Received: 17 May 2021 | Revised: 20 June 2022 | Accepted: 24 March 2023
DOI: 10.1002/for.2987

RESEARCH ARTICLE

WILEY

On bootstrapping tests of equal forecast accuracy for nested models

Firmin Doko Tchatoka[1] | **Qazi Haque**[1,2]

[1]School of Economics, The University of Adelaide, 10 Pulteney St, Adelaide, South Australia, 5005, Australia

[2]Centre for Applied Macroeconomic Analysis, Australian National University, Canberra, Australia

Correspondence
Firmin Doko Tchatoka, School of Economics, The University of Adelaide, 10 Pulteney St, Adelaide, SA 5005, Australia.
Email: firmin.dokotchatoka@adelaide.edu.au

Funding information
The University of Adelaide; Australian Research Council, Grant/Award Number: DP200101498

Abstract

The asymptotic distributions of the recursive out-of-sample forecast accuracy test statistics depend on stochastic integrals of Brownian motion when the models under comparison are nested. This often complicates their implementation in practice because the computation of their asymptotic critical values is burdensome. Hansen and Timmermann (2015, Econometrica) propose a Wald approximation of the commonly used recursive *F*-statistic and provide a simple characterization of the exact density of its asymptotic distribution. However, this characterization holds only when the larger model has one extra predictor or the forecast errors are homoscedastic. No such closed-form characterization is readily available when the nesting involves more than one predictor and heteroscedasticity or serial correlation is present. We first show through Monte Carlo experiments that both the recursive *F*-test and its Wald approximation have poor finite-sample properties, especially when the forecast horizon is greater than one and forecast errors exhibit serial correlation. We then propose a hybrid bootstrap method consisting of a moving block bootstrap and a residual-based bootstrap for both statistics and establish its validity. Simulations show that the hybrid bootstrap has good finite-sample performance, even in multi-step ahead forecasts with more than one predictor, and with heteroscedastic or autocorrelated forecast errors. The bootstrap method is illustrated on forecasting core inflation and GDP growth.

KEYWORDS
bootstrap consistency, moving block bootstrap, out-of-sample forecasts

1 | INTRODUCTION

Out-of-sample tests of predictive accuracy have received considerable attention in the literature.[1] Such testing procedures often involve comparing the out-of-sample mean squared forecast error (MSFE) of alternative models to select the one that minimizes this criterion. The case of nested models is particularly interesting because the test statistics often used—such as the recursively generated *F*-statistic [McCracken (2007) and Clark and McCracken (2001, 2005)]—have nonstandard asymptotic distributions that depend on stochastic integrals of Brownian motion; see Clark and McCracken (2012, 2014, 2015), and Hansen and Timmermann (2015). Many studies have developed methods for approximating the quantiles of the limiting distributions of these statistics, mainly by using simulation methods; see Rossi and Inoue (2012) and Hansen and Timmermann (2012). However, these simulation methods can be computationally burdensome, especially in the multivariate setting, because it requires a discretization of both the underlying (multivariate) Brownian motion and the support of the nuisance parameters (such as the relative size of the initial estimation sample versus the out-of-sample evaluation period).

Recently, Hansen and Timmermann (2015) show that the recursively generated F-statistic of McCracken (2007) can be approximated by a Wald-type statistic whose asymptotic distribution is a convolution of dependent $\chi^2(1)$-distributed random variables, thus simplifying the computation of test critical values. When the underlying data generating process (DGP) is homoscedastic, their characterization yields a closed-form expression of the exact density of the limiting distribution of the F-statistic, even when the number of extra predictors in the larger model is greater than one; see Hansen and Timmermann ((2015), Theorem 5). However, no closed-form characterization of the density of the limiting distribution of this statistic is available in the multivariate setting (i.e., when there are more than one extra predictors in the larger model) if the underlying DGP is heteroscedastic or serially correlated.

This paper contributes to this research area in two main ways. First, we show through Monte Carlo simulations that even for moderate sample sizes, both the recursively generated F-test of McCracken (2007) and its Wald approximation of Hansen and Timmermann (2015) are often oversized, especially when the forecast errors exhibit heteroscedasticity or serial correlation. The size distortions of both tests increase with the forecast horizon. For example, in a simple framework where there is only one extra predictor in the larger model, our simulations show that under serially correlated forecast errors the rejection frequencies under the null hypothesis of the F-test (at the 5% nominal level) can jump from 10.6% when $T=50$, 7.5% when $T=100$, and 6.2% when $T=200$ for *1-step ahead* forecasts to 27.6% when $T=50$, 17.4% when $T=100$, and 13.9% when $T=200$ for *4-step ahead* forecasts.

Second, we propose a hybrid bootstrap method consisting of a moving block bootstrap (henceforth MBB, which is nonparametric) and a residual-based bootstrap (which is parametric) for both the recursively generated F-test and its equivalent Wald statistic.[2] Our bootstrap method builds on earlier work by Corradi and Swanson (2007), but it differs from theirs in two important aspects. First, whereas (Corradi & Swanson, 2007) (henceforth CS) bootstrap is purely nonparametric in the sense that level data are re-sampled (pairs bootstrap), ours is semi-parametric and is based on resampling the residuals of the unrestricted regression that includes the extra predictors. Re-sampling the residuals is paramount to recovering an eventual pattern of serial correlation in the regression errors, which is not always the case with the *pairs bootstrap*. Second, CS establish the conditions on the block length under which their MBB is consistent, but there is no practical guidance on the choice of this block length in their study. Their Monte Carlo experiments [see Tables 2-3 in Corradi and Swanson (2007)] provide a clear evidence on the importance of choosing the block length that fits the data better, as the performance of the bootstrap CS test varies largely across alternative choices of block lengths. In this paper, we suggest a data-dependent approach to select the block length. Specifically, we propose setting it equal to the optimal lag length of the Newey and West's (1987) HAC estimator used in the expressions of the statistics. As the choice of the block length aims to capture the dependence structure of the data, we believe matching it to the optimal lag length of the HAC estimator is reasonable. However, we do not claim optimality of this choice, for example, in the sense of maximizing test power. Rather, we follow Andrews and Monahan (1992) and the recommendations of Newey and West (1994) to select the kernel bandwidth of the HAC estimator and then use it as the block length in our bootstrap DGP. This choice satisfies the conditions under which our bootstrap consistency is established, thus guaranteeing that type I error is controlled for. From this perspective, our bootstrap method can be viewed as complementary to Corradi and Swanson (2007).

We show that our proposed bootstrap is consistent under both the null hypothesis of equal forecast accuracy and the alternative hypothesis, irrespective of the forecast horizon and the underlying DGP exhibiting heteroscedasticity or serial correlation. The proof of our bootstrap is innovative and different from the one in CS. Indeed, due to nesting, the standard Gaussian approximation used in CS no longer holds, so one has to resort to the functional central limit theorem; see Davidson (1994). We present simulation evidence indicating that the bootstrap approximation performs well in small samples, even with heteroscedastic or serially correlated errors. These results are qualitatively the same across forecast horizons, confirming our theoretical findings. We illustrate our theoretical results with empirical applications which look at forecasting core inflation and GDP growth.

Important contributions on residual MBB are Efron ((1982), pp.35–36) and Fitzenberger (1998). However, their bootstrap schemes assume that the regressors are strictly exogenous, and are therefore kept fixed (not re-sampled) in the bootstrap algorithm. For weakly dependent time series with lagged dependent variables, as is the case in most applications of out-of-sample tests of equal forecast accuracy, this type of MBB will not have the desired size property. Recent contributions on bootstrapping out-of-sample tests of equal forecast accuracy include Clark and McCracken (2012, 2014, 2015), who suggest using the "fixed regressor wild bootstrap (FRWB)." This bootstrap algorithm maintains the usual assumption that forecast errors exhibit an $MA(\tau-1)$ structure for forecast horizons $\tau>1$.

Throughout this paper, convergence almost surely is symbolized by " $a.s.$," "$\xrightarrow{p}$" stands for convergence in probability, whereas "$\xrightarrow{d}$" means convergence in distribution. The usual stochastic orders of magnitude are denoted by $O_p(\cdot)$ and $o_p(\cdot)$. $\mathbb{P}$ denotes the relevant probability measure and $\mathbb{E}$ is the expectation operator under $\mathbb{P}$. The "$*$" on all these symbols and other variables (for example $\mathbb{P}^*$) indicates the bootstrap world. $o_p*(1)$-$\mathbb{P}$ denotes a term converging to zero in $\mathbb{P}^*$-probability, conditional on the sample, and for all samples except a subset with probability measure approaching zero, and $O_p*(1)$-$\mathbb{P}$ is for a term that is bounded in $\mathbb{P}^*$-probability, conditional on the sample, and for all samples except a subset with probability measure approaching zero. Similarly, $o_{a.s^*}(1)$ and $O_{a.s^*}(1)$ denote the terms that approach zero almost surely and the terms that are almost surely bounded, according to the probability law $\mathbb{P}^*$ and conditional on the sample. The notation I_q stands for the identity matrix of order q and $\|U\|$ denotes the usual Euclidean or Frobenius norm for a matrix U. Finally, $\sup_{\omega\in\Omega}|f(\omega)|$ is the supremum norm on the space of bounded continuous real functions with topological space Ω.

The remainder of the paper is organized as follows. Section 2 presents the setup, formulates the null hypothesis as well as the assumptions used, and summarizes briefly the asymptotic properties of the tests studied. Section 3 presents our proposed bootstrap method and proves the validity of the bootstrap. Section 4 presents Monte Carlo results on the finite-sample performance of our proposed bootstrap as well as the asymptotic tests and the FRWB of Clark and McCracken (2012, 2014, 2015). Section 5 applies our bootstrap test to forecasts of core inflation and real GDP growth for US data. Finally, Section 6 concludes.

2 | FRAMEWORK

We first introduce the setup and the testing problem of interest as well as their asymptotic properties.

2.1 | Setup

Let $\{Y_t:\ 1\le t\le T\}$ be a stochastic process defined on $(\Omega,\mathcal{B},\mathcal{F})$, where $\mathcal{B}$ is a σ-algebra on Ω, $\mathcal{F}$ is the class of distributions under consideration, and Y_t has support on a compact subset of $\mathbb{R}^p$ for some positive integer p. Consider the partition $Y_t:=(y_t,X'_{2t})'$, where $y_t:1\times1$ and $X_{2t}:k\times1$ $(k=p-1)$ may contains lags of y_t. By convention, we assume that a vector (or a matrix) does not appear in the model if its number of columns or rows is zero. For example, X_{2t} does not appear in the above partition of Y_t if $p=1$. Let $s=\max\{q,\tau\}+1$, where q denotes the maximum lag length of the variables in X_{2t} and $\tau\ge1$ is the forecast horizon of interest.

Consider the predictive regression model (see Hansen & Timmermann, 2015)

$$\begin{aligned} y_t &= X'_{2,t-\tau}\beta_2+\varepsilon_{2t} \\ &= X'_{1,t-\tau}\beta_{21}+\tilde{X}'_{2,t-\tau}\beta_{22}+\varepsilon_{2t},\ t=s,\ldots,T, \end{aligned} \tag{2.1}$$

where $X_{2t}=(X'_{1,t},\tilde{X}'_{2,t})'$: $X_{1,t}\in\mathbb{R}^{k_1}$ and $\tilde{X}_{2,t}\in\mathbb{R}^{k_2}$ $(k=k_1+k_2)$; $\beta_2=(\beta'_{21},\beta'_{22})'\in\mathbb{R}^k$: $\beta_{21}\in\mathbb{R}^{k_1}$ and $\beta_{22}\in\mathbb{R}^{k_2}$ are unknown, and ε_{2t} is an error term. We are interested in testing whether $\tilde{X}_{2t}$ has predictive power in forecasting y_t τ-periods ahead.

Several studies[3] show that this testing problem can be formulated as the comparison of mean squared error (MSE) of the forecast of $y_{t+\tau}$ generated using the unrestricted model (2.1) to the one resulting from the restricted regression

$$y_t=X'_{1,t-\tau}\beta_1+\varepsilon_{1t},\ t=s,\ldots,T. \tag{2.2}$$

In this setting, the null hypothesis of equal predictive performance takes the form

$$H_0:\ \mathbb{E}_F[(y_t-X'_{2,t-\tau}\beta_2^0)^2-(y_t-X'_{1,t-\tau}\beta_1^0)^2]=0 \tag{2.3}$$

for some $F\in\mathcal{F}$, where $\beta_j^0=\arg\min_{\beta_j}\ \mathbb{E}_F[(y_t-X'_{j,t-\tau}\beta_j)^2]$ denote the unknown true values of β_j $(j=1,2)$ in (2.1)–(2.2).

Common to most studies in this literature is the assumption that the resulting forecast errors exhibit an $MA(\tau-1)$ structure for $\tau>1$ (see, e.g., Clark & McCracken, 2012, 2015). In this paper, we emphasize the possibility that the τ-period-ahead forecast errors may be autocorrelated and the order of the autocorrelation is unknown. For example, in empirical applications using vector autoregression (VAR) framework, the true data generating process (DGP) in (2.1)–(2.2) may include a certain number of lags of the dependent variable y_t as predictors but a researcher may only use a smaller number of lags than the true one, thus leading to autocorrelated errors ε_{jt}. Allowing for autocorrelated forecast errors helps to reduce bias in the estimates that results from this lag misspecification.

