International Journal of Forecasting 33 (2017) 1–10

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Monte Carlo forecast evaluation with persistent data

Lynda Khalaf[a,b,c,*], Charles J. Saunders[d]

[a] *Department of Economics and Centre for Monetary and Financial Economics (CMFE), Carleton University, Canada*

[b] *Centre interuniversitaire de recherche en économie et analyse quantitative (CIREQ), Canada*

[c] *Centre de Recherche en économie de l'Environnement, de l'Agroalimentaire, des Transports et de l'Énergie (CREATE), Université Laval, Canada*

[d] *Department of Economics, University of Western Ontario, Canada*

ARTICLE INFO

Keywords:
Evaluating forecasts
Model selection
Stationarity
Statistical tests
Monte Carlo
Finite sample tests

ABSTRACT

Persistent processes, including local-to-unity and random walks, are commonly considered as forecasting models of interest. However, the associated forecast errors follow non-standard distributions that complicate forecast evaluation tests. We propose a finite sample simulation-based solution to this problem. The method requires a flexible parametric null model that can be simulated as long as a finite dimension nuisance parameter can be specified. The size control of our method is robust to non-standard limiting distributions, such as degenerate asymptotic distribution problems that arise from nested and unit root models. Our simulation studies demonstrate that many of the existing forecast evaluation methods, including various bootstraps, over-reject for highly persistent data. In contrast, our method is level correct and has good power. We extend our approach to the inversion of forecast evaluation statistics in order to construct exact confidence sets for the benchmark model. Confidence sets provide much more information than tests, particularly in the case of the persistence-adjusted relevance of predictive regressors (Rossi, 2005).

1. Introduction

Forecast evaluation methods and statistics allow for the ranking and comparison of models, but inference in time series contexts is related strongly to the degree of persistence. Local-to-unity and unit roots models are persistent processes that are considered commonly as models of interest; see Alquist and Kilian (2010), Baumeister et al. (Forthcoming), and Bernard et al. (2012) for some applications to commodity prices and macroeconomic data.

The forecast errors from persistent processes are known to follow non-standard distributions, see Kemp (1999) and Phillips (1998). Diebold and Kilian (2000) suggest using a unit root pre-test to choose a linear or first-difference forecasting model design. Their method provides some improvement over an arbitrary selection of the model structure, but the improvements rely on low-power tests for unit roots. The forecast evaluation tests for cointegrated and unit root models that were examined by Berkowitz and Giorgianni (2001) and Corradi et al. (2001) rely on non-standard critical values for inference.

Rossi (2005) employs a Bonferroni method, based on the work of Cavanagh et al. (1995) and Stock and Watson (1988), to account for the non-standard distribution of forecast evaluation statistics. Rossi's method focuses on a local-to-unity definition of the autoregressive parameters underlying the predictive model, which approaches the random walk forecast near the boundary even when the predictive covariates are not irrelevant. More broadly,

[*] Corresponding author at: Department of Economics and Centre for Monetary and Financial Economics (CMFE), Carleton University, Canada.

E-mail addresses: Lynda.Khalaf@carleton.ca (L. Khalaf), csaund9@uwo.ca (C.J. Saunders).

http://dx.doi.org/10.1016/j.ijforecast.2016.06.004
0169-2070/© 2016 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Bonferroni bounds are known to suffer from low power, due to their conservative nature.

Outside the context of forecast evaluation, the methods for constructing confidence intervals for autoregressive parameters at or near unity are challenged by Phillips (2014). Specifically, the confidence intervals based on local-to-unity approaches surveyed by Phillips are shown to be invalid in the stationary case, with zero asymptotic coverage probability. This includes methods of the form considered by Rossi (2005). Such discontinuities provide the motivation for this paper.

Forecasting evaluations under the assumption of stationarity have resulted in several successful bootstrap approaches; see for example Giacomini and White (2006), Hansen (2005), Harvey and Newbold (2000), Hubrich and West (2010) and White (2000). However, stationary and strongly persistent series cause a deterioration of the properties of these bootstrap methods, leading to forecast evaluation tests that are severely oversized. The poor performances of bootstrap methods for local-to-unity and unit root processes are not unexpected; Andrews (2000) and Mikusheva (2007) demonstrate that bootstrap and subsampling methods can be inconsistent.

This paper proposes a finite sample motivated approach for addressing the above problems, building on the Monte Carlo (MC) test methods of Dufour (2006). This simulation-based procedure is exact when the null distribution of the statistic considered can be simulated under the null hypothesis. Complicated finite or limiting distributions, which covers various asymptotic discontinuities, cause no concern.

Thus, our approach leads to exact *p*-values for forecast evaluation statistics, independent of the degree of persistence of the data, whether stationary, local-to-unity, or a unit root process, in spite of possible underlying non-standard, asymmetric or degenerate distributions, and regardless of whether the alternatives consist of a single or multiple models, in which case sup-type statistics are simulated. The limiting distribution of a forecast evaluation statistic does not even need to be known *a priori*.

