0.363
MSE-*F*	Fixed regressor	0.481	0.609	0.445	0.593	0.715	0.518	0.659
MSE-*t*	Non-parametric	0.296	0.385	0.295	0.372	0.457	0.311	0.397
MSE-*t*	Fixed regressor	0.360	0.487	0.280	0.412	0.534	0.294	0.425
MSE-*t*	Normal	0.282	0.374	0.270	0.352	0.448	0.285	0.380
MSE-*t*, 2-sided	Normal	0.178	0.233	0.172	0.232	0.300	0.184	0.251

DGP 4, 1-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.282	0.434	0.268	0.429	0.569	0.319	0.484
MSE-*F*	Fixed regressor	0.527	0.697	0.491	0.685	0.821	0.579	0.763
MSE-*t*	Non-parametric	0.349	0.485	0.346	0.497	0.616	0.396	0.547
MSE-*t*	Fixed regressor	0.426	0.601	0.329	0.533	0.680	0.366	0.568
MSE-*t*	Normal	0.331	0.474	0.319	0.476	0.606	0.368	0.527
MSE-*t*, 2-sided	Normal	0.200	0.322	0.207	0.320	0.451	0.241	0.370

DGP 5, 4-step forecasts

Statistic	Source of critical values	$R = 40$	$R = 40$	$R = 80$	$R = 80$	$R = 80$	$R = 120$	$R = 120$
		$P = 80$	$P = 120$	$P = 40$	$P = 80$	$P = 120$	$P = 40$	$P = 80$
MSE-*F*	Non-parametric	0.290	0.349	0.315	0.347	0.421	0.342	0.383
MSE-*F*	Fixed regressor	0.467	0.563	0.428	0.557	0.649	0.509	0.611
MSE-*t*	Non-parametric	0.324	0.379	0.328	0.375	0.442	0.366	0.406
MSE-*t*	Fixed regressor	0.360	0.440	0.270	0.380	0.487	0.286	0.399
MSE-*t*	Normal	0.332	0.393	0.339	0.385	0.460	0.373	0.419
MSE-*t*, 2-sided	Normal	0.244	0.279	0.281	0.284	0.336	0.299	0.302

Notes:
1. See the notes to Table 1.
2. In these experiments, the DGP coefficients μ or b_{ij} are set to values (given in Section 4.1) large enough that the alternative model is expected to be more accurate than the null model.

inflation variables are computed as two-period averages of quarterly (relative) inflation rates. For both inflation forecast horizons, our model estimation sample uses a start date of 1968:Q3.

Results for the stock return and inflation forecast applications are reported in Tables 5 and 6. The tables provide, for each alternative model, the ratio of the MSE of forecasts from the benchmark model to the alternative model's forecast MSE. The tables also provide, for the MSE-*F* test, *p*-values obtained under three different approaches.[23] Two of these approaches are the same ones included in last section's Monte Carlo analysis: the non-parametric bootstrap and our proposed fixed regressor bootstrap.

A third approach – a no-predictability version of the fixed regressor bootstrap – is included to illustrate some differences in testing equal accuracy in the finite sample versus equal accuracy in population. The theoretical results in Clark and McCracken (2012) on comparisons of forecasts from multiple nested models establish the asymptotic validity of the no-predictability fixed regressor bootstrap for the null of equal forecast accuracy in population, which is equivalent to a null hypothesis of $\beta_w = 0$. This no-predictability fixed regressor bootstrap takes the same form as described in Section 3.5, with the sole difference being that in step 1, $\hat{d} = 0$, which is equivalent to simply estimating the null forecasting model by OLS (model 0, which includes only the variables $x_{0,t}$) rather than the alternative model (model 1, which includes the variables $x_{0,t}$ and $x_{w,t}$).

In the case of excess stock returns, the evidence in Table 5 is consistent with much of the literature: return predictability is limited. Of the 16 alternative forecasting models, only two – the first two in the table – have MSEs lower than the benchmark

[23] In all three approaches, we use 9999 replications in computing the bootstrap *p*-values.

Table 5
Tests of equal accuracy for monthly stock returns.