The form of H_0 in (2.3) suggests building test statistics based on the MSE loss differential $\mathbb{E}_F[(y_t - X'_{2,t-\tau}\beta_2)^2 - (y_t - X'_{1,t-\tau}\beta_1)^2]$. This is usually done out-of-sample using a recursive estimation of the model parameters; see, for example, Diebold and Mariano (1995); West (1996); Clark and McCracken (2001). In this framework, the test statistics often suggested have limiting distributions that depend on stochastic integrals of Brownian motions, which makes the computation of their critical values cumbersome.

In this study, we focus particularly on the recursively generated F-test of Clark and McCracken (2001), whose critical values are easier to compute in some cases due to its equivalence to a Wald-type statistic (Hansen & Timmermann, 2015). In particular, when the DGP is homoscedastic, Hansen and Timmermann (2015) provide closed-form expressions of the exact density of the limiting distributions of this F-statistic, even when the unrestricted model (2.1) contains more than one extra predictors. However, in addition to its large-sample approximation nature, there is no such closed-form characterization when the underlying DGP is heteroscedastic and (2.1) includes multiple extra predictors. This setting is of great relevance in empirical work, thus providing a valid statistical procedure that accounts for it is arguably of interest to applied researchers.

To introduce the recursively generated F-statistic, suppose that P_T out-of-sample predictions are available, where the first is based on a parameter vector estimated using data from s to R_T, the second on a parameter vector estimated using data from s to R_T+1, and so on, and the last is based on a parameter vector estimated using data from s to $R_T + P_T - \tau \equiv T - \tau$. Let $\hat{y}_{t|t-\tau}(\hat{\beta}_{2,t-\tau}) := \hat{y}_{t|t-\tau} = X'_{2t-\tau}\hat{\beta}_{2t-\tau}$ denote the τ-step ahead forecast generated from model (2.1) and $\tilde{y}_{t|t-\tau}(\hat{\beta}_{1,t-\tau}) := \tilde{y}_{t|t-\tau} = X'_{1,t-\tau}\hat{\beta}_{1,t-\tau}$ be the one that results from model (2.2), where $\hat{\beta}_{j,t}$ $(j=1,2)$ are the recursive OLS estimators of β_j from (2.1)–(2.2), that is,

$$\hat{\beta}_{j,t} = \arg\min_{\beta_j} \frac{1}{t}\sum_{n=s}^{t}\left(y_n - X'_{j,n-\tau}\beta_j\right)^2, \quad R_T \le t \le T-\tau;\ j=1,2. \tag{2.4}$$

The recursively generated F-statistic for H_0 (see Hansen & Timmermann, 2015) takes the form

$$\mathcal{T}_T = \frac{1}{\hat{\sigma}^2_\varepsilon}\sum_{t=R_T+1}^{T}\left[\left(y_t - X'_{2,t-\tau}\hat{\beta}_{2,t-\tau}\right)^2 - \left(y_t - X'_{1,t-\tau}\hat{\beta}_{1,t-\tau}\right)^2\right], \tag{2.5}$$

where $\hat{\sigma}^2_\varepsilon$ is a consistent estimator of the variance of the unrestricted error in (2.1).[4]

Let $H_2 = \underset{T\to\infty}{p\lim}\left(\frac{1}{T}\sum_{t=s}^{T} X_{2,t-\tau}X'_{2,t-\tau}\right)$ (assuming that the limit exists and also X_2 includes a column vector of ones) be partitioned as

$$H_2 = \begin{bmatrix} H_1 & H'_{21} \\ H_{21} & \tilde{H}_2 \end{bmatrix};\ H_1 : k_1 \times k_1,$$
$$H_{21} : k_2 \times k_1,\ \tilde{H}_2 : k_2 \times k_2,$$

and define $\check{H}_2 = \tilde{H}_2 - H_{21}H_1^{-1}H'_{21}$, $Z_{t-\tau} = \tilde{X}_{2,t-\tau} - H_{21}H_1^{-1}X_{1,t-\tau}$. Also, let $\check{\Gamma}_n$ denote the n-th autocovariance (suppose for now that it exists) of the stochastic process $\{Z_{t-\tau}\varepsilon_{2t}\}$, that is,

$$\check{\Gamma}_n = \underset{T\to\infty}{p\lim}\frac{1}{T}\sum_{t=s}^{T} Z_{t-\tau}\varepsilon_{2t}\varepsilon'_{2,t-n}Z'_{t-\tau-n},$$

and define $\check{\Omega} = \sum_{n=-\tau+1}^{\tau-1}\check{\Gamma}_n$. Let $Z_{T,t-\tau}$ denote the residual from the multivariate regression of $\tilde{X}_{2,t-\tau}$ on $X_{1,t-\tau}$, that is,

$$Z_{T,t-\tau} = \tilde{X}_{2,t-\tau} - \sum_{t=s}^{T}\tilde{X}_{2,t-\tau}X'_{1,t-\tau}\left(\sum_{t=s}^{T}X_{1,t-\tau}X'_{1,t-\tau}\right)^{-1}X_{1,t-\tau}.$$

The Wald statistic for the null hypothesis $\beta_{22}=0$ in (2.1) is given by

$$\hat{S}_T = T\hat{\beta}'_{22\cdot T}\hat{V}_T^{-1}\hat{\beta}_{22\cdot T}, \tag{2.6}$$

where $\hat{\beta}_{22\cdot T} = \left(\sum_{t=s}^{T}Z_{T,t-\tau}Z'_{T,t-\tau}\right)^{-1}\sum_{t=s}^{T}Z_{T,t-\tau}y_t$ and $\hat{V}_T \equiv \hat{V}_T(\hat{\beta}_{22\cdot T})$ is a consistent estimator of the variance of $\lim_{T\to\infty} var(\sqrt{T}\hat{\beta}_{22\cdot T})$. Under homoscedastic errors, we can express $\hat{V}_T = \hat{\sigma}^2_\varepsilon\left(\sum_{t=s}^{T}Z_{T,t-\tau}Z'_{T,t-\tau}\right)^{-1}$ and $\hat{S}_T$ in (2.6) as

$$\begin{aligned}\hat{S}_T &= \tilde{S}_T/\hat{\sigma}^2_\varepsilon(T), \text{ where } \tilde{S}_T \equiv \tilde{S}(T) \\ &= \sum_{t=s}^{T}y_tZ'_{T,t-\tau}\left(\sum_{t=s}^{T}Z_{T,t-\tau}Z'_{T,t-\tau}\right)^{-1}\sum_{t=s}^{T}Z_{T,t-\tau}y_t,\end{aligned} \tag{2.7}$$

where $\hat{\sigma}^2_\varepsilon(T)$ is a consistent estimator of $\sigma^2_\varepsilon = var(\varepsilon_{2,t})$ based on the full sample. Similarly, let $\hat{S}_{R_T}$ denote the Wald statistic for $\beta_{22}=0$, computed using the first R_T observations in the sample. Hansen and Timmermann (2015) show that $\mathcal{T}_T$ in (2.5) is asymptotically equivalent to the difference between two Wald-type statistics, that is, $\mathcal{T}_T = \mathcal{W}_T + o_p(1)$, where

$$\mathcal{W}_T = \hat{S}_T - \hat{S}_{R_T} + \sigma_\varepsilon^{-2}\check{\kappa}\log(\rho), \tag{2.8}$$

with $\check{\kappa} = tr[\check{H}_2^{-1}\check{\Omega}]$ under H_0, $\check{H}_2$ and $\check{\Omega}$ are defined above. It is clear from (2.8) that $\mathcal{W}_T$ is related to the homoscedastic Wald statistics of the null hypothesis $\beta_{22} = 0$, even when the underlying true data generating process is heteroscedastic. This means that the recursive F-statistic $\mathcal{T}_T$ is not robust to heteroscedasticity, as $\mathcal{W}_T$ is not robust to heteroscedasticity and both are equivalent under H_0. However, the Wald formulation (2.8) is more interesting than the F-type one (2.5) because it allows to accommodate heteroscedasticity or serial correlation, for example, by using a HAC estimator of the covariance matrix of the estimator of β_{22} with the full and reduced samples. But, even if such a correction was implemented, computing asymptotic critical values of the resulting statistic will still involve discretization of the underlying multivariate Brownian motions when (2.1) includes multiple extra predictors (Hansen & Timmermann, 2015), thus making it cumbersome to compute test critical values. This provides a strong motivation for our bootstrap method that not only alleviates the limitations of the F-statistic $\mathcal{T}_T$ but also makes its implementation easier.

It is worth noting that in this paper, we consider the recursively generated F-test of McCracken (2007) and Clark and McCracken (2001, 2005) that Hansen and Timmermann (2015) showed is equivalent to a Wald-type statistic. However, Clark and West (2007) suggest that for nested models the appropriate test statistic to use is the "MSPE-adjusted" test statistic. This statistic is the difference between MSPE of the parsimonious model and the larger model plus an adjustment term. They show that the adjustment term is needed as the MSPE from the parsimonious model is expected to be smaller than that of the larger model, in turn because the larger model introduces noise into its forecasts by estimating parameters whose population values are zero. Clark and West (2007) show that standard normal inference for the raw unadjusted difference in MSPEs—what they call "MSPE-normal"—performs abysmally. Moreover, they show that the MSPE-adjusted test statistic is not asymptotically normal but rather has a non-standard distribution. However, using the standard normal critical values leads to a conservative test asymptotically (i.e., the actual size less than the nominal size as the sample increases); see Clark and McCracken (2001, 2005). They also show that simulation methods, such as bootstrap, work well for both MSPE-adjusted and MSPE-normal test statistics. Clark and McCracken (2005) also compare the MSE-F statistic to the MSPE-adjusted statistic (what they call "ENC-F") based on both asymptotic critical values from non-standard limiting distribution and bootstrap method. They find that the MSE-F test has better size property than the ENC-F test, with both the asymptotic and bootstrap critical values. More generally, the simulations indicate the ENC-F test is more over-sized than the MSE-F test.

2.2 | Notations and assumptions

Throughout the study, the following notations are used. For any $j \in \{1,2\}$, let

$$\begin{aligned} s_{jt}(\beta_j) &= X_{j,t-\tau}(y_t - X'_{j,t-\tau}\beta_j) \equiv \left(s_{j,p}(t)\right)_{1\le p\le k_j}, \\ h_{jt} &= X_{j,t}X'_{j,t} \equiv \left[h_{j,pl}(t)\right]_{1\le p,l\le k_j}, \quad H_{jt} = \frac{1}{t}\sum_{n=s}^{t} h_{j,n-\tau}, \end{aligned}$$

where $\left(s_{j,p}(t)\right)_{1\le p\le k_j}$ is a k_j-dimensional vector with elements $s_{j,p}(t)$ and $\left[h_{j,pl}(t)\right]_{1\le p,l\le k_j}$ is a $k_j \times k_j$ matrix with entries $h_{j,pl}(t)$.

Define $\beta_2 = (\beta_1', 0')'$ and consider the selection matrix $J = \left[I_{k_1\times k_1} \vdots 0_{k_1\times k_2}\right]'$ such that $J's_{2t}(\beta_2) = s_{1t}(\beta_1)$ and $J'h_{2t}J = h_{1t}$. Also, let

$$\begin{aligned} \tilde{s}_{2t}(\beta_2) &= \sigma_\varepsilon^{-1}\tilde{A}H_2^{-1/2}s_{2t}(\beta_2), \\ \tilde{A} &\in \mathbb{R}^{k_2\times k} : \tilde{A}'\tilde{A} = H_2^{1/2}\left(-JH_1^{-1}J' + H_2^{-1}\right)H_2^{1/2}, \end{aligned} \tag{2.9}$$

where $\sigma_\varepsilon^2 = var(\varepsilon_{2,t})$ and $H_j = E_F[h_{jt}]$ for all $j \in \{1,2\}$. Define $B(r) = [B_1(r), ..., B_{k_2}(r)]' \in \mathbb{R}^{k_2}$; the standard Brownian motion on $\mathbb{D}^{k_2}_{[0,1]}$, $\mathbb{D}^{k_2}_{[0,1]}$ is the space of Cadlag mappings from [0,1] to $\mathbb{R}^{k_2}$. For any positive definite real matrix $\Sigma : q\times q, B(\Sigma)$ stands for a q-dimensional Brownian motion with covariance matrix Σ (see, e.g., Davidson, 1994, Section 27.7). We now make the following assumptions on the model variables and parameters.

Assumption 1.

(*i*) $U_{jt} = \left[s_{jt}(\beta_j)', vec\left(h_{jt} - H_j\right)'\right]'$ is covariance stationary such that $\mathbb{E}_F(U_{jt}) = 0$ and $H_j \equiv \mathbb{E}_F(h_{jt})$ is positive definite for all t and j;

(*ii*) U_{jt} is $3(2+1/\psi)$-dominated[5] uniformly in β_j

for some $\psi > 0$ and all t,j;

(*iii*) U_{jt} is $L_{2+\delta}$-NED[6] on some sequence $\{V_{jt}\}$ uniformly in β_j of size $-2(1+\psi)$, where $\{V_{jt}\}$ is α-mixing of size $-2(2+\delta)(1+2\psi)$ for t,j and some $\delta > 0$, $\psi > 0$.