In addition to testing, we also use the MC method to produce confidence intervals for intervening parameters. For example, for the problem analyzed by Rossi (2005), we provide simultaneous confidence sets for the persistence parameter and for the coefficient of the predictors under test. As was argued by Rossi (2005), near-unit roots confound the contributions of predictors, even when the latter are relevant. The joint MC confidence intervals that we propose provide much more information than tests, an advantage that we illustrate empirically via the well-known Meese-Rogoff puzzle.

Outside the forecasting context, the MC and MMC test methods have also been shown to solve complications that arise from unidentified nuisance parameters, see Dufour et al. (2004). The MMC method has been applied successfully in a range of areas of econometrics, with a focus on level correction or the computation of confidence sets; refer to Beaulieu et al. (2007), Beaulieu et al. (2010b), Beaulieu et al. (2013), Bernard et al. (2012), and Bernard et al. (2007). To the best of our knowledge, this is the first extension of the MMC method to the forecasting of evaluation statistics.

Our second contribution is two simulation studies that demonstrate the rejection frequency properties of our proposed approach. The first study examines the predictive ability of a random walk null model, where the forecast evaluation statistic is based on a pair of models: the random walk benchmark model and a single alternative model. The simulation design is inspired by the work of Rossi (2005). In this case, our MC test method is applied to a scaled Diebold and Mariano (1995) type statistic (denoted MSE*t*) and the encompassing statistic outlined by Clark and West (2007) (denoted ENC*t*). To maintain a focus on the forecast evaluation statistics, we implement the benchmark method defined by Rossi (2005, p. 83) as the "infeasible test", where the confidence interval for the local-to-unity parameter is assumed to include only the true value. We find that all methods provide level control in terms of size. In terms of power, the ENC*t* MC test has a higher power than either the MSE*t* MC test or the infeasible Rossi approach for all of our simulation settings. The MSE*t* MC test method dominates the infeasible Rossi method for larger sample sizes and for lower persistent processes.

The second simulation study considers a highly persistent and parsimonious benchmark model and compare it against multiple alternative models. The design is based on the work of Hubrich and West (2010). Under the null of our proposed MMC method, the rejection frequency demonstrates level control. In contrast, alternative methods, including those of Giacomini and White (2006), Hansen (2005), Harvey and Newbold (2000), Hubrich and West (2010), and White (2000), and two encompassing reality checks of Clark and McCracken (2012), are over-sized. The MMC methods demonstrate good power that improve as the sample size increases.

Lastly, we propose to invert the forecast evaluation statistic, based on the MC test method, in order to obtain exact confidence intervals on the parameters in the forecast period. Traditional approaches to obtaining confidence intervals for the benchmark model parameters assume that the in-sample estimation properties are suitable for the construction of out-of-sample bands, which may be problematic as the persistence approaches unity. In contrast, our confidence set is constructed by collecting the set of 'benchmark' models that satisfy the data. Test inversion theory has been applied to time series and for forecasting, as well as when at or approaching unity, see Cavanagh et al. (1995), Stock (1991), and Stock and Watson (1988), with the in-sample properties being exploited in each case. The idea of constructing a model confidence set was put forward by Hansen et al. (2011) in the context of a finite and discrete set of models. Our inversion produces confidence sets for the parameters of one class of alternative models, namely a class that can be defined broadly as parameterizing the "alternative" to the null hypothesis under test. To the best of our knowledge, with the notable exception of the study by Hansen et al. (2011), this is the first study to construct out-of-sample confidence sets using MC and inversion theories applied to forecast evaluation statistics.

We apply our MC test inversion to the well-known Meese-Rogoff puzzle, and confirm that the Deutsche Mark to US exchange rate fails to reject the null of a random walk. The confidence intervals based on MSE*t*, ENC*t* and our inversion do not differ much, and cover the random walk model.

The remainder of the paper is arranged as follows. Section 2 details our econometric procedure. Section 3 summarizes a small set of forecast evaluation statistics that are relevant to this study, including the Diabold-Mariano, MSE*t*, ENC*t*, and maxENC*t* statistics. Section 4 outlines the simulation studies, with Section 4.1 considering a single alternative model, and Section 4.2 allowing for multiple alternative models. Section 5 examines the Meese-Rogoff puzzle using the MC test method and the MC inversion. The final section brings together the findings of the paper.