Alternative model variable	MSE(null)/ MSE(altern.)	MSE-F Bootstrap p–values		
		Non-param.	No predictability fixed regressor	Fixed regressor
Cross-sectional premium	1.009	0.136	0.001	0.071
Return on long-term Treasury	1.005	0.381	0.024	0.177
BAA-AAA yield spread	0.996	0.688	0.828	0.487
BAA-AAA return spread	0.995	0.824	0.933	0.779
Net equity expansion	0.994	0.648	0.358	0.659
CPI inflation	0.993	0.646	0.587	0.776
Stock variance	0.992	0.773	0.512	0.230
Dividend–payout ratio	0.991	0.681	0.572	0.724
Term (yield) spread	0.987	0.724	0.939	0.984
Earnings–price ratio	0.985	0.938	0.383	0.933
10-year earnings–price ratio	0.983	0.876	0.985	0.984
3-month T-bill rate	0.982	0.739	0.952	0.993
Dividend–price ratio	0.981	0.843	0.550	0.993
Dividend yield	0.981	0.836	0.436	0.996
Yield on long-term Treasury	0.978	0.796	0.988	0.995
Book–market ratio	0.965	0.996	0.967	0.994

Notes:
1. As described in Section 5, monthly forecasts of excess stock returns in period $t + 1$ are generated recursively from a null model that includes just a constant and 15 alternative models that include a constant and the period R ($t - 1$ in the case of CPI inflation) value of each of the variables listed in the first column. Forecasts from January 1970 to December 2002 are obtained from models estimated with a data sample starting in January 1954.
2. For each alternative model, the table reports the ratio of the null model's forecast MSE to the alternative model's MSE and bootstrapped p-values for the null hypothesis of equal accuracy, based on the MSE-F statistic. Section 3.5 details the bootstrap methods. The RMSE of the null model is 0.046.

Table 6
Tests of equal accuracy for core inflation.

Alternative model variables	MSE(null)/MSE(altern.)	MSE-F Bootstrap p-values		
		Non-param.	No predictability fixed regressor	Fixed regressor
1-quarter horizon				
CFNAI	1.016	0.343	0.092	0.293
CFNAI, food, imports	1.098	0.100	0.001	0.062
4-quarter horizon				
CFNAI	0.921	0.675	0.881	0.915
CFNAI, food, imports	1.279	0.317	0.000	0.031

Notes:
1. As described in Section 5, 1-quarter and 4-quarter ahead forecasts of core PCE inflation (specified as a period $t + \tau$ predictand) are generated recursively from a null model that includes a constant and lags of inflation (from period R and earlier) and alternative models that include one lag (period R values) of the variables indicated in the table (defined further in Section 5). The 1-quarter forecasts are of quarterly inflation; the 4-quarter forecasts are of 4-quarter inflation. Forecasts from 1985:Q1 + $\tau - 1$ through 2008:Q2 are obtained from models estimated with a data sample starting in 1968:Q3.
2. For each of the alternative models, the table reports the ratio of the null model's forecast MSE to the alternative model's MSE and bootstrapped p-values for the null hypothesis of equal accuracy, based on the MSE-F statistic. Section 3.5 details the bootstrap methods. The RMSE of the null model is 0.613 at the 1-quarter horizon and 0.444 at the 4-quarter horizon.

(that is, MSE ratios greater than 1). The no-predictability fixed regressor bootstrap p-values reject the null model in favor of the alternative for each of these two models. These test results indicate the predictor coefficients on the cross-sectional premium and return on long-term Treasuries are non-zero. However, p-values based on the fixed regressor bootstrap imply weaker evidence of forecastability, with the null of equal forecast accuracy rejected for the cross-sectional premium, but not the Treasury return (at a 10% significance level). This pattern suggests that, while the coefficient on the Treasury return may differ from zero, the coefficient is not large enough that a model including the Treasury return would be expected to forecast better than the null model over a sample of the size considered. Critical values based on the non-parametric bootstrap yield no rejections, presumably (given our Monte Carlo evidence) reflecting lower power.

The inflation results reported in Table 6 yield similarly mixed evidence of predictability. By itself, the CFNAI improves the accuracy of 1-quarter ahead forecasts but not 4-quarter ahead forecasts. At the 1-step horizon, the no-predictability fixed regressor bootstrap p-values reject the null model in favor of the alternative — indicating the predictor coefficients on the CFNAI to be non-zero. However, p-values based on the fixed regressor bootstrap fail to reject the null of equal accuracy. So while the coefficient on the CFNAI may differ from zero, it is not large enough that a model including the CFNAI would be expected to forecast better than the null model in a sample of the size considered. Including not only the CFNAI but also relative food and import price inflation yields larger gains in forecast accuracy, at both horizons. In this case, critical values from both the no-predictability fixed regressor and fixed regressor bootstrap reject the null (at a 10% significance level). This suggests the relevant coefficients are non-zero and large enough to make the alternative model more accurate than the null. Here, too, critical values based on the non-parametric bootstrap yield fewer rejections.