Assumption 2. There is a kernel function $K(\cdot)$ with bandwidth q_T+1 satisfying:

(*i*) $K(\cdot): \mathbb{R} \to [-1,1]$, $K(0)=0, K(x)=K(-x)\, \forall\, x \in \mathbb{R}$,

$$\int_{-\infty}^{+\infty} K(x)^2 dx < \infty, \int_{-\infty}^{+\infty} |K(x)| dx < \infty,$$

$K(\cdot)$ is continuous at 0 and at all but a number of other points in $\mathbb{R}, \sup_{x \geq 0} |K(x)| < \infty$;

(*ii*) as $T \to \infty, q_T \to \infty$ and $q_T / \sqrt[q]{T} \to 0$ for some $q \in [0,\infty)$ such that

$$\|f^{(q)}\| \in [0,\infty) \text{ where } f^{(q)} = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |j|^q \mathbb{E}_F[X_{2,t-j} X_{2,t}'];$$

(*iii*) $\int_0^{+\infty} \overline{K}(x) dx < \infty$ where $\overline{K}(x) = \sup_{y \geq x} |K(y)|$.

Assumption 3.

(**i**) $T = R_T + P_T$ and $R_T = \lfloor \rho T \rfloor$ for some $\rho \in (0,1)$;

(**ii**) $P_T / R_T \to \pi = (1-\rho)/\rho$ as $T \to \infty$.

Remarks. **1.** The covariance stationary condition of both the score vector $s_{jt}(\beta_j)$ and the Hessian matrix h_{jt} in Assumption 1-(i) is standard in the literature (see, e.g., Clark & McCracken, 2012). Assumption 1-(ii) requires the existence of at least the first six moments for both $s_{jt}(\beta_j)$ and h_{jt}. In addition, the NED and strong α-mixing conditions in Assumption 1-(iii) ensures that the process $\{y_t, X_{2t}'\}$ is ergodic in both the mean and covariance, whereby the central limit theorem can be applied. Unlike previous studies,[7] Assumption 1 allows for correlation of order more than $\tau - 1$. The usual assumption from previous studies that $s_{jt}(\beta_j)$ and $s_{j,t-h}(\beta_j)$ are uncorrelated for order more than $\tau - 1$ is mainly justified by the fact that the τ-period-ahead forecast errors exhibit an $MA(\tau-1)$ structure. Therefore, the correlation between $s_{jt}(\beta_j)$ and $s_{j,t-h}(\beta_j)$ vanishes for $h \geq \tau$. However, as we emphasize on cases where the τ-period-ahead forecast errors may be autocorrelated, the usual assumption cannot be sustained.

2. Conditions (i) and (ii) of Assumption 2 ensure the consistency of the HAC estimator with a rate derived in Andrews (1991). Under these conditions, the bandwidth parameter q_T satisfies

$$\limsup_{T\to\infty} \sup_{0<\nu<\nu_u} (q_T+1)^{-1} \sum_{n=1}^{T-1} \left| K\left(\frac{n}{\nu(q_T+1)}\right) \right| < \infty \quad (2.10)$$

for any $0 < \nu_u < \infty$ (see, e.g., Jansson, 2002, Lemma 1). Note that (2.10) holds for all the kernels in class $\mathcal{K}_3$ of Andrews (1991, eq. (7.1)) and those satisfying Assumptions 1 and 3 in Newey and West (1994).

3. Assumption 3 is frequently used in the literature of tests of predictive accuracy. It implies that $0 < \pi < \infty$, that is, R_T and P_T grow at the same rate as T increases. It can be extended to $\pi = 0$, that is, P_T grows at a lower rate than R_T. Inference in this case is straightforward as it yields pivotal statistics (see McCracken, 2007, Theorem 3.2-(b)), meaning that our bootstrap method will yield a higher-order refinement in that case.

2.3 | Asymptotic distributions

Under Assumptions 1–3, Hansen and Timmermann (2015) provide a characterization of the asymptotic distribution of $\mathcal{T}_T$ under the null hypothesis $\beta_{22} = 0$ and local alternatives of the form $\beta_{22} = cT^{-1/2} b$ for some constant scalar c and vector b. More precisely, they show that

(a) if $\beta_{22} = 0$, then

$$\mathcal{T}_T \xrightarrow{d} \sum_{l=1}^{k_2} \lambda_l \left[2\int_\rho^1 r^{-1} B_l(r) dB_l(r) - \int_\rho^1 r^{-2} B_l^2(r) d(r) \right] \tag{2.11}$$

$$\stackrel{d}{\equiv} \sum_{l=1}^{k_2} \lambda_l \left[B_l^2(1) - \rho^{-1} B_l^2(\rho) + log(\rho) \right]; \tag{2.12}$$

(b) and if $\beta_{22} = cT^{-1/2}b$ for some c and b (b is such that $b'\check{\Sigma}b = \sigma_\varepsilon^2 \kappa$), then

$$\mathcal{T}_T \xrightarrow{d} \sum_{l=1}^{k_2} \lambda_l \left[2\int_\rho^1 r^{-1} B_l(r) dB_l(r) - \int_\rho^1 r^{-2} B_l^2(r) d(r) + (1-\rho)c^2 + 2ca_l[B_p(1) - B_l(\rho)] \right]$$

$$\stackrel{d}{\equiv} \sum_{l=1}^{k_2} \lambda_l \left[B_l^2(1) - \rho^{-1} B_l^2(\rho) + log(\rho) + (1-\rho)c^2 + 2ca_l[B_l(1) - B_l(\rho)] \right]; \tag{2.13}$$

where $a = b'\check{H}_2 \check{\Omega}_\infty^{-1/2} Q' \equiv (a_l)_{1 \le l \le k_2}$, Q is an orthogonal matrix such that $Q'Q = I_{k_2}$ and $Q'\Lambda Q = \sigma_\varepsilon^{-2} \check{\Omega}_\infty^{1/2} \check{H}_2^{-1} \check{\Omega}_\infty^{1/2}$, $\Lambda = diag(\lambda_1, \ldots, \lambda_{k_2})$, and $\check{\Omega}_\infty$ is the asymptotic variance of the stochastic process $\{Z_{t-\tau}\varepsilon_{2t}\}$ where $Z_{t-\tau} = \tilde{X}_{2,t-\tau} - H_{21}H_1^{-1}X_{1,t-\tau}$.

Remarks. Several observations are of order.

1. The expression of the limiting distribution of $\mathcal{T}_T$ in (2.11) is well known in the literature (see, e.g., McCracken, 2007) and the difficulties in computing test critical values using this formula are well-documented. Hansen and Timmermann (2015) show that this integral of stochastic Brownian motions can be expressed as a convolution of dependent $\chi^2(1)$ variables, as shown in (2.12).
2. In (2.12), the eigenvalues λ_l $(l = 1, \ldots k_2)$ measure the degree of heteroscedasticity in the model (Hansen & Timmermann, 2015). Under homoscedasticity, $\lambda_l = 1$ for all l, and (2.11) reduces to the earlier result in McCracken (2007). However, under heteroscedasticity $\lambda_l \neq 1$ for some $l = 1, \ldots, k_2$. Clearly, (2.12) illustrates that $\mathcal{T}_T$ (thus $\mathcal{W}_T$) is not robust to heteroscedasticity.
3. If $B(r)$ is a univariate standard Brownian motion (i.e., if (2.1) contains only one extra predictor), the limiting distribution in (2.12) is identical to that of the random variable $\sqrt{1-\rho}(\mathbf{Z}_1^2 - \mathbf{Z}_2^2) + log(\rho)$, where $\mathbf{Z}_j \stackrel{i.i.d.}{\sim} N(0,1), j = 1,2$ (Hansen & Timmermann, 2015, Theorem 4). Hence, test critical values can be simulated easily given ρ. This case is somewhat restrictive as it implies that $\tilde{X}_2$ in (2.1) contains only one regressor (i.e., $k_2 = 1$). Similarly, if $B(r)$ is a multivariate standard Brownian motion (i.e., if $\tilde{X}_2$ contains more than one regressor) and the DGP is homoscedastic (i.e., $\lambda_1 = \ldots = \lambda_{k_2} = 1$), a closed-form expression of the pdf of the exact density of the asymptotic limit variable in (2.12) is provided in Hansen and Timmermann ((2015), theorem 5). As such, the asymptotic critical values of $\mathcal{T}_T$ can be simulated under H_0 using this pdf formula. However, we are not aware of a closed-form characterization of the pdf of the limit variable in (2.12) under heteroscedasticity or serial correlation in the multivariate nested setting (i.e., $k_2 > 1$). Moreover, the equivalence between $\mathcal{T}_T$ and $\mathcal{W}_T$ is asymptotic in nature, and as such the finite-sample behavior of both statistics may differ even in the case where the DGP is homoskedastic, thus leading to size distortions in small samples when the simulated critical values from (2.12) are used. Our bootstrap procedure not only simplifies the computational burdens of simulating critical values of $\mathcal{T}_T$ and $\mathcal{W}_T$, especially when $k_2 > 1$ and the DGP is heteroskedastic, but also provides a framework to improve the finite-sample performance of the tests.

3 | BOOTSTRAP TEST

In the multistep-ahead forecasting framework (i.e., $\tau > 1$), bootstrap methods often fail to control the size even for a well-specified homoscedastic model due to lack of accounting for the serial correlation structure of the resulting forecast errors. Building on earlier work by Corradi and Swanson (2007), we propose a bootstrap method that performs quite well for both $\mathcal{T}_T$ and $\mathcal{W}_T$ irrespective of the forecast horizon τ.

3.1 | Bootstrap DGP

Let $\hat{\varepsilon}_{2,t}$ denote the residuals from the OLS of (2.1), and define $\{W_t = (X_{2,t}^\dagger, \hat{\varepsilon}_{2,t}) : t = s, \ldots, T\}$ where $X_{2,t}^\dagger$ contains the variables in $X_{2,t}$ other than the lags of the dependent variable y_t. Let $\ell_T \in \mathbb{N}$ be a block length $(1 \le \ell_T \le T-s)$, $B_{t,\ell_T} = \{W_t, W_{t+1}, ..., W_{t+\ell_T-1}\}$ be the block of ℓ_T consecutive observations starting at W_t for $t = s, ..., T$. Assume that $T - s = b_T \ell_T$ so that the moving block bootstrap (MBB) procedure consists of drawing $b_T = (T-s)/\ell_T$ blocks, $\{B^*_{1,\ell_T}, B^*_{2,\ell_T}, ..., B^*_{b_T,\ell_T}\}$, randomly with replacement from the set of overlapping blocks $\{B_{s,\ell_T}, ..., B_{T-\ell_T+1,\ell_T}\}$. The first ℓ_T observations in the pseudo-time series are the sequence of ℓ_T values in B^*_{s,ℓ_T}, the next ℓ_T observations in the pseudo-time series are the ℓ_T values in B^*_{s+1,ℓ_T}, and so on, that is, $W^* \equiv \left(W_s^{*'}, W_{s+1}^{*'}, ..., W_T^{*'}\right)' = \left(B_{1,\ell_T}^{*'}, B_{2,\ell_T}^{*'}, ..., B_{b_T,\ell_T}^{*'}\right)'$, where $W_t^{*'} := \left[X_{2,t}^{\dagger *'}, \varepsilon^*_{2,t}\right]$ for all $t = s, ...T$. Let $\mathbf{I}_1, ..., \mathbf{I}_{b_T}$ be i.i.d. random variables distributed uniformly on $\{s-1, s, ..., T-\ell_T+1\}$. The resulting bootstrap sample can be defined as $\{W_t^* := W_{\tau_t}, t = s, ..., T\}$ where τ_t is a random array, that is, $\{\tau_t\} := \{\mathbf{I}_1+1, ..., \mathbf{I}_1+\ell_T, ..., \mathbf{I}_{b_T}+1, ..., \mathbf{I}_{b_T}+\ell_T\}$. To construct the bootstrap dependent variable y_t^*, we proceed as follows.