2. Econometric setting

We consider a set of forecasting models, denoted by $\mathcal{M} = \{0, 1, \ldots, m\}$ and indexed by i, that are used to generate predictions of a time series y_{t+h}, where t is the time index and h denotes the number of time periods ahead that are predicted. For each time period, an information set, denoted Ψ_t, contains both present and past observations of relevant series. Let $Y_t = \{y_1, \ldots, y_t\}$ and $X_t = \{x_1, \ldots, x_t\}$ be the observations of the dependent and predictive regressors, respectively, leading to the information set:

$$\Psi_t \equiv \{Y_t, X_t\}\,. \tag{1}$$

We assume a flexible parametric form of the models, $g_{i,h}(\bullet)$, that allows the model parameters to be estimated based on historical information, specifically:

$$\hat{\beta}_{i,t|R} = g_{i,h}\left(y_t, \Psi_{t-1}|R\right), \tag{2}$$

where R indicates the number of prior observations used in the estimation, and $\hat{\beta}_{i,t|R}$ is a column vector with the length determined by the ith model. Predictions of the series are generated using an updated information set

$$\hat{y}_{i,t+h|t,R} = G_{i,h}\left(\hat{\beta}_{i,t|R}, \Psi_t\right), \tag{3}$$

where $G_i(\bullet)$ is a function for obtaining h-step-ahead predictions for model i.

The h-step-ahead prediction errors are given by:

$$\hat{e}_{i,t+h|t,R} = y_{i,t+h} - \hat{y}_{i,t+h|t,R}. \tag{4}$$

This estimation-prediction approach is repeated from $t = R$ to $t = T - h$, obtaining P predictions of the dependent series, where the sample size is $T = (R + h) + P + (h - 1)$. Eq. (4) forms the basis of the forecast evaluation statistics that we denote more generally as S. Selected forecast evaluation statistics and methods are presented in Section 3.

Let ξ_K represent the nuisance parameters that interfere with the distribution of the forecast evaluation statistic under the null. The set of nuisance parameters is $\xi_K = \left\{\xi_1, \ldots, \xi_k\right\}$, where k represents the total number of nuisance parameters under the null, and $\xi_K \in \Omega_K$, where Ω_K is the nuisance parameter space. The forecast evaluation statistics compare forecast errors from two or more models, but the nuisance parameter set is defined based solely on the benchmark ($i = 0$) model, for example the model parameters and/or the variance of the errors.

The Monte Carlo test procedure is summarized based on the methodology and theory presented by Dufour (2006). The maximized Monte Carlo (MMC) p-value, denoted $\hat{p}_N(S_0)$, is obtained as follows.

1. Compute the forecast evaluation statistic from the data, S_o.
2. Draw N random draws from an assumed distribution (e.g., normal or t distribution).
3. Under the null, simulate N Monte Carlo series based on the benchmark model, N random draws, and a given value of the nuisance parameters (ξ_K).
4. Compute the forecast evaluation statistic for each Monte Carlo simulation, $S_w(\xi_K)$, where $w = 1, \ldots, N$. Our notation for the simulated statistic underscores the conditioning on a given nuisance parameter value.
5. Count the number of simulated forecast evaluation statistics that equal or exceed the statistic from the data,

$$\hat{G}_N(S_0|\xi_K) = \sum_{w=1}^{N} I_{[0,\infty]}(S_w(\xi_K) \geq S_0|\xi_K), \tag{5}$$

where $I_{[0,\infty]}(S_w(\xi_K) \geq S_0|\xi_K)$ is an indicator function of the form:

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A. \end{cases} \tag{6}$$

6. The p-value conditional on the given nuisance parameter value is

$$\hat{p}_N(S_0|\xi_K) = \left(\frac{\hat{G}_N(S_0|\xi_K) + 1}{N + 1}\right). \tag{7}$$

7. Finally, the MMC p-value is obtained by maximizing over the nuisance parameter space,

$$\hat{p}_N(S_0) = \sup_{\xi_K \in \Omega_K} \left(\hat{p}_N(S_0|\xi_K)\right). \tag{8}$$

The p-value is used to define the critical region $\hat{p}_N(S_0) < \alpha$, where α is the desired significance level and satisfies $0 < \alpha < 1$. The primary benefit of using the MMC p-value procedure is that it is valid even when the asymptotic null distribution is non-standard and depends on nuisance parameters. The main requirement of the method is that the null distribution of the statistic can be simulated given a nuisance parameter set of finite dimensions. The MMC p-value will be exact in the sense that the rejection probability is less than or equal to α. The MMC method can be specified as either a one- or two-tailed test. If Ω_K is the empty set, then the benchmark model is nuisance parameter free, and Step 7 can be skipped.

The MC draws in Step 2 rely on a parametric assumption such as the standard normal, which is considered in the next sections. Coudin and Dufour (2009) show that when a MC p-value is computed given a distributional assumption and the underlying test statistic converges to any distribution (say $\hat{F}$) that does not depend on this assumption, the associated MC test will remain asymptotically valid under any set of weaker distributional assumptions for which the statistic still converges (not necessarily at the same rate) to the same $\hat{F}$; see also Beaulieu et al. (2010a).

Regarding Step 7, normal gradient-based maximization methods are ineffective at maximizing the MMC p-value because both N and $\hat{G}_N(S_0|\xi_K)$ are discrete, resulting in non-differentiable points in $\hat{p}_N(S_0|\xi_K)$. Non-gradient maximization routines must be employed, such as *simulated annealing* (Goffe et al., 1994) and *particle swarm optimization* (Kennedy and Eberhart, 1995), both of which have been shown to be effective for functions with plateaus or multiple local maxima, as well as for non-differentiable functions. Both maximization routines are used in our simulation studies in Section 4.

The MMC method may be computationally expensive for a simulation study, due to the requirement for a non-gradient optimization routine. In practice, the time required for a single call to *simulated annealing* (minutes) and *particle swarm optimization* (seconds) diminishes substantially for an empirical forecasting study.

3. Forecast evaluation statistics and bootstraps

The forecast evaluation statistics are derived from the estimated forecast errors, $\hat{e}_{i,t+h|t,R}$, where the main objective is to find the model with the lowest forecast variance or to determine whether all models have equivalent forecast variances. The Diebold and Mariano (1995) statistic (DM) is constructed as the mean squared prediction error (MSPE),

$$\hat{\sigma}^2_{i|h,R} = P^{-1}\sum_{t=R}^{T-h}\hat{e}^2_{i,t+h|t,R}, \tag{9}$$

of a benchmark model, indexed by $i = 0$, less that of an alternative model, indexed by $j = 1, \dots, m$, and is given by

$$DM_{j|h,R} = \hat{\sigma}^2_{0|h,R} - \hat{\sigma}^2_{j|h,R}. \tag{10}$$

The DM statistic is suitable for pairwise comparisons of forecasting models. Diebold and Mariano (1995) present a variant of the DM statistic that is scaled by the long-run variance. The mean squared error t-statistic (MSEt) scales the errors by the variance of the difference in squared errors:

$$MSEt_{j|h,R} = \frac{P^{-1}\sum\limits_{t=R}^{T-h}\left(\hat{e}^2_{0,t+h|t,R} - \hat{e}^2_{j,t+h|t,R}\right)}{\sqrt{V\left[\hat{e}^2_{0,t+h|t,R} - \hat{e}^2_{j,t+h|t,R}\right]}}. \tag{11}$$

An alternative loss function is based on forecast encompassing, and takes into account the covariance of the benchmark and the alternative forecast errors. The encompassing t-statistic (ENCt) is computed as

$$ENCt_{j|h,R} = \frac{P^{-1}\sum\limits_{t=R}^{T-h}\left(\hat{e}_{0,t+h|t,R}(\hat{e}_{0,t+h|t,R} - \hat{e}_{j,t+h|t,R})\right)}{\sqrt{V\left[\hat{e}_{0,t+h|t,R}(\hat{e}_{0,t+h|t,R} - \hat{e}_{j,t+h|t,R})\right]}}. \tag{12}$$

These statistics are used to compare pairs of competing models, and form the basis for statistics that compare more than two models against a benchmark model. When comparing multiple models, the null hypothesis is based on whether or not at least one alternative model exhibits a statistically lower forecast variance than the benchmark.

For multiple models, we focus on the maximum encompassing t-statistic test (maxENCt), which tests whether any alternative model beats the benchmark model, and is computed as

$$maxENCt_{h,R} = \underset{\{j\in 1,\dots,m\}}{\text{argmax}}\{ENCt_{j|h,R}\}, \tag{13}$$

for a set of $m \geq 2$ competing models. The conditional predictive ability (CPA) test put forward by Giacomini and White (2006) and the equal predictive ability (EPA) test of Harvey and Newbold (2000) are also available for testing multiple competing models. While these statistics may fall asymptotically into the family of normal distributions, their distributional properties are known to be nonstandard in the presence of high persistence and finite samples.

Several bootstrap methods have been put forward to help account for distorted critical values. Hubrich and West (2010) present a non-parametric bootstrap method which exploits Clark and West's (2007) proposition that the adjusted squared predicted errors share critical values with the standard normal distribution.

The reality check (*RCMSE*) outlined by White (2000) uses the *MSPE* as a basis for model comparisons, combining this with the block bootstrap of Politis and Romano (1994). Hansen (2005) provides an enhanced version of the reality check, applying a Student-type adjustment (*RCMSEt*) that improves upon White's version in terms of power, while retaining similar size properties. In keeping with the spirit of Clark and McCracken (2012), we include two encompassing reality check bootstrap methods that are constructed by replacing the *MSPE* with the *ENC* statistic, or the *ENCt* for a Student-type version.

4. Simulation study

We demonstrate the properties of the MMC test method using two simulation studies. The first study examines the case of a single alternative model with a random walk benchmark model, which is inspired by Rossi (2005). The second study considers multiple alternative models with a local-to-unity benchmark model.