6. Conclusion

This paper develops bootstrap methods for testing whether, in a finite sample, competing out-of-sample forecasts from nested models are equally accurate. Most prior work on forecast tests for nested models has focused on a null hypothesis of equal accuracy in population — basically, whether coefficients on the extra variables in the larger, nesting model are zero. We instead use an asymptotic approximation that treats the coefficients as non-zero but small, such that, in a finite sample, forecasts from the small model are expected to be as accurate as forecasts from the large model. While an unrestricted, correctly specified model might have better population-level predictive ability than a misspecified restricted model, it need not do so in finite samples due to imprecision in the additional parameter estimates. In the presence of these "weak" predictors, we show how to test the null of equal average predictive ability over a given sample size.

Under our asymptotic approximation of weak predictive ability, we first derive the asymptotic distributions of two tests for equal out-of-sample predictive ability. We then develop a parametric bootstrap procedure – a fixed regressor bootstrap – for testing the null of equal finite-sample forecast accuracy. We next conduct a range of Monte Carlo simulations to examine the finite-sample properties of the tests and bootstrap procedures. Our proposed fixed regressor bootstrap works reasonably well: when the null of equal finite-sample predictive ability is true, the testing procedure yields approximately correctly sized tests. Moreover when an alternative model is, in truth, more accurate than the null, the testing procedure has reasonable power. In contrast, when applied to nested models, the non-parametric method of White (2000) often does not work as well, in a size or power sense.

In the final part of our analysis, we apply our proposed methods for testing equal predictive ability to forecasts of excess stock returns and core inflation, using US data. In both applications, our methods for testing equal finite sample accuracy yield weaker evidence of predictability than do methods for testing equal population-level accuracy. There remains some evidence, but only modest. Using non-parametric bootstrap methods that are technically invalid with nested models – methods that have poorer size and power properties – yields much less evidence of predictability.

Acknowledgments

Much of the research for this paper was completed while Clark was on the staff of the Federal Reserve Bank of Kansas City. We gratefully acknowledge excellent research assistance from Stephen Terry and helpful comments from two referees and the associate editor, Bruce Hansen, Lutz Kilian, Norm Swanson, and seminar or conference participants at the University of Kansas, University of Michigan, Rutgers University, HEC Montreal, Bank of Spain, Midwest Econometrics Group, Missouri Economics Conference, NBER Summer Institute, and the ECB/Bundesbank/CFS Joint Lunchtime Seminar. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Banks of Cleveland or St. Louis or the Federal Reserve Board of Governors.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2014.06.016.

References

Atkeson, A., Ohanian, L.E., 2001. Are Phillips curves useful for forecasting inflation? Federal Reserve Bank of Minneapolis Quarterly Review 25, 2–11.

Bachmeier, L., Swanson, N.R., 2005. Predicting inflation: does the quantity theory help? Econ. Inquiry 43, 570–585.

Bordo, M., Haubrich, J.G., 1875-1997. The yield curve as a predictor of growth: long-run evidence. Rev. Econ. Stat. 90, 182–185.

Bruneau, C., De Bandt, O., Flageollet, A., Michaux, E., 2007. Forecasting inflation using economic indicators: the case of France. J. Forecast. 2, 1–22.

Butler, A.W., Grullon, G., Weston, J.P., 2005. Can managers forecast aggregate market returns? J. Finance 60, 963–986.

Campbell, J.Y., Thompson, S.B., 2008. Predicting excess stock returns out of sample: can anything beat the historical average? Rev. Financ. Stud. 21, 1509–1531.

Chen, Y., Rogoff, K.S., Rossi, B., 2010. Can exchange rates forecast commodity prices? Q. J. Econ. 125, 1145–1194.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. J. Econometrics 105, 85–110.

Clark, T.E., McCracken, M.W., 2005a. Evaluating direct multistep forecasts. Econometric Rev. 24, 369–404.

Clark, T.E., McCracken, M.W., 2005b. The power of tests of predictive ability in the presence of structural breaks. J. Econometrics 124, 1–31.

Clark, T.E., McCracken, M.W., 2006. The predictive content of the output gap for inflation: resolving in-sample and out-of-sample evidence. J. Money Credit Banking 38, 1127–1148.