1. If $X_{2,t}$ does not contain any lag of y_t, then set $X_{2,t}^{\dagger *'} := X_{2,t}^{*'} = \left[X_{1,t}^{*'}, \tilde{X}_{2,t}^{*'}\right]$ and generate y_t^* as:

$$y_t^* = X_{1,t-\tau}^{*'} \hat{\beta}_{1,T-\tau} + \varepsilon^*_{2,t}, t = s, s+1, ..., T. \tag{3.1}$$

2. If $X_{2,t}$ contains lags of y_t, then proceed as follows. First, set $y_t^* = y_t$ for all $t = 1, ..., s-1$ (initial values) and form $y^*_{t-\tau}$ for $t = s, ..., s+\tau-1$,[8] and partition the bootstrap draws $X_{2,t}^{\dagger *'}$ appropriately to form $X_{1,t}^{*'}$. Then, compute y_t^* for $t = s+\tau, ..., T$ as:

$$y_t^* = X_{1,t-\tau}^{*'} \hat{\beta}_{1,T-\tau} + \varepsilon^*_{2,t}, t = s+\tau, ..., T. \tag{3.2}$$

Remarks. **1.** The above bootstrap scheme is a hybrid of moving block bootstrap and residual-based bootstrap. As such, it differs from the previous literature on the topic in a number of ways. In particular, it differs from Corradi and Swanson (2007) in three important aspects. - *First*, whereas the bootstrap of Corradi and Swanson (2007) is non-parametric in the sense that level data are re-sampled (pairs bootstrap) in their DGP, ours is semi-parametric in the sense that it is based on resampling the residuals. Resampling the residuals is important for recovering the serial correlation pattern of regression errors. For example, the Monte Carlo experiment in Corradi and Swanson (2007) illustrates that *the pairs* bootstrap does not always mimic well, not only the serial correlation pattern in regression errors but also the persistence of the data, thus leading to valid but conservative bootstrap tests. Looking at Corradi and Swanson ((2007), tables 2 and 3)], we see that with moderate autocorrelation of the errors in their DGP ($a_3 = 0.3$) and mild persistence in the data ($a_2 = 0.6$), the *size2* results indicate a conservative moving block bootstrap procedure at the 10% nominal level for 1-step-ahead forecasts. The under-rejections of this moving block bootstrap worsens as the persistence in the data increases (see Corradi & Swanson, 2007, tables 2 and 3, panel C). *Second*, to ensure that our bootstrap provides asymptotically valid estimates of the appropriate critical values regardless of whether the null hypothesis holds, we generate the bootstrap samples using the residuals from the unrestricted model (similar to Clark & McCracken, 2012), and then form the bootstrap DGP as $y_t^* = X_{1,t-\tau}^{*'} \hat{\beta}_{1,T-\tau} + \varepsilon^*_{2,t}$. *Third*, Corradi and Swanson (2007) establish the conditions on the block length ℓ_T under which the MBB is consistent. We establish similar conditions but also suggest a data-dependent method to select the bootstrap block length. As the choice of the block length should capture the structure of dependence in the data, we believe equalizing it to the optimal lag length of the HAC estimator for the variance of the errors is a reasonable choice. We are not aware of any study on this topic that simultaneously addresses the problem of autocorrelation and the selection choice of the bootstrap block length in a data-dependent manner. For example, Corradi and Swanson (2007, tables 2 and 3) show clear evidence that the choice of the bootstrap block length influences the performance of the bootstrap CS test. Tables S4 and S5 and the discussion therein highlight the impact of the block length choice on the bootstrap test properties. In particular, the bootstrap test can be under-sized for a choice of the block length that exceeds the true one, and over-sized for a choice of the block length that is shorter than the true one. From this perspective, our bootstrap method is an important contribution to the literature.

2. An important contribution on residual MBB is Efron (1982, pp. 35–36), but its scheme considers the regressors to be strictly exogenous, thus those regressors are not re-sampled in the bootstrap algorithm. This type of MBB is not appropriate for weakly dependent time series with lagged dependent variables. Fitzenberger (1998) proposes a MBB where the regressors are re-sampled, but as in Corradi and Swanson (2007), the choice of block length is not addressed. Recent contributions on bootstrapping out-of-sample tests of predictive accuracy include Clark and McCracken (2012, 2014, 2015). Their bootstrap algorithm relies on a variant of the wild bootstrap that maintains the $MA(\tau-1)$ structure of the forecast errors.
3. Although our framework does not directly address the multiple hypothesis testing problems analyzed in Clark and McCracken (2012), our bootstrap method can be generalized to this framework. We leave this extension for future research.

Let $\hat{\beta}^*_{j,t}$ be the recursive bootstrap estimator similar to $\hat{\beta}_{j,t}$ in (2.4), that is,

$$\hat{\beta}^*_{j,t} = \arg\min_{\beta_j} \frac{1}{t}\sum_{n=s}^{t}\left(y^*_n - X^{*\prime}_{j,n-\tau}\beta_j\right)^2,\ R_T \le t \le T-\tau; j=1,2. \tag{3.3}$$

Lemma A.3 in Appendix A establishes that

$$\mathbb{E}_{F^*}\left[\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}\left(\hat{\beta}^*_{j,t} - \hat{\beta}_{j,t}\right)\right] = O_{p^*}(1)\ \text{pr} - \mathbb{P}\ \text{for all}\ j = 1,2.$$

That is, the limiting distribution of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\hat{\beta}^*_{j,t} - \hat{\beta}_{j,t})$ is not centered at zero but is rather characterized by a location bias. This means that a bootstrap test based on $\hat{\beta}^*_{j,t}$ may not have desirable size properties and some adjustments are required. Several studies, including Politis and Romano (1994) and Corradi and Swanson (2007), have proposed methods to eliminate this location bias from $\hat{\beta}^*_{j,t}$. Due to its simplicity, we adapt the approach of Corradi and Swanson (2007).

Define the adjusted recursive estimator

$$\tilde{\beta}^*_{j,t} = \arg\min_{\beta_j} \frac{1}{t}\sum_{n=s}^{t}\left[\left(y^*_n - X^{*\prime}_{j,n-\tau}\beta_j\right)^2 + 2\beta_j'\left(\mu_T\sum_{n=s}^{T-\tau}s_{j,n}(\hat{\beta}_{j,t})\right)\right],$$
$$R_T \le t \le T-\tau, \tag{3.4}$$

where $\mu_T = 1/(T-\tau-s+1)$ and $s_{j,n}(\hat{\beta}_{j,t}) = X_{j,n-\tau}(y_n - X'_{j,n-\tau}\hat{\beta}_{j,t})$ for all $j=1,2$. We can solve (3.4) explicitly for $\tilde{\beta}^*_{j,t}$ to get

$$\begin{aligned}\tilde{\beta}^*_{j,t} &= \left(\frac{1}{t}\sum\nolimits_{n=s}^{t}h^*_{j,n-\tau}\right)^{-1} \\ &\times\left(\frac{1}{t}\sum\nolimits_{n=s}^{t}\left[X^*_{j,n-\tau}y^*_n - \mu_T\sum\nolimits_{n=s}^{T-\tau}s_{j,n}(\hat{\beta}_{j,t})\right]\right), \\ &R_T \le t \le T-\tau,\end{aligned} \tag{3.5}$$

where $h^*_{j,n-\tau} = X^*_{j,n-\tau}X^{*\prime}_{j,n-\tau}$ is the bootstrap analog of $h_{j,n-\tau} = X_{j,n-\tau}X'_{j,n-\tau}$.

Before moving on to the bootstrap tests, we first state the following result on the asymptotic behavior of $\tilde{\beta}^*_{j,t}$ in (3.5)

Theorem 3.1. *Suppose Assumptions 1–3 are satisfied and $\ell_T = o\left(T^{\frac{1}{4}}\right)$. Then*

$$\lim_{T\to\infty}\mathbb{P}\left[\omega:\ \sup_{v_j\in\mathbb{R}^{k_j}}\left|\mathbb{P}^*\left(\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}^*_{j,t} - \hat{\beta}_{j,t}) \le v_j\right) - \mathbb{P}\left(\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\hat{\beta}_{j,t} - \beta^0_j) \le v_j\right)\right| > \zeta\right] = 0$$

for any $\zeta > 0$.

Remark 4. Theorem 3.1 establishes the consistency of the limiting distribution of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{j,t}^{*}-\hat{\beta}_{j,t})$ to that of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\hat{\beta}_{j,t}-\beta_j^0)$ for all $j=1,2$. It is similar to Corradi and Swanson ((2007), theorem 1); however, its proof differs from theirs mainly because the asymptotic distribution of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\hat{\beta}_{j,t}-\beta_j^0)$ is a mixture of Brownian motions (see Lemma A.2(b) in Appendix A), rather than a standard Gaussian process as in Corradi and Swanson (2007). Under the conditions of Theorem 3.1, the Mann and Wald's (1943) theorem implies that any bootstrap statistic that is a continuous function of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{j,t}^{*}-\hat{\beta}_{j,t})$ should control test size. This result is important because it implies that our proposed hybrid bootstrap statistics in Section 3.2 provide a good approximation of the limiting distribution of the standard test statistics $\mathcal{T}_T$ and $\mathcal{W}_T$ without the usual re-centering as in Corradi and Swanson (2007).

Section 3.2 presents our bootstrap statistics and characterizes their asymptotic behavior under both the null hypothesis (size) and the alternative hypothesis (power).

3.2 | Bootstrap statistics

We suggest the following recursive bootstrap F-statistic and its equivalent Wald formulation:

$$\mathcal{T}_T^{*}=\frac{1}{\hat{\sigma}_{\varepsilon}^{*2}}\sum_{t=R_T+1}^{T}\left[\left(y_t^{*}-X_{2,t-\tau}^{*'}\tilde{\beta}_{2,t-\tau}^{*}\right)^2-\left(y_t^{*}-X_{1,t-\tau}^{*'}\tilde{\beta}_{1,t-\tau}^{*}\right)^2\right] \tag{3.6}$$

$$\mathcal{W}_T^{*}=\hat{S}_T^{*}-\hat{S}_{R_T}^{*}+\hat{\sigma}_{\varepsilon}^{*-2}\check{\kappa}^{*}\log(\rho), \tag{3.7}$$

where $\tilde{\beta}_{j,t}^{*}$ is given in (3.5), $\hat{S}_m^{*}=\tilde{S}_m^{*}/\hat{\sigma}_{\varepsilon}^{*2}$ is the bootstrap counterpart of $\hat{S}_m$ for $m\in\{T,R_T\}$ with $\tilde{S}_m^{*}\equiv\tilde{S}^{*}(m)=\sum_{t=s}^{m}y_t^{*}Z_{m,t-\tau}^{*'}\left(\sum_{t=s}^{m}Z_{m,t-\tau}^{*}Z_{m,t-\tau}^{*'}\right)^{-1}\sum_{t=s}^{m}Z_{m,t-\tau}^{*}y_t^{*}$, and $\hat{\sigma}_{\varepsilon}^{*2}$ and $\check{\kappa}^{*}$ are the bootstrap counterparts of $\hat{\sigma}_{\varepsilon}^{2}$ and $\check{\kappa}$, respectively.[9] We implement the bootstrap test using the following algorithm:

1. Given the observed data, construct an estimate of the test statistics $\mathcal{T}_T$ and $\mathcal{W}_T$ defined in (2.5)-(2.8).
2. Construct $j=1,\ldots,N$ bootstrap pseudo-samples independently as described in Section 3.1, and compute the bootstrap statistics $\mathcal{T}_T^{*}$ and $\mathcal{W}_T^{*}$ as in (3.6)–(3.7).
3. Reject the null hypothesis at the α% level if the test statistic ($\mathcal{T}_T$ or $\mathcal{W}_T$) is greater than the $(100-\alpha)$ percentile of the empirical distribution of the N simulated test statistics.

We can now state the following theorem on the validity of the bootstrap:

Theorem 3.2. *Suppose Assumptions 1–3 are satisfied and* $\ell_T=o\left(T^{\frac{1}{4}}\right)$. *Then:*

$$\lim_{T\to\infty}\mathbb{P}\left[\omega:\sup_{v_j\in\mathbb{R}^{k_j}}\left|\mathbb{P}^{*}(\mathcal{T}_T^{*}\le v_j)-\mathbb{P}(\mathcal{T}_T\le v_j)\right|>\zeta\right]=0,$$

$$\lim_{T\to\infty}\mathbb{P}\left[\omega:\sup_{v_j\in\mathbb{R}^{k_j}}\left|\mathbb{P}^{*}(\mathcal{W}_T^{*}\le v_j)-\mathbb{P}(\mathcal{W}_T\le v_j)\right|>\eta\right]=0$$

for any $\zeta>0$ *and* $\eta>0$.

Remark 5. Theorem 3.2 holds irrespective of whether the null hypothesis H_0 is satisfied or not. Moreover, our bootstrap approximates well the limiting distribution of the standard test statistics $\mathcal{T}_T$ and $\mathcal{W}_T$ even if the data generating process is heteroscedastic or weakly dependent. This contrasts with the *fixed regressor* bootstrap for which only the $MA(\tau-1)$ structure of forecast errors are accounted for (see Clark & McCracken, 2012, 2015). In particular, Theorem 3.2 allows for a more general ARMA structure for the forecast errors, which as discussed previously, can of be great importance in applied work when $\tau>1$.

Now, let $c_{\mathcal{T}}^{*}(\alpha)$ and $c_{\mathcal{W}}^{*}(\alpha)$ denote the $(1-\alpha)$ quantiles under H_0 of the bootstrap statistics $\mathcal{T}_T^{*}$ and $\mathcal{W}_T^{*}$ respectively for some $\alpha\in(0,1)$. Theorem 3.3 characterizes the behavior of the statistics $\mathcal{T}_T$ and $\mathcal{W}_T$ under the alternative hypothesis $H_1:\beta_{22}\neq 0$ when the bootstrap critical values $c_{\mathcal{T}}^{*}(\alpha)$ and $c_{\mathcal{W}}^{*}(\alpha)$ are utilized.

Theorem 3.3. *Suppose Assumptions 1–3 are satisfied,* $\ell_T=o\left(T^{\frac{1}{4}}\right)$ *and* $\beta_{22}\neq 0$ *is fixed. Then:*

$$\mathbb{P}\left[\mathcal{T}_T>c_{\mathcal{T}}^{*}(\alpha)\right]\to 1,\quad \mathbb{P}\left[\mathcal{W}_T>c_{\mathcal{W}}^{*}(\alpha)\right]\to 1\quad\text{as } T\to\infty.$$

Note that although Theorem 3.3 only focuses on fixed alternative hypotheses, there is no impediment to extending it to local alternative hypotheses of the form given in (2.13).

Next, we investigate the finite-sample performance of both the asymptotic and proposed bootstrap through Monte Carlo experiments in Section 4.

4 | FINITE-SAMPLE PERFORMANCE OF THE TESTS

We consider the following DGP described by a bi-variate vector autoregression (VAR) (similar to Hansen & Timmermann, 2015):

$$y_t = 0.3y_{t-\tau} + \beta_{22}x_{t-\tau} + u_{yt}, \tag{4.1}$$

$$x_t = 0.5x_{t-\tau} + u_{xt}, \tag{4.2}$$

where u_{yt}, u_{xt} are the error terms and τ is the forecast horizon. To simplify the presentation, the experiments are run with $\tau \in \{1,4\}$, but the results are qualitatively the same for alternative forecast horizons. In all experiments, the null forecast model is of "*no-influence*" of $x_{t-\tau}$ in (4.1), and the alternative (unrestricted) forecast model takes the form of (4.1)–(4.2) with $\beta_{22} \neq 0$. Therefore, the testing problem of interest can be formulated equivalently as

$$H_0 : \beta_{22} = 0 \ vs. \ H_1 : \beta_{22} \neq 0. \tag{4.3}$$

We consider four DGPs for the error vector $(u_{yt}, u_{xt})'$ as follows.

DGP 1 (Homoscedasticity): $(u_{yt}, u_{xt})' \overset{i.i.d.}{\sim} N(0, I_2)$;

DGP 2 (Heteroscedasticity alone): $u_{jt} \sim N(0, h_{jt})$ for $j \in \{x,y\}$, where $h_{jt} = \alpha_0 + \alpha_1 u_{jt-1}^2 + \alpha_2 h_{jt-1}$ and $\alpha_1 = 0.1, \alpha_2 = 0.8$;

DGP 3 (Autocorrelation only): $u_{jt} = 0.5u_{j,t-1} + \varepsilon_{jt}$ for $j \in \{x,y\}$, where $(\varepsilon_{yt}, \varepsilon_{xt})' \overset{i.i.d.}{\sim} N(0, I_2)$ when $\tau = 1$, whereas

$$u_{jt} = 0.5u_{j,t-1} + \varepsilon_{jt} + 0.95\varepsilon_{jt-1} + 0.90\varepsilon_{jt-2} + 0.80\varepsilon_{jt-3},$$

where $(\varepsilon_{yt}, \varepsilon_{xt})' \overset{i.i.d.}{\sim} N(0, I_2)$ when $\tau = 4$[10];

DGP 4 (Heteroscedasticity and autocorrelation): DGP 2 + DGP 3.

The simulations are run with $N =$ 10,000 simulated samples and $B = 199$ bootstrap pseudo samples of sizes $T \in \{50, 100, 200\}$. For both the asymptotic and bootstrap tests, the nominal level α is set at 5%.[11] The sample split points, $\pi = (1-\rho)/\rho$, varies in $\{0.2, 0.8, 1.4, 2.0\}$. For example, if $\pi = 0.2$, then $\rho = \frac{5}{6}$ that is, $\lfloor \frac{5T}{6} \rfloor$ observations are used in the initial estimation. This allows us to compare our results with previous studies (see, e.g., Clark & McCracken, 2001, 2005; McCracken, 2007).

4.1 | Size results

Table 1 shows the rejection frequencies for the standard asymptotic test and our proposed moving block bootstrap test for 1-step-ahead ($\tau = 1$) and 4-step-ahead ($\tau = 4$) forecasts for sample sizes $T \in \{50, 200\}$ at nominal 5% level.[12] The first column of the table presents the fraction π of the sample used in the initial estimation period. The other columns present, for each DGP, the rejection frequencies of the tests under H_0 at the nominal 5% level.

Consider first the rejections of the standard tests with usual asymptotic critical values (see parts of Table 1 labelled "*Asymptotic*"). We see that when $\tau = 1$ both tests are oversized for small sample sizes. In particular, when $T = 50$, the over-rejections under DGP 3 (autocorrelation alone) and DGP 4 (heteroscedasticity and autocorrelation) are large. For example, the maximal rejection frequencies under DGP 4 for the F-test can be as high as twice the nominal level (i.e., 10.3%), whereas that of the Wald test is even worse (13.7%). More precisely, the rejection frequencies of the recursive F-test in DGP 4 range from 7.5% to 10.3% and that of the Wald test range from 12.1% to 13.7%. Similar results are observed in DGP 3 (autocorrelation alone). As shown in Appendix S1, the size distortions persist for both tests in DGP 3 and DGP 4 for $T = 100$. However, the tests show better size when $\tau = 1$ and $T = 200$. Looking at the case when $\tau = 4$, we observe that both tests over-reject the null hypothesis substantially, and their size distortions persist in DGP 3 and DGP 4 even when $T = 200$. In particular, the empirical rejection frequency of $\mathcal{T}_T$ can be as large as 27.6% when $T = 50$, 17.6% when $T = 100$, and 13.9% when $T = 200$. Similar results are seen for $\mathcal{W}_T$ as well but its rejection frequencies under DGP 3 and DGP 4 are slightly less than that of $\mathcal{T}_T$.

Consider next the bootstrap tests' rejections (parts of Table 1 labelled "*Moving Block Bootstrap*"). To enable comparison with recent bootstrap methods, we also report the fixed regressor wild bootstrap MSE-F test results (Clark & McCracken, 2012; 2015). To shorten the presentation, the rejection frequencies for the fixed regressor MSE-F test are reported in Table S2. We can see that the moving block bootstrap performs well for both $\mathcal{T}_T$ and $\mathcal{W}_T$, whereas the fixed regressor MSE-F bootstrap tends to over-reject, especially in DGPs 3 and 4.[13]

TABLE 1 Rejection frequencies with sample sizes $T = 50$ and 200, $\alpha = 5\%$.

$\tau = 1$									$\tau = 4$							
	DGP 1		DGP 2		DGP 3		DGP 4		DGP1		DGP 2		DGP 3		DGP 4	
π	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald
$T = 50$																
Asymptotic																
0.2	7.9	8.6	7.1	7.9	10.6	12.0	10.3	12.7	10.8	12.1	10.5	11.7	21.5	21.0	21.8	21.1
0.8	8.0	8.7	7.4	8.8	9.7	13.2	10.0	13.7	10.0	12.0	9.4	11.6	25.2	19.6	25.5	19.8
1.4	6.8	8.8	6.7	8.0	8.7	12.1	8.9	12.1	8.7	11.0	9.0	11.5	25.3	17.0	26.0	17.4
2.0	7.0	8.6	6.5	8.5	7.9	11.5	7.5	12.2	8.5	10.4	7.5	10.4	27.6	15.6	27.1	15.2
Moving Block Bootstrap																
0.2	4.5	4.4	4.7	4.1	5.7	5.3	4.9	5.0	4.6	4.0	3.8	3.4	6.9	6.5	8.2	7.3
0.8	4.7	3.7	4.5	3.8	5.2	5.6	5.9	5.0	5.1	3.4	4.3	3.1	7.6	7.6	9.0	8.0
1.4	5.3	4.0	5.4	4.9	6.0	5.5	4.6	5.5	5.3	3.8	3.6	3.6	9.9	7.2	9.1	7.8
2.0	4.7	4.5	4.1	3.8	5.6	5.1	5.1	5.4	3.9	4.0	4.2	3.9	10.8	7.0	10.2	6.8
$T = 200$																
Asymptotic																
0.2	5.8	5.9	5.3	5.6	6.0	6.6	5.9	6.4	6.4	6.8	6.3	6.9	10.2	9.4	10.5	9.9
0.8	5.6	5.4	5.5	5.5	6.0	6.5	6.0	6.7	5.9	6.3	6.0	6.4	12.2	9.6	12.1	9.6
1.4	5.6	5.7	5.6	5.6	6.2	7.0	6.1	7.2	6.6	7.1	5.3	5.9	13.9	9.8	13.1	9.5
2.0	5.5	5.5	5.4	5.1	6.1	7.2	5.7	6.6	5.8	6.5	6.1	7.0	13.8	9.1	13.9	9.5
Moving Block Bootstrap																
0.2	4.2	4.8	4.2	4.5	5.0	4.9	5.1	5.3	5.3	5.0	4.5	4.4	4.4	5.1	6.2	6.0
0.8	5.4	4.9	4.2	4.7	4.8	5.8	5.3	5.7	4.0	5.0	4.3	4.9	4.2	5.9	5.3	5.4
1.4	5.6	4.9	5.3	5.0	5.2	6.0	5.4	4.8	4.5	5.1	5.5	5.1	4.8	4.9	7.1	5.3
2.0	5.6	5.2	5.5	5.0	5.1	5.0	5.4	5.0	4.3	4.7	4.6	4.8	6.6	4.8	5.0	5.3

Specifically, looking first at DGP 1 and DGP 2, the empirical size of our moving block bootstrap is mostly around the 5% nominal level irrespective of the forecast horizon, the sample size (including when $T = 50$), and the cut-off point π. Although the fixed regressor MSE-F bootstrap shows some over-rejections in small samples, it is largely comparable with our MBB bootstrap. We expect the fixed regressor bootstrap to work relatively well in DGP 1 and DGP 2 because the assumptions underlying the fixed regressor bootstrap are not violated. However, looking at DGP 3 and DGP 4, it is obvious that the over-rejections of the fixed regressor MSE-F bootstrap are very large, especially for 4-step-ahead forecasts ($\tau = 4$) and sample sizes $T \in \{50,100\}$. For example, its maximal rejection frequency in DGP 4 when $\tau = 4$ is 20.9% for $T = 50$ and 12.2% for $T = 100$. In most cases considered, our moving block bootstrap with both $\mathcal{T}_T$ and $\mathcal{W}_T$ outperforms the fixed regressor MSE-F bootstrap.

Comparing the relative size performance between moving block bootstrap with $\mathcal{T}_T$ and that with $\mathcal{W}_T$, both perform relatively well for all forecast horizons and cut-off points π considered, even when $T \in \{50,100\}$.

In Table 1, the DGPs include only one extra predictor ($k_2 = 1$) in the larger model. One of the advantages of our bootstrap procedure is that it simplifies the computational burdens of the asymptotic critical values of the statistics $\mathcal{T}_T$ and $\mathcal{W}_T$, especially when the DGP is not homoscedastic. As such, limiting the Monte Carlo experiment to the case where $k_2 = 1$ (as seen in Table 1) is not necessary for our purpose. Accordingly, we have also run the Monte Carlo experiment with larger model containing two extra predictors ($k_2 = 2$). More specifically, we use the following DGP when $k_2 = 2$:

$$y_t = 0.3y_{t-\tau} + \beta_{22}x_{1,t-\tau} + \beta_{23}x_{2,t-\tau} + u_{yt}, \tag{4.4}$$

$$x_{j,t} = 0.5x_{j,t-\tau} + u_{xjt},\ j = 1,2 \tag{4.5}$$

where the same four DGPs for the error vector (u_{yt}, u_{xjt}) are covered. In this setting, the null forecast model is of "*no influence*" of the predictors $x_{j,t-\tau}(j=1,2)$ in (4.4), whereas the alternative (unrestricted) forecast model takes the form of (4.4)–(4.5) with the parameters β_{22} and β_{23} being nonzero. Table 2 reports the results for the bootstrap tests. We only focus on the bootstrap tests in this table because the asymptotic critical values of the standard tests are difficult to compute in DGP 3 and DGP 4 despite a relatively small number of extra predictors in the larger model ($k_2=2$). As seen in Table 2, the rejection frequencies of the moving block bootstrap tests are in line with the results in Table 1. Specifically, our proposed moving block bootstrap has an overall good finite-sample performance and it outperforms the fixed regressor MSE-F bootstrap, especially in DGPs 3 and 4.[14]

4.2 | Power results

We now examine the power properties of the proposed moving block bootstrap $\mathcal{T}_T$ and $\mathcal{W}_T$ tests. To enable comparison, we also include the power analysis of the counterpart asymptotic tests, even though the latter are size distorted in most DGPs as shown in the size analysis section above. To simplify the presentation, we restrict the focus to $T\in\{50,200\}$ and cut-off point $\pi=0.8$.

Figure 1 shows the plots of the empirical rejection frequencies (vertical axis) for 4-step-ahead forecasts for $(a)\,T=50$ (top four subfigures) and $(b)\,T=200$ (bottom four subfigures) for values of $\beta_{22}\in[-1,1]$ (horizontal axis). Appendix S1 contains the power curves for 1-step-ahead forecasts; see Figure A1. In these plots, 0 in the horizontal axis corresponds to the null hypothesis $H_0:\beta_{22}=0$, that is, the rejection frequencies for this value of 0 are test empirical size. For $\beta_{22}\neq 0$, the rejection frequencies represent test empirical power. Each figure shows the empirical power of the asymptotic and the MBB tests for both the recursive-F $\mathcal{T}_T$ and the Wald $\mathcal{W}_T$ test statistics.

Several results stand out from these figures. First, the bootstrap empirical power is close to 1 even for moderate deviations from the null hypothesis when $T=200$, thus supporting the bootstrap consistency result in Theorem 3.3. The convergence of the bootstrap test is faster for both test statistics in DGPs 1-2 compared with DGPs 3-4. Also, the convergence seems to be faster in 1-step-ahead forecasts as seen in Figure A1 in Appendix A. Second, the bootstrap test has good power when $T=50$, irrespective of the forecast horizon, and this is the case even for small deviations from the null hypothesis. Third, in all cases considered (DGPs, sample sizes and forecast horizons), the MBB test with $\mathcal{T}_T$ has an edge in terms of power over that with $\mathcal{W}_T$. Finally, the empirical power curves of the standard $\mathcal{T}_T$ and $\mathcal{W}_T$ tests with asymptotic critical values are way above the 5% nominal level line, thus underscoring the lack of controlling size with these tests. As such, their power advantage shown in Figure 1 is attributable to their asymptotic size distortions and therefore must be ignored.

A important contribution of Hansen and Timmermann's (2015) Wald approximation is that it facilitates the computation of the asymptotic critical values. Seeing the lack of size control of their proposed Wald-type test for moderate sample sizes (e.g., $T=200$), our hybrid bootstrap method is more appealing than the Wald approximation. In addition, even when the sample size is large, the asymptotic critical values of the Wald statistic

TABLE 2 Moving block bootstrap rejection frequencies with $k_2=2, \alpha=5\%$.

	$\tau=1$								$\tau=4$							
	DGP 1		DGP 2		DGP 3		DGP 4		DGP1		DGP 2		DGP 3		DGP 4	
π	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald	$\mathcal{T}_T$	Wald
$T=50$																
0.2	3.7	3.1	5.2	4.0	5.7	5.9	6.5	5.6	3.4	2.7	3.9	3.0	10.0	8.1	9.3	8.2
0.8	4.4	3.7	3.8	3.3	5.5	5.5	4.5	5.6	5.0	2.6	4.3	3.2	12.4	7.9	13.1	7.7
1.4	4.7	3.7	5.6	3.7	5.1	4.8	6.0	5.3	4.1	2.9	5.2	3.0	13.2	7.2	12.5	7.5
2.0	6.1	3.9	4.6	4.6	4.8	5.9	5.5	5.1	4.5	2.9	5.1	3.0	16.2	7.6	14.1	7.3
$T=200$																
0.2	4.5	4.5	4.7	4.9	4.9	5.1	5.3	5.7	4.6	5.0	4.6	4.2	5.5	5.8	5.0	4.8
0.8	4.7	4.8	5.0	4.9	4.5	5.4	4.9	5.7	4.5	5.3	4.2	5.5	5.6	5.5	5.7	5.0
1.4	4.9	4.5	4.6	5.2	4.6	5.6	6.3	5.9	4.8	4.7	5.2	4.8	5.6	5.6	5.6	5.5
2.0	5.2	4.9	5.2	5.1	4.7	5.0	4.6	5.6	4.8	5.1	5.0	4.8	5.2	5.0	4.8	5.2

can only be simulated easily in specific cases (as discussed above in Section 2.3), whereas our bootstrap method applies in more general settings, including when the larger model has more than one extra predictor and the DGP is heteroscedastic or weakly dependent. Finally, the bootstrap tests do not suffer much from power loss compared with the asymptotic tests, and they provide better size control.

5 | EMPIRICAL APPLICATIONS

We illustrate our theoretical results through two applications. The first application examines the predictive ability of Chicago Fed National Activity Index (CFNAI) and other inflation measures for forecasting core PCE inflation (similar to Clark & McCracken, 2015). The second is drawn from Stock and Watson (2003) and Clark and

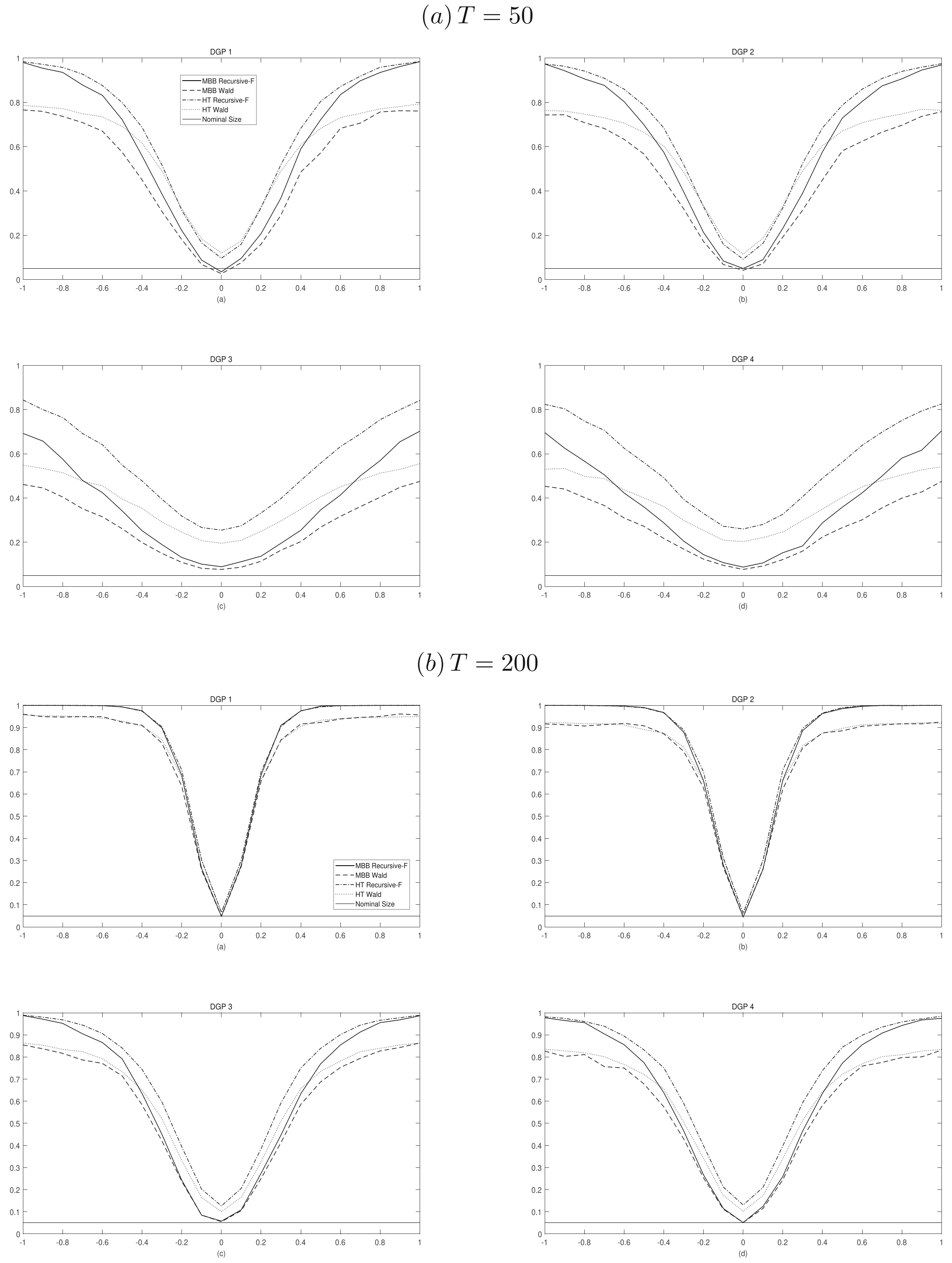

FIGURE 1 Power of the MBB tests at level 5%, four-step ahead forecasts: $\tau = 4$.

McCracken (2012) and looks at forecasting quarterly US GDP growth using a range of potential indicators.

5.1 | Forecasting core inflation

In this application, we compare 1-quarter and 4-quarter ahead forecasts of inflation from two models. In the 1-quarter ahead forecasting exercise, the baseline (restricted) model relates the change in inflation at $t+1$ to current and one lagged value of inflation change, that is,

$$y_{t+1} = b_0 + b_1 y_t + b_2 y_{t-1} + u_{y,t+1}, \tag{5.1}$$

where $y_t = \Delta\pi_t, \pi_t = 400 ln\left(\frac{P_t}{P_{t-1}}\right)$ and P_t is the aggregate price index at t. The alternative (unrestricted) model includes CFNAI, PCE food price inflation less core inflation, and import price inflation less core inflation, that is,

$$y_{t+1} = b_0 + b_1 y_t + b_2 y_{t-1} + x_t' b_3 + u_{x,t+1}, \tag{5.2}$$

where x_t contains period t values of CFNAI, PCE food price inflation less core inflation, and import price inflation less core inflation. In the 4-quarter ahead forecasting case, the baseline (restricted) model relates $y_{t+4}^{(4)} - y_t^{(4)}$ to a constant and $y_t^{(4)} - y_{t-4}^{(4)}, y_t^{(4)} = 100 ln\left(\frac{P_t}{P_{t-4}}\right)$. The alternative (unrestricted) model adds the period t values of CFNAI, relative food price inflation, and relative import price inflation to the baseline model.[15] In both cases, the data sample spans $1983:Q3$ through $2008:Q2$ ($T=100$) and we use the cut-off point $\pi = 1.4$ for the initial estimation. Thus, out-of-sample forecasts from $1994:Q2+\tau-1$ through $2008:Q2$ ($\tau \in \{1,4\}$) are obtained and the corresponding statistics computed. The bootstrap critical values are obtained with 9999 replications. The results are presented in Table 3. The first column of the table shows the variables included in the alternative model, whereas the other columns show for each forecast horizon the p-values of the proposed bootstrap test along with the *fixed regressor* wild bootstrap (FRWB) MSE-F test of Clark and McCracken (2015).

The main findings from this table can be summarized as follows. First, including CFNAI, PCE food price inflation and import price inflation do not improve forecast accuracy of core inflation at 1-quarter ahead horizon, whereas it does for 4-quarter ahead forecasts. Second, the p-value of the FRWB is 0.000 for 4-quarter ahead forecasting, whereas those of the MBB F- and Wald-statistics stand at 0.066 and 0.041, respectively. This means that using the FRWB leads to rejecting the baseline model at 1% nominal level, whereas the MBB $\mathcal{T}$ and $\mathcal{W}$ tests fail to reject the baseline model at 1% nominal level. In particular, the MBB $\mathcal{T}$ test, which has higher power than the MBB $\mathcal{W}$ test for small to moderately large sample sizes (see the power analysis in Section 4.2), even fails to reject the baseline model at the usual 5% nominal level. The findings are also consistent with our size analysis in Section 4.1 where we show that the fixed regressor MSE-F bootstrap tends to over-reject even with a moderately large sample size, especially in multi-step ahead forecasts and in the presence of autocorrelation in the errors. Therefore, this application demonstrates that the choice of the bootstrap is important as it can lead to different conclusions.

TABLE 3 Test of equal accuracy for core inflation.

Restricted variables	MBB $\mathcal{T}_T$	MBB $\mathcal{W}_T$	Fixed regressor MSE-F
One-quarter ahead: $\tau = 1$			
CFNAI, food, imports	0.186	0.957	0.314
Four-quarter ahead: $\tau = 4$			
CFNAI, food, imports	0.066	0.041	0.000

5.2 | Forecasting real GDP growth

In this application, we examine the performance of 13 alternative models with respect to the baseline model in forecasting real GDP growth. As in the previous application, the comparison is done for τ-period ahead forecasts with $\tau \in \{1,4\}$. The baseline model includes a constant and one lag of real GDP growth, where GDP growth between t and $t-\tau$ is measured as $y_t = (400/\tau) ln(GDP_t/GDP_{t-\tau})$, that is,

$$y_t = \beta_0 + \beta_1 y_{t-\tau} + u_{yt}, \tag{5.3}$$

whereas each of the 13 alternative models adds a potential leading indicator x_t to (5.3), that is,

$$y_t = \beta_0 + \beta_1 y_{t-\tau} + \beta_2 x_{t-\tau} + u_{xt}, \tag{5.4}$$

where u_{xt} and u_{yt} are error terms. The set of leading indicators used are shown in Table 4. It includes the change in consumption's share in GDP (measured with nominal data), weekly hours worked in manufacturing, building permits, purchasing manager indexes for supplier delivery times and orders, new claims for unemployment insurance, growth in real stock prices, change in 3-month Treasury bill rate, change in 1-year Treasury bond yield, change in 10-year Treasury bond yield,

TABLE 4 Test of equal accuracy for GDP.

	One-quarter ahead: $\tau = 1$			Four-quarter ahead: $\tau = 4$		
	MBB		Fixed regressor	MBB		Fixed regressor
Restricted variables	$\mathcal{T}_T$	$\mathcal{W}_T$	MSE-F	$\mathcal{T}_T$	$\mathcal{W}_T$	MSE-F
Δ (C/Y)	0.000	0.000	0.000	0.201	0.271	0.059
Δ ln Permits	0.001	0.000	0.000	0.000	0.000	0.000
Δ ln S&P500	0.002	0.000	0.000	0.197	0.019	0.000
Spread, Baa-Aaa	0.066	0.092	0.072	0.374	0.608	0.841
PMI Orders	0.092	0.007	0.000	0.298	0.119	0.011
Unemployment claims	0.183	0.373	0.217	0.338	0.393	0.307
Δ 3-month treasury	0.192	0.470	0.026	0.507	0.551	0.997
Δ 1-year treasury	0.232	0.280	0.451	0.656	0.560	0.906
Hours	0.344	0.535	0.420	0.157	0.140	0.189
PMI deliveries	0.333	0.644	0.905	0.741	0.187	0.999
Δ 10-year treasury	0.425	0.371	0.530	0.932	0.860	0.889
Spread, 10 years - 3 months	0.999	0.995	0.995	1.000	0.999	0.999
Spread, 10 years - 1 year	1.000	1.000	0.997	1.000	1.000	0.998

3-month to 10-year yield spread, 1-year to 10-year yield spread, and spread between Aaa and Baa corporate bond yields from Moody's. The data span the period $1961:Q2$ through $2009:Q4$ ($T=195$) and out-of-sample forecasts from $1981:Q4+\tau-1$ through $2009:Q4$ ($\tau \in \{1,4\}$) are obtained and the corresponding statistics computed.

The results are reported in Table 4.[16] The first column of the table shows the extra predictor added to the baseline model (thus determining alternative (or unrestricted) model), whereas the other columns show each test's p-value from the pairwise forecast comparisons. The top half of the table reports results for one-quarter ahead forecasts, whereas the bottom half shows results for four-quarter ahead forecasts.

Considering first one-quarter ahead forecasts, we see that tests based on our MBB suggest that five models—those including change in consumption share, growth in building permits, growth in stock prices, Baa-Aaa interest rate spread and PMI new orders - improve the accuracy of forecasts relative to the benchmark AR(1) model. In addition to the above five models, the FRWB test finds that the alternative model that adds change in the 3-month Treasury bill rate to the baseline model also improves slightly the one-quarter ahead forecasts of GDP growth (p-value of 2.6%), whereas our MBB fails to reject the baseline in that case at 10% nominal level (p-values of 19.2% and 47.0% for the MBB F and Wald tests, respectively).

Next, looking at the four-quarter ahead forecasts, our MBB suggests that two models - those including growth in building permits and growth in stock prices - forecast better than the benchmark AR(1) model at 5% nominal level.[17] On the other hand, the FRWB also adds the models with change in consumption share and PMI orders to the above models in terms of their forecast performance in comparison to the baseline model. According to our MBB, the higher predictive power of the change in consumption share and PMI orders seem to disappear in the four-quarter ahead forecasts even at 10% nominal level. Meanwhile, the FRWB fails to pick this up suggesting that the test over-rejects in some cases, which is in line with the Monte Carlo evidence reported earlier.

6 | CONCLUSION

In this paper, we examine the finite-sample performance (size and power) of the recursively generated F-test of out-of-sample predictive accuracy (McCracken, 2007) and its equivalent Wald approximation (Hansen & Timmermann, 2015). We show through Monte Carlo experiments that even for moderate sample sizes, both tests can be oversized, especially when the forecast errors exhibit serial correlation. We then propose a bootstrap method for both statistics and establish its consistency even when the forecast errors are autocorrelated, irrespective of the forecast horizon. Our bootstrap method is valid and easy to implement in cases where the larger model contains many extra predictors and the data generating process is heteroscedastic or weakly dependent, situations under which the asymptotic critical values of the standard recursive F or Wald statistics are difficult to simulate.

The proposed bootstrap is a hybrid of a moving block bootstrap (which is nonparametric) and a residual-based bootstrap (which is parametric). We suggest a practical means of choosing the block length in a data-dependent way. In particular, we argue that in order to capture the autocorrelation structure of regression residuals practitioners should choose the block length that mimics the optimal lag length of the Newey and West's (1987) HAC estimator. Monte Carlo simulations show that the proposed bootstrap test has overall good finite-sample performance. The method is also illustrated with applications on forecasting core inflation and real GDP growth.

ACKNOWLEDGEMENTS

The authors thank Prosper Dovonon, James A. Duffy, Jean-Marie Dufour, Mardi Dungey, Leandro Magnusson, Sophocles Mavroeidis, Adrian Pagan, Peter C.B. Phillips, Richard Smith, Mark Weder, and the participants of the 27th NZESG conference at the Otago Business School, Annual University of Tasmania (UTAS) macroeconomics workshop 2018, and the seminar participants at University of Oxford. This work was supported with super computing resources provided by the Phoenix HPC service at The University of Adelaide. Doko Tchatoka acknowledges financial support from the Australian Research Council through the Discovery Grant DP200101498. Open access publishing facilitated by The University of Adelaide, as part of the Wiley - The University of Adelaide agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

The data that support the empirical applications have been obtained from openly available database of Federal Reserve Economic Data (FRED) and from Clark, T. E. and M. W. McCracken (2012) Reality checks and comparisons of nested predictive models, Journal of Business & Economic Statistics 30 (1), 53-66, available from https://doi.org/10.1198/jbes.2011.10278.

ORCID

Firmin Doko Tchatoka https://orcid.org/0000-0003-1876-0633
Qazi Haque https://orcid.org/0000-0003-0694-4443

ENDNOTES

[1] See Diebold and Mariano (1995), West (1996), White (2000), Stock and Watson (2003), Giacomini and White (2006), Corradi and Swanson (2007), McCracken (2007), Clark and McCracken (2001, 2005, 2012, 2014, 2015), Rossi and Inoue (2012), Hansen and Timmermann (2012), among others.

[2] See Kunsch (1989).

[3] See, for example, Clark and McCracken (2005); Giacomini and White (2006); Clark and West (2007); Clark and McCracken ((2012), (2015)); Hansen and Timmermann (2015).

[4] The HAC estimator with the Bartlett kernel is utilized in the simulations and empirical applications, but any kernel in class $\mathscr{K}_3$ of Andrews ((1991), eq.(7.1)) could be employed. The block length of the kernel bandwidth is selected following the recommendations of Andrews and Monahan (1992) and Newey and West (1994).

[5] That is, there exists $\bar{U}_{jt}$ such that $|U_{j,p}(t)| < \bar{U}_{jt}$ and $\mathbb{E}_F\left[|\bar{U}_{jt}|^{3(2+1/\psi)}\right] < \infty$, for all t,j and p, where $U_{jt} := (U_{j,p}(t))_{1\le p\le k}$.

[6] Let $\{V_t\}$ be a stochastic process and $\mathscr{F}_{t-n}^{t+n} := \sigma(V_{t-n},\ldots,V_{t+n})$ denote the σ-field generated by $V_{t-n},\ldots,V_{t+n}$. We define a process $\{W_t\}$ to be NED (Near Epoch Dependent) on a mixing process $\{V_t\}$ if $\mathbb{E}_F[\|W_t\|^2] < \infty$ and $\nu_n := \sup_t \|W_t - E_{t-n}^{t+n}(W_t)\|_2 \to 0$ as $n\to\infty$, where $\|\cdot\|_p$ is the L_p norm and $E_{t-n}^{t+n}(\cdot) \equiv \mathbb{E}_F[\cdot|\mathscr{F}_{t-n}^{t+n}]$. $\{W_t\}$ is NED on $\{V_t\}$ of size $-a$ if $\nu_n = O(n^{-a-\delta})$ for some $\delta > 0$. We say that $\{V_t\}$ is strong mixing with coefficients $\alpha_n \equiv \sup_m \sup_{A\in\mathscr{F}_{-\infty}^m, B\in\mathscr{F}_{m+n}^\infty} |P(A\cap B) - P(A)P(B)|$ if $\alpha_n \to 0$ as $n\to\infty$ suitably fast.

[7] See, for example, Clark and McCracken (2001, 2012, 2014, 2015); McCracken (2007), and Hansen and Timmermann (2015), among others.

[8] Note that by definition, $s = \max\{q,\tau\} + 1$, with q being the maximum lag length of y_t included in $X_{2,t}$.

[9] Note that the HAC estimator σ_ϵ^2 is computed using the full bootstrap sample.

[10] Note that in addition to the $AR(1)$ property, the forecast errors also exhibit the usual $MA(\tau-1)$ form when $\tau = 4$. The form of the $MA(\tau-1)$ is identical to the one in Clark and McCracken (2012) but the AR(1) part is new.

[11] We also run the experiments with $\alpha \in \{1\%, 10\%\}$ and the main findings remain qualitatively unchanged.

[12] Table S1 contains the results for $T = 100$.

[13] Note that DGP 3 and 4 violate the assumptions of the fixed regressor MSE-F bootstrap, which only allows for an MA($\tau-1$) structure.

[14] The rejection frequencies for the fixed regressor MSE-F tests are reported in Table S3.

[15] To simplify the lag structure, the relative food and import price inflation variables are computed as two-period averages of quarterly (relative) inflation rates, similar to Clark and McCracken (2015).

[16] 9999 replications were used to approximate the bootstrap critical values.

[17] In case of growth in stock prices, test based on recursive F statistics fails to reject the baseline at 10% nominal level, therefore providing mixed evidence regarding this variable.

REFERENCES

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, *59*(3), 817–858.

Andrews, D. W. K., & Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, *60*(4), 953–966.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*(1), 85–110.

Clark, T. E., & McCracken, M. W. (2005). Evaluating direct multistep forecasts. *Econometric Reviews*, *24*(4), 369–404.

Clark, T. E., & McCracken, M. W. (2012). Reality checks and comparisons of nested predictive models. *Journal of Business & Economic Statistics*, *30*(1), 53–66.

Clark, T. E., & McCracken, M. W. (2014). Tests of equal forecast accuracy for overlapping models. *Journal of Applied Econometrics*, *29*(3), 415–430.

Clark, T. E., & McCracken, M. W. (2015). Nested forecast model comparisons: A new approach to testing equal accuracy. *Journal of Econometrics*, *186*(1), 160–177.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, *138*(1), 291–311.

Corradi, V., & Swanson, N. R. (2007). Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *International Economic Review*, *48*(1), 67–109.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*. Oxford: Oxford Uuniversity Press.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, Vol. 38: SIAM.

Fitzenberger, B. (1998). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics*, *82*(2), 235–287.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Hansen, P., & Timmermann, A. (2012). Choice of sample split in out-of-sample forecast evaluation. (*Economics Working Papers ECO2012/10*): European University Institute.

Hansen, P. R., & Timmermann, A. (2015). Equivalence between out-of-sample forecast comparisons and wald statistics. *Econometrica*, *83*(6), 2485–2505.

Jansson, M. (2002). Consistent covariance matrix estimation for linear processes. *Econometric Theory*, *18*(6), 1449–1459.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241.

Mann, H. B., & Wald, A. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, *14*(3), 217–226.

McCracken, M. W. (2007). Asymptotics for out of sample tests of granger causality. *Journal of Econometrics*, *140*(2), 719–752.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708.

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

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, *89*(428), 1303–1313.

Rossi, B., & Inoue, A. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economic Statistics*, *30*(3), 432–453.

Stock, J. H., & Watson, M. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, *41*(3), 788–829.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*(5), 1067–1084.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Doko Tchatoka, F., & Haque, Q. (2023). On bootstrapping tests of equal forecast accuracy for nested models. *Journal of Forecasting*, *42*(7), 1844–1864. https://doi.org/10.1002/for.2987

APPENDIX A

In this appendix, we present the proofs of the main results. Additional lemmas and simulation results are contained in Appendix S1.

A.1 | Proof of main results

Proof of Theorem 3.1. From (1.12) in the proof of Lemma 1.4, we have

$$\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{}\beta^*_{j,t}-\hat{\beta}_{j,t})=\frac{H_j^{-1}}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau} \times\left(\frac{1}{t}\sum_{n=s}^{t}\left[s^*_{j,n}(\hat{\beta}_{j,t})-\mu_T\sum_{n=s}^{T-\tau}s_{j,n}(\hat{\beta}_{j,t})\right]\right)+o_{p^*}(1)\text{ pr-}\mathbb{P}. \tag{A1}$$

We can express $s^*_{j,n}(\hat{\beta}_{j,t})$ and $s_{j,n}(\hat{\beta}_{j,t})$ as

$$\begin{aligned}s^*_{j,n}(\hat{\beta}_{j,t})&=X^*_{j,n-\tau}[y^*_n-X^{*\prime}_{j,n-\tau}(\hat{\beta}_{j,t}-\beta^0_j)]\\&=s^*_{j,n}(\beta^0_j)-h^*_{j,n-\tau}(\hat{\beta}_{j,t}-\beta^0_j)\end{aligned}$$

$$s_{j,n}(\hat{\beta}_{j,t}) = X_{j,n-\tau}[y_n - X'_{j,n-\tau}(\hat{\beta}_{j,t} - \beta_j^0)] = s_{j,n}(\beta_j^0) - h_{j,n-\tau}(\hat{\beta}_{j,t} - \beta_j^0)$$

so that (A1) can be written as

$$\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{j,t}^* - \hat{\beta}_{j,t}) = \frac{H_j^{-1}}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}\left(\frac{1}{t}\sum_{n=s}^{t}\left[s_{j,n}^*(\beta_j^0) - \mu_T\sum_{n=s}^{T-\tau}s_{j,n}(\beta_j^0)\right]\right) - \frac{H_j^{-1}}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}\left(\frac{1}{t}\sum_{n=s}^{t}\left(h_{j,n-\tau}^* - \mu_T\sum_{n=s}^{T-\tau}h_{j,n-\tau}\right)\left(\hat{\beta}_{j,t} - \beta_j^0\right)\right) + o_{p^*}(1) \text{ pr-}\mathbb{P}. \quad (A2)$$

From Lemma 1.5, the second term of the RHS of (A2) is $o_{p^*}(1)$ pr-$\mathbb{P}$. Therefore, we have

$$\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{j,t}^* - \hat{\beta}_{j,t}) = \frac{H_j^{-1}}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}\left(\frac{1}{t}\sum_{n=s}^{t}\left[s_{j,n}^*(\beta_j^0) - \mu_T\sum_{n=s}^{T-\tau}s_{j,n}(\beta_j^0)\right]\right) + o_{p^*}(1) \text{ pr-}\mathbb{P}. \quad (A3)$$

Now, from Fitzenberger (1998, lemma 1), we have

$$\frac{1}{t}\sum_{n=s}^{t}\mathbb{E}_{F^*}[s_{j,n}^*(\beta_j^0)] = \mu_T\sum_{n=s}^{T-\tau}s_{j,n}(\beta_j^0) + O_p(\ell_T T^{-1}),$$

thus we can express (A3) as

$$\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{j,t}^* - \hat{\beta}_{j,t}) = \frac{H_j^{-1}}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}\frac{1}{t}\sum_{n=s}^{t}\left(s_{j,n}^*(\beta_j^0) - \mathbb{E}_{F^*}[s_{j,n}^*(\beta_j^0)]\right) + o_{p^*}(1) \text{ pr-}\mathbb{P}. \quad (A4)$$

By following the same steps as in Lemma 1.2, we can write the RHS of (A4) as

$$\frac{H_j^{-1}}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}\frac{1}{\sqrt{T}}\frac{(T/t)}{\sqrt{T}}\sum_{n=s}^{t}\left(s_{j,n}^*(\beta_j^0) - \mathbb{E}_{F^*}[s_{j,n}^*(\beta_j^0)]\right) + o_{p^*}(1) \text{ pr-}\mathbb{P} = \sqrt{1-\rho}H_j^{-1}\frac{(T/t)}{\sqrt{T}}\sum_{n=s}^{t}\left(s_{j,n}^*(\beta_j^0) - \mathbb{E}_{F^*}[s_{j,n}^*(\beta_j^0)]\right) + o_{p^*}(1) \text{ pr-}\mathbb{P}.$$

We deal with $j=2$ and $j=1$ separately. First, note as in the proof of Lemma 1.2-(b) that for $j=2$, Lemma 1.2-(a) along with the bootstrap sampling implies that

$$\sqrt{1-\rho}H_2^{-1}\frac{(T/t)}{\sqrt{T}}\sum_{n=s}^{t}(s_{2,n}^*(\beta_2^0) - \mathbb{E}_{F^*}[s_{2,n}^*(\beta_2^0)]) \Rightarrow \sqrt{1-\rho}r^{-1}H_2^{-1}\Upsilon^{-1}H_2^{-1/2}\tilde{A}'B^*(r) + o_{as^*}(1)\ a.s - \mathbb{P},$$

where $B^*(r)$ is a k_2 dimensional vector of standard Brownian motion.

$$\text{As } var\left[\sqrt{1-\rho}r^{-1}H_2^{-1}\Upsilon^{-1}H_2^{-1/2}\tilde{A}'B^*(r)\right] = r^{-3}(1-\rho)H_2^{-1}\Upsilon^{-1}H_2^{-1},$$

it is clear that

$$\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{2,t}^* - \hat{\beta}_{2,t}) \Rightarrow B\left[r^{-3}(1-\rho)H_2^{-1}\Upsilon^{-1}H_2^{-1}\right]\ a.s - \mathbb{P}, \quad (A5)$$

which is the distribution of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\hat{\beta}_{2,t} - \beta_2^0)$ given in Lemma 1.2. Similarly, for $j=2$, we find

$$\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{1,t}^* - \hat{\beta}_{1,t}) \Rightarrow B\left[r^{-3}(1-\rho)H_1^{-1}J'\Upsilon^{-1}JH_1^{-1}\right]\ a.s - \mathbb{P}, \quad (A6)$$

which also is the distribution of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\hat{\beta}_{1,t} - \beta_1^0)$ given in Lemma 1.2.

Overall, this results show that $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\tilde{\beta}_{j,t}^* - \hat{\beta}_{j,t})$ converges almost surely to the asymptotic distribution of $\frac{1}{\sqrt{P_T}}\sum_{t=R_T}^{T-\tau}(\hat{\beta}_{j,t} - \beta_j^0)$ for all $j=1,2$, thus e□blishing Theorem 3.1. ■

Proof of Theorem 3.2. $\mathcal{T}_T$ and $\mathcal{W}_T$ are asymptotically equivalent, it suffices to establish the result for $\mathcal{W}_T$. Also, as the MSE loss differential in the numerator of $\mathcal{T}_T$ is related to the homoskedastic Wald statistics (see Hansen & Timmermann, 2015, Corollary 1) regardless of whether the underlying process is homoskedastic and regardless of whether the null hypothesis holds or not, we consider $\mathcal{W}_T = \hat{S}_T - \hat{S}_{R_T} + \sigma_\varepsilon^{-2}\check{\kappa}log(\rho)$ where $\hat{\sigma}_\varepsilon^2(\hat{S}_T - \hat{S}_{R_T})$ is equal to

$$\tilde{S}_T - \tilde{S}_{R_T} = \check{U}'_{T,T}\check{H}_2^{-1}\check{U}_{T,T} - \frac{T}{R_T}\check{U}'_{T,R_T}\check{H}_2^{-1}\check{U}_{T,R_T} + o_p(1) + \beta'_{22}\sum_{t=R_T+1}^{T} Z_{t-\tau}Z'_{t-\tau}\beta_{22} + 2\sqrt{T}\beta'_{22}(\check{U}_{T,T} - \check{U}_{T,R_T}), \tag{A7}$$

$\tilde{S}_T$ given in (2.7), $\check{U}_{T,T} = \frac{1}{\sqrt{T}}\sum_{t=s}^{T} Z_{t-\tau}\varepsilon_{2t}$, $\tilde{S}_{R_T}, \check{U}_{T,R_T}$ are the corresponding of both respectively in the sub-sample with R_T observations.

Suppose first that $\beta_{22} = 0$. From the proof of Theorem 3 in Hansen and Timmermann (2015, p. 2503),

$$\tilde{S}_T - \tilde{S}_{R_T} \xrightarrow{d} B(1)'\check{\Omega}_\infty^{1/2}\check{H}_2^{-1}\check{\Omega}_\infty^{1/2}B(1) - \rho^{-1}B(\rho)'\check{\Omega}_\infty^{1/2}\check{H}_2^{-1}\check{\Omega}_\infty^{1/2}B(\rho) \tag{A8}$$

where $\check{\Omega}_\infty$ is the log-run variance of the process $\{Z_{t-\tau}\varepsilon_{2t}\}$ and $B(r) \in \mathbb{D}_{[0,1]}^{k_2}$. Because $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + o_p(1)$, then $\mathcal{W}_T$ is distributed as σ_ε^{-2} times the limit distribution of $\tilde{S}_T - \tilde{S}_{R_T}$ in (A8). To establish the validity of the bootstrap for $\mathcal{W}_T$ when $\beta_{22} = 0$, it suffices to establish that $\mathcal{W}_T^*$ converges to σ_ε^{-2} times the limit distribution in (A8) a.s.-$\mathbb{P}$.

First, note that $\mathcal{W}_T^* = \hat{S}_T^* - \hat{S}_{R_T}^* + \hat{\sigma}_\varepsilon^{*\,-2}\check{\kappa}^* log(\rho)$, and under H_0 along with the results of Lemmas 1.1-1.5, we can express $\hat{\sigma}_\varepsilon^{*\,2}(\hat{S}_T^* - \hat{S}_{R_T}^*)$ as:

$$\tilde{S}_T^* - \tilde{S}_{R_T}^* = \check{U}^{*\prime}_{T,T}\check{H}_2^{-1}\check{U}^*_{T,T} - \frac{T}{R_T}\check{U}^{*\prime}_{T,R_T}\check{H}_2^{-1}\check{U}^*_{T,R_T} + \beta'_{22}\sum_{t=R_T+1}^{T} Z^*_{t-\tau}Z^{*\prime}_{t-\tau}\beta_{22} + 2\sqrt{T}\beta'_{22}\left(\check{U}^*_{T,T} - \check{U}^*_{T,R_T}\right) + o_{p^*}(1)\ pr - \mathbb{P}, \tag{A9}$$

where the various quantities in stars are the analogues to the ones in (A7) in the bootstrap sample. It is easy to see from the model assumptions, along with the results of Lemmas 1.1–1.5, that

$$\tilde{S}_T^* - \tilde{S}_{R_T}^* \xrightarrow{d^*} B^*(1)'\check{\Omega}_\infty^{1/2}\check{H}_2^{-1}\check{\Omega}_\infty^{1/2}B^*(1) - \rho^{-1}B^*(\rho)'\check{\Omega}_\infty^{1/2}\check{H}_2^{-1}\check{\Omega}_\infty^{1/2}B^*(\rho),\ a.s - \mathbb{P}^* \tag{A10}$$

under H_0, where $B^*(r) \in \mathbb{D}_{[0,1]}^{k_2}$. Because $B^*(r)$ in (A10) and $B(r)$ in (A8) have the same distribution, it is the case that (A10) holds a.s.-$\mathbb{P}$ with $B^*(r)$ replaced by $B(r)$, that is,

$$\tilde{S}_T^* - \tilde{S}_{R_T}^* \xrightarrow{d} B(1)'\check{\Omega}_\infty^{1/2}\check{H}_2^{-1}\check{\Omega}_\infty^{1/2}B(1) - \rho^{-1}B(\rho)'\check{\Omega}_\infty^{1/2}\check{H}_2^{-1}\check{\Omega}_\infty^{1/2}B(\rho),\ a.s - \mathbb{P}. \tag{A11}$$

Therefore, $\mathcal{W}_T^*$ has the same asymptotic distribution as $\mathcal{W}_T$ a.s-$\mathbb{P}$ under H_0.

Suppose now that $\beta_{22} \neq 0$. If $\beta_{22} = cT^{-1/2}b$ (local-to-zero alternative) for some c and b as in (2.13), then we can show as in the case under H_0 that $\tilde{S}_T^* - \tilde{S}_{R_T}^*$ converge almost surely to the asymptotic distribution of $\tilde{S}_T - \tilde{S}_{R_T}$. If $\beta_{22} \neq 0$ is fixed, it easy to see from (A7) and the model assumptions that $\tilde{S}_T - \tilde{S}_{R_T} \Rightarrow \infty$ because the first two terms in the RHS of (A7) are $O_p(1)$, whereas the last two terms diverge. As such, we also have $\tilde{S}_T - \tilde{S}_{R_T} \xrightarrow{d} \infty$ because weak convergence implies convergence in distribution. Because $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + o_p(1)$ irrespective of the value of β_{22}, it follows that the above result implies that

$$\mathcal{W}_T \xrightarrow{d} \infty \ \text{ if } \beta_{22} \neq 0. \tag{A12}$$

□

Similarly, we have that both $\tilde{S}_T^* - \tilde{S}_{R_T}^* \xrightarrow{d} \infty$ a.s- $\mathbb{P}$ so that $\mathcal{W}_T^* \xrightarrow{d} \infty$ a.s- $\mathbb{P}$. Because $\mathcal{W}_T^*$ and $\mathcal{W}_T$ also diverge when $\beta_{22} \neq 0$ is fixed, it is clear that Theorem 3.2 holds in that case.

■

Proof of Theorem 3.3. Let $c^\infty_{0\cdot\mathcal{T}}(\alpha)$ and $c^\infty_{0\cdot\mathcal{W}}(\alpha)$ denote the $(1-\alpha)^{th}$ quantiles under H_0 of the asymptotic distributions of $\mathcal{T}_T$ and $\mathcal{W}_T$ respectively. We know from (2.11)–(2.12) that $c^\infty_{0\cdot\mathcal{T}}(\alpha) < \infty$ and $c^\infty_{0\cdot\mathcal{W}}(\alpha) < \infty$. From Theorem 3.2, we have

□

$$c_{\mathcal{T}}^*(\alpha) \xrightarrow{p} c_{0\cdot\mathcal{T}}^\infty(\alpha) < \infty,\ c_{\mathcal{W}}^*(\alpha) \xrightarrow{p} c_{0\cdot\mathcal{W}}^\infty(\alpha) < \infty \ \text{ as } T \to \infty. \tag{A13}$$

We know from the above proof of Theorem 3.2 that $\mathcal{W}_T \Rightarrow \infty$ if $\beta_{22} \neq 0$ is fixed. Then, Theorem 3.3 follows by combining this convergence result with (A13). ■

AUTHOR BIOGRAPHIES

Firmin Doko Tchatoka is a Professor at the School of Economics and Public Policy, The University of Adelaide (Australia). His main research interests are in Econometrics and Statistics, where he has developed deep knowledge of both theory and applied. His research spans a wide range of topics including causal inference in structural models, model selection and optimal inference in moment condition models, bootstrap and Monte Carlos methods, treatment effect and networks, forecasting and identification of macroeconomic structural models, the housing market, etc. Professor Doko Tchatoka has published in top economics and statistics journals, including the Journal of Econometrics and Econometric Theory. He is the lead investigator of multiple national and international competitive research grant projects, Australian Research Council (ARC) grants. He held visiting scholar positions internationally, such as at McGill University (Canada), Yale University (US), University of Toronto (Canada), Concordia University (Canada). He served on many Committees of the prestigious Econometric Society meetings, and he was twice the keynote lecturer of the Summer Institute of the Econometric Society (Benin in 2018 and Ivory Coast in 2021).

Qazi Haque is a Lecturer in the School of Economics and Public Policy at The University of Adelaide. His research interests are in Macroeconomics and Macroeconometrics with a focus on policy-relevant issues. He holds a PhD in Economics from The University of Adelaide. Prior to joining The University of Adelaide as a Lecturer, he worked as an Economist at the Reserve Bank of Australia and as a Post-Doctoral researcher at The University of Western Australia. He is also affiliated as a Research Associate in the Centre for Applied Macroeconomic Analysis at the Australian National University and has been a visiting researcher at renowned academic and policy institutions including the University of Oxford, Reserve Bank of New Zealand and De Nederlandsche Bank. His research has been published in top-ranked journals including, among others, in the European Economic Review, Journal of Money, Credit and Banking, Journal of Economic Dynamics and Control, International Journal of Central Banking. He is currently serving as the Secretary in the Executive Board of The Australasian Macroeconomics Society.