4.1. A single alternative model

We employ the simulation design described by Rossi (2005), which is

$$y_{1,t} = \beta_1 y_{2,t} + u_{1,t}, \qquad u_{1,t} = \rho_1 u_{1,t-1} + \epsilon_{1,t},$$
$$y_{2,t} = u_{2,t}, \qquad u_{2,t} = \rho_2 u_{2,t-1} + \epsilon_{2,t},$$
$$\rho_1 = 1 - \frac{c_1}{T}, \quad \text{and} \quad \rho_2 = 1 - \frac{c_2}{T},$$

where the objective is to determine whether the regressor $(y_{2,t})$ has any predictive power. The errors are drawn independently from an uncorrelated standard normal distribution; all methods are scale-invariant, so a unit variance is appropriate both for this design and for the Monte Carlo method outlined in Section 2. Thus, draws from this model require values to be set for ρ and β_1. Following Rossi, and without loss of generality, we assume

Table 1
Rejection frequency under the null: one-step-ahead predictions, 10% level.

	$P = 40$	$P = 100$	$P = 200$
MC-ENC*t*			
$R = 40$	0.101	0.084	0.083
$R = 100$	0.087	0.097	0.100
$R = 200$	0.110	0.097	0.107
MC-MSE*t*			
$R = 40$	0.101	0.093	0.088
$R = 100$	0.088	0.090	0.097
$R = 200$	0.108	0.108	0.107
Infeasible Rossi (2005)			
$R = 40$	0.064	0.066	0.041
$R = 100$	0.047	0.052	0.052
$R = 200$	0.049	0.050	0.041

that $c_1 = c_2 = c$, and thus $\rho_1 = \rho_2 = \rho$; however, this could be relaxed, and is designed merely to simplify the simulation settings. The benchmark model is the random walk forecast, which can be derived from the model above via the boundary restriction

$$H_0 : \beta_1 = 0, \quad \text{and} \quad \rho = 1. \tag{14}$$

Nevertheless, a (right-tailed) MC p-value can be obtained easily as described above, by drawing from the random walk benchmark. The latter is nuisance-parameter free except for the scale, so assuming scale invariance of the MSE*t* and ENC*t* statistics, we obtain MC p-values for both using unit variance.

This method extends naturally to a test inversion framework, where the null hypothesis sets β_1 and ρ to any combination of relevant values and recomputes the statistic conformably. Inverting the resulting test involves retaining the parameter values that are not rejected via our proposed MC p-values, which produces exact simultaneous confidence sets for β_1 and ρ. Section 5 describes this MC inversion approach in greater detail via an empirical example using exchange and interest rates.

The Bonferroni approach of Rossi (2005) is also included in our study. Rossi focuses on the contributions of the included regressors to predictions in the context of local-to-unity models. We allow for comparability by simulating the infeasibility test, as defined by Rossi (2005, p. 83), where the local-to-unity parameter is fixed to the true value but the underlying level correction is maintained (to 5%). Thus, c is zero or the random walk in Table 1, and can be inferred from $c = T(\rho - 1)$ in Table 2. Methods used to construct confidence intervals for the local-to-unity parameter have come a long way over the last decade. Even recently, Phillips (2014) questioned the practice of constructing confidence sets under a local-to-unity setting, proving that stationary series could have zero asymptotic coverage probabilities. By simulating the infeasiblity test, we side-step this debate regarding local-to-unity parameter estimation and return our focus to the objective of forecast evaluation methods.

The rejection frequencies presented in Tables 1–3 are based on 1000 simulations and draws from the standard normal distribution, and we use $N = 199$ for the MC test method. For simulations under the null, the benchmark model is the random walk, so $\rho = 1$ or $c = 0$. Under the alternative, we present two settings $\rho = 0.9$ and $\rho = 0.99$, so the local-to-unity parameter can be inferred from $c = T(\rho - 1)$. We allow for additional predictive ability under the alternative by setting $\beta_1 = -0.05$ to parallel the simulation design of Rossi (2005), while Table 3 also sets $\beta_1 = -0.2$, to examine the effect of a stronger predictive ability. For the infeasible Rossi (2005) method, the critical values for the DM statistic are computed via 1000 simulations under the null, and as a right-tail test the critical value is the 95th percentile of these simulations.

The simulation results for the rejection frequency under the null are presented in Table 1. When applied to the MSE*t* and ENC*t* statistics, the proposed MC test method demonstrates rejection frequencies that are level correct. The infeasible Rossi (2005) method is close to the 5% level, which would be 10% once the level (5%) of the local-to-unity bounds estimation is taken into consideration. In summary, all of the methods presented provide adequate level control.

Tables 2 and 3 present the rejection frequencies under the alternative. When applied to the MSE*t* and ENC*t* statistics, our method results in good power, with similar patterns for a variety of P and R settings. Table 2 presents the case with weak predictive ability, where we find that the ENC*t* and MSE*t* MC test methods dominate the infeasible test. For the stronger predictive ability presented in Table 3, the ENC*t* MC test method dominates the other approaches. The MSE*t* MC test method dominates the infeasible test of Rossi for $\rho = 0.9$, but this dominance is only for longer time series with higher levels of persistence ($\rho = 0.99$).

4.2. *Multiple alternative models*

The simulation study of Hubrich and West (2010) employs a VAR data generation process (DGP) where the null model parameter is 0.5. This study uses the same framework but increases the null model parameter to 0.99, which introduces a much higher degree of persistence into the DGP. We examine the possibility that the bootstrap method of Hubrich and West (2010) may be biased in the presence of highly persistent data. This may have implications for their US inflation analysis, since their results indicate a highly persistent process, due to changes in the mean and volatility of price series in general, see Hendry and Hubrich (2011) and Stock and Watson (2007).

This simulation study determines whether or not a subset of the disaggregate series provides additional information when forecasting the aggregate, when all series are highly persistent. Consider the aggregate series y_t as the sum of D disaggregate components, $y_{d,t}$, where $d = 1, \ldots, D$, so that

$$y_t = \sum_{d=1}^{D} y_{d,t}. \tag{15}$$

The benchmark model is

$$y_t = \mu_0 + \phi_{0,0,t|R} y_{t-1} + e_{0,t}, \tag{16}$$

which is nested by m alternative models (say $m = 2$ or $m = 4$) of the following form,

$$y_t = \mu_j + \phi_{0,j,t|R} y_{t-1} + \beta_{1,j,t|R} y_{j,t-1} + e_{j,t}. \tag{17}$$

Table 2
Rejection frequency under the alternative: one-step-ahead predictions, 10% level and $\beta = -0.05$.

	$\rho = 0.9$			$\rho = 0.99$		
	$P = 40$	$P = 100$	$P = 200$	$P = 40$	$P = 100$	$P = 200$
MC-ENC*t*						
$R = 40$	0.272	0.441	0.576	0.120	0.108	0.126
$R = 100$	0.393	0.717	0.924	0.135	0.140	0.161
$R = 200$	0.530	0.881	0.995	0.146	0.167	0.225
MC-MSE*t*						
$R = 40$	0.260	0.416	0.563	0.118	0.122	0.121
$R = 100$	0.309	0.570	0.839	0.135	0.140	0.153
$R = 200$	0.346	0.618	0.917	0.141	0.172	0.194
Infeasible Rossi (2005)						
$R = 40$	0.055	0.072	0.066	0.065	0.067	0.057
$R = 100$	0.072	0.063	0.060	0.070	0.051	0.073
$R = 200$	0.061	0.060	0.065	0.061	0.074	0.080

Table 3
Rejection frequency under the alternative: one-step-ahead predictions, 10% level and $\beta = -0.20$.

	$\rho = 0.9$			$\rho = 0.99$		
	$P = 40$	$P = 100$	$P = 200$	$P = 40$	$P = 100$	$P = 200$
MC-ENC*t*						
$R = 40$	0.496	0.693	0.868	0.288	0.385	0.510
$R = 100$	0.617	0.935	0.993	0.327	0.567	0.666
$R = 200$	0.646	0.955	1.000	0.315	0.513	0.778
MC-MSE*t*						
$R = 40$	0.443	0.665	0.834	0.244	0.330	0.447
$R = 100$	0.439	0.820	0.967	0.260	0.466	0.587
$R = 200$	0.412	0.736	0.970	0.249	0.381	0.620
Infeasible Rossi (2005)						
$R = 40$	0.150	0.242	0.272	0.172	0.222	0.289
$R = 100$	0.237	0.322	0.403	0.234	0.331	0.480
$R = 200$	0.249	0.349	0.432	0.264	0.394	0.512

The DGP for the disaggregates is a VAR(1) with a $D \times D$ matrix of autoregressive parameters, Φ, a mean vector $\mu \equiv (u_1, \ldots, u_D)'$, where $\mu_d = 1$ for all d, and zero mean i.i.d. disturbances $U_t \equiv (u_{1,t}, \ldots, u_{D,t})'$, giving:

$$Y_t \equiv (y_{1,t}, \ldots, y_{D,t})' = \mu + \Phi Y_{t-1} + U_t. \tag{18}$$

For the simulations, the disturbances are drawn from an i.i.d. normal distribution, though using a Student's t-distribution with low degrees of freedom results in notionally similar results. When determining the size of each of the test statistics with persistent data, we assume a common value of $\phi = 0.99$ for the diagonal elements of Φ, and $D = 3$; specifically:

$$\Phi = \begin{bmatrix} 0.99 & 0 & 0 \\ 0 & 0.99 & 0 \\ 0 & 0 & 0.99 \end{bmatrix}. \tag{19}$$

Furthermore, each disaggregate of y_t follows an AR(1) process. Since y_t is the arithmetic sum of the disaggregates, it too will follow an AR(1) process with a lag parameter value, $\phi = 0.99$.

It is assumed in the power simulations that at least one of the disaggregate components Granger-causes the aggregate in Eq. (18), and as such Φ is updated to:

$$\Phi = \begin{bmatrix} 0.99 & -0.008 & 0 \\ 0.2 & 0.5 & 0 \\ 0 & 0 & 0.99 \end{bmatrix}. \tag{20}$$

This design allows for low persistence in one of the series, and high persistence in the other disaggregates.

A similar design is used when the number of alternative models is expanded to $m = 4$ and the number of disaggregates also expands to $D = 4$. In this case, the size simulations in the matrix in Eq. (19) are replaced with $\Phi = 0.99 I_D$. In the power simulations, the matrix in Eq. (20) is replaced with:

$$\Phi = \begin{bmatrix} 0.99 & -0.008 & 0 & 0 \\ 0.2 & 0.5 & 0 & 0 \\ 0 & 0 & 0.99 & 0 \\ 0 & 0 & 0 & 0.99 \end{bmatrix}. \tag{21}$$

For both the size and power simulations, 1000 replications of the null DGP are used to compute the p-values associated with each test, and other test-specific settings follow.

The MMC procedure allows for some flexibility; for the simulation study, $N = 99$, and the simulated annealing program was edited to reduce computation time. Specifically, if the evaluation of the p-value exceeded the nominal size (α), then the simulated annealing procedure was halted and returned an indicator to retain the null hypothesis. The maxENC*t* test statistic from the null DGP can be shown by simulation to be invariant in terms of the standard deviation of the error term draws (U_t), the constant term (μ), and the initial value of the disaggregates ($y_{d,0}$).

Table 4
Rejection frequency under the null: one-step-ahead predictions, 10% level.

		$m = 2$			$m = 4$		
		$P = 40$	$P = 100$	$P = 200$	$P = 40$	$P = 100$	$P = 200$
$R = 40$	HW2010	0.183	0.258	0.332	0.189	0.265	0.346
	CPA	0.173	0.200	0.238	0.240	0.226	0.240
	EPA	0.339	0.382	0.425	0.275	0.309	0.339
	RCMSE	0.060	0.027	0.005	0.097	0.045	0.018
	RCMSEt	0.060	0.028	0.005	0.093	0.037	0.017
	RCENC	0.223	0.287	0.348	0.326	0.380	0.450
	RCENCt	0.237	0.297	0.369	0.315	0.370	0.442
$R = 100$	HW2010	0.168	0.161	0.226	0.193	0.183	0.233
	CPA	0.232	0.166	0.174	0.325	0.227	0.208
	EPA	0.413	0.352	0.365	0.364	0.325	0.297
	RCMSE	0.115	0.047	0.031	0.167	0.088	0.056
	RCMSEt	0.121	0.050	0.030	0.178	0.089	0.057
	RCENC	0.217	0.229	0.280	0.278	0.288	0.334
	RCENCt	0.227	0.243	0.286	0.276	0.285	0.331
$R = 200$	HW2010	0.132	0.143	0.134	0.188	0.202	0.158
	CPA	0.210	0.195	0.168	0.324	0.289	0.228
	EPA	0.389	0.373	0.353	0.372	0.398	0.336
	RCMSE	0.113	0.081	0.035	0.189	0.145	0.088
	RCMSEt	0.129	0.085	0.038	0.215	0.164	0.089
	RCENC	0.188	0.170	0.163	0.269	0.274	0.281
	RCENCt	0.205	0.185	0.170	0.271	0.279	0.267

Note: the statistics are from Hubrich and West (2010) [HW2010], Giacomini and White (2006) [CPA], Harvey and Newbold (2000) [EPA], White (2000) [RCMSE], Hansen (2005) [RCMSEt], and Clark and McCracken (2012) [RCENC and RCENCt].

The only nuisance parameter determined is the common value ϕ.

The Hubrich and West (2010) bootstrap method (HW2010) is implemented in the same manner as in their paper, where the p-value is computed based on the ordinal rank of the maxENCt statistic in a set of 50,000 bootstrap replications. The precise p-value was not required for the simulation study, so the replications were terminated when the decision on the null hypothesis could not be reversed. Their approach was replicated in this study in order to isolate the effect of highly persistent data on the p-value.

The critical values for the CPA test are taken from the standard $\chi^2(m)$ at the 10% level. The critical values for the EPA test are taken from the standard $F(m-1, P-m+1)$ distribution at the 10% level.

All four reality check simulations, *RCMSE*, *RCMSE*t, *RCENC* and *RCENC*t, use 1000 stationary bootstrap replications and a geometric mean block size of two, which are the same assumptions used by both Hubrich and West (2010) and White (2000).

4.2.1. Simulation results

Tables 4 and 5 present the simulation results for forecasting models that are nested at the unit root boundary. The rejection frequency under the nulls of various forecast evaluation methods are given in Table 4. The encompassing-based methods, HW2010, CPA, EPA, RCENC and RCENCt, are almost universally oversized, with modest improvements as R increases. The two MSPE-based reality checks (RCMSE and RCMSEt) appear to achieve level control for specific combinations of in-sample and out-of-sample sizes. However, these rejection frequencies decrease substantially as P rises.

The HW2010 approach is oversized for all combinations of R, P and m examined in this study, and the rejection frequencies rise as P increases. They also increase with m, but this change is modest for the small set of models examined in this study. As R increases, the severity of the oversized results in their nonparametric bootstrap method diminishes, to the point that it is almost correctly sized with $R = 200$ and $m = 2$, and all of the values of P tested. It is not universally apparent that the properties shown by Hubrich and West (2010) can be achieved in this local-to-unity setting, even with a very large R.

The CPA method's rejection frequency results are considerably oversized for all combinations of R, P and m examined here. Unlike the findings of Hubrich and West (2010), the CPA results do not display any observable patterns. For $m = 2$, the rejection frequency results generally improve as R increases, but such is not the case for $m = 4$, where the patterns are erratic. Increasing the number of alternative models uniformly increases the rejection frequencies under the null, but the relative increase varies. When $m = 4$ and $R > 40$, an increase in P results in an improvement in rejection frequencies, however, with the $m = 2$ results, this pattern is only apparent for $R = 100$ and $R = 200$. Hubrich and West (2010) found that the rejection frequencies were higher under the null of the CPA method than under that of the HW2010, but such was not the case for the results shown in Table 4. Although this is somewhat surprising, since CPA is a two-tailed test and HW2010 is one-tailed test, it is not improbable, since the latter statistic takes the supremum.

The commonly utilized reality check, RCMSE, is normally a very conservative test, with suitable level control. However, the rejection frequencies shrink as P increases, but rise as m and R increase. Similar properties are observed for RCMSEt. These tests appear to have reasonable

Table 5
Maximized Monte Carlo rejection frequency: one-step-ahead predictions, 10% level.

	$m = 2$			$m = 4$		
	$P = 40$	$P = 100$	$P = 200$	$P = 40$	$P = 100$	$P = 200$
Under the null						
$R = 40$	0.076	0.089	0.092	0.080	0.091	0.071
$R = 100$	0.068	0.067	0.081	0.086	0.076	0.084
$R = 200$	0.067	0.061	0.063	0.087	0.079	0.065
Under the alternative						
$R = 40$	0.362	0.588	0.787	0.192	0.310	0.465
$R = 100$	0.337	0.645	0.851	0.194	0.344	0.543
$R = 200$	0.332	0.601	0.834	0.177	0.323	0.504

Rejection frequencies based on the maxENC*t* statistic.

level control when $P/R = 1$, but over-reject for $P < R$ and under-reject for $P > R$.

The MSPE-based statistics used in the common reality check methods are inappropriate for nested models, due to non-standard distributions. The encompassing reality checks provide more sensible responses to increases in R, P and m, but the tests are heavily oversized under all conditions examined. The ENC-based reality check bootstrap outlined by Clark and McCracken (2012) displays good level control properties for a set of nested models, but their simulation setting was stationary and did not consider a case that was close to the unit root boundary.

The EPA test is heavily oversized under all conditions examined. The empirical size results have the following properties: the size decreases as m increases, but increases in P and/or R fail to show any systematic effect on the size.

These methods do not provide level control, as they are based on standard asymptotic critical values or bootstrap methods that become inconsistent at the boundary. The rejection frequencies for the leading methods under the alternative will be uninformative without level control.

Table 5 presents the rejection frequencies of the MMC method applied to the maxENC*t* statistic. The rejection frequencies under the null are modestly conservative for low values of R and generally more conservative as R increases, and are level correct in the sense of Dufour (2006). The simulations indicate that the MMC method is robust to increases in the number of alternative models. It also exhibits good level control, as P increases for a given R. The MMC method provides simulations under the conditions of finite values of R, P, and T, with the former two being at the discretion of the analyst.

The rejection frequency of the MMC method under the alternative (power) is promising even under the highly persistent process examined. As the number of out-of-sample observations (P) increases, the rejection frequency increases under all simulation settings. As the number of alternative models increases, we observe a reduction in the rejection frequency for all combinations of R and P considered in our study. The number of in-sample observations has a marginal and erratic effect on the rejection frequencies under the alternative. In our simulation design, the only nuisance parameter is the lagged dependent coefficient. The supremum of the MMC method ensures that the resulting p-value is not conditional on nuisance parameters, which appears to promote a degree of independence from the in-sample observations.

5. Inverting the Meese-Rogoff puzzle

Simply put, the Meese-Rogoff puzzle is the fact that a random walk model provides better forecasts of real exchange rates (ϵ) than a model based on fundamentals. Thus, a structural forecasting model of the quarterly real Deutsche Mark–US Dollar exchange rate from 1973 to 1998 is defined as

$$\epsilon_{t+h} = \beta_0 + \rho\epsilon_t + \beta_1 X_t + \nu_{t+h}, \tag{22}$$

where X is the real interest rate differential between the two countries. The benchmark forecasting model is the random walk for forecasting the real exchange rates:

$$\epsilon_{t+h} = \epsilon_t + \nu_{t+h}. \tag{23}$$

Because of scale-invariance, MC random walk draws with a unit variance are appropriate. The MC test is based on one-step-ahead predictions at the 5% level, and all parameters for the alternative model are estimated via OLS. The number of in-sample observations is