Clark, T.E., McCracken, M.W., 2009. Combining forecasts from nested models. Oxford Bulletin of Economics and Statistics 71, 303–329.

Clark, T.E., McCracken, M.W., 2012. Reality checks and comparisons of nested predictive models. J. Bus. Econom. Statist. 30, 53–66.

Cooper, M., Gulen, H., 2006. Is time-series-based predictability evident in real time? J. Bus. 79, 1263–1292.

Corradi, V., Swanson, N.R., 2002. A consistent test for nonlinear out of sample predictive accuracy. Journal of Econometrics 110, 353–381.

de Jong, R.M., Davidson, J., 2000. The functional central limit theorem and weak convergence to stochastic integrals I: weakly dependent processes. Econometric Theory 16, 621–642.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econom. Statist. 13, 253–263.

Diebold, F.X., Rudebusch, G.D., 1991. Forecasting output with the composite leading index: a real-time analysis. J. Amer. Statist. Assoc. 86, 603–610.

Ferreira, M.A., Santa-Clara, P., 2011. Forecasting stock market returns: the sum of the parts is more than the whole. J. Financ. Econ. 100, 514–537.

Giacomini, R., Rossi, B., 2006. How stable is the forecasting performance of the yield curve for output growth? Oxford Bull. Econ. Stat. 68, 783–795.

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

Goncalves, S., Kilian, L., 2007. Asymptotic and bootstrap inference for AR(∞) processes with conditional heteroskedasticity. Econometric Rev. 26, 609–641.

Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21, 1455–1508.

Groen, J.J., 1999. Long horizon predictability of exchange rates: is it for real? Empir. Econom. 24, 451–469.

Guo, H., 2006. On the out-of-sample predictability of stock market returns. J. Bus. 27, 645–670.

Hansen, B.E., 1996. Erratum: The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. J. Appl. Econometrics 11, 195–198.

Hansen, B.E., 2008. Generalized Shrinkage Estimators. University of Wisconsin, Manuscript.

Hjalmarsson, E., 2009. Should we expect significant out-of-sample results when predicting stock returns? In: Ellison, G.I. (Ed.), Stock Returns: Cyclicity, Prediction and Economic Consequences. Nova Science Publishers, New York, pp. 269–274.

Inoue, A., Kilian, L., 2004. In-sample or out–of-sample tests of predictability? Which one should we use? Econometric Rev. 23, 371–402.

Inoue, A., Rossi, B., 2008. Monitoring and forecasting financial crises. J. Money Credit Banking 40, 523–534.

Kilian, L., 1999. Exchange rates and monetary fundamentals: what do we learn from long-horizon regressions? J. Appl. Econometrics 14, 491–510.

Kilian, L., Taylor, M.P., 2003. Why is it so difficult to beat the random walk forecast of exchange rates? J. Int. Econ. 60, 85–107.

Lettau, M., Ludvigson, S., 2001. Consumption, aggregate wealth, and expected stock returns. J. Finance 56, 815–849.

Mark, N.C., 1995. Exchange rates and fundamentals: evidence on long-horizon predictability. Amer. Econ. Rev. 85, 201–218.

McCracken, M.W., 2007. Asymptotics for out-of-sample tests of granger causality. J. Econometrics 140, 719–752.

Meese, R., Rogoff, K., 1988. Was it real? The exchange rate-interest differential relation over the modern floating-rate period. J. Finance 43, 933–948.

Molodtsova, T., Papell, D.H., 2009. Out-of-sample exchange rate predictability with Taylor Rule fundamentals. J. Int. Econ. 77, 167–180.

Pesaran, M.H., Timmermann, A., 1995. Predictability of stock returns: robustness and economic significance. J. Finance 50, 1201–1228.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Amer. Statist. Assoc. 89, 1303–1313.

Rapach, D., Wohar, M.E., 2006. In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. J. Empir. Finance 13, 231–247.

Stock, J.H., Watson, M.W., 1999. Forecasting inflation. J. Monetary Econ. 44, 293–335.

Stock, J.H., Watson, M.W., 2003. Forecasting output and inflation: The role of asset prices. J. Econ. Lit. 41, 788–829.

Trenkler, G., Toutenburg, H., 1992. Pre-test procedures and forecasting in the regression model under restrictions. J. Statist. Plann. Inference 30, 249–256.

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

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1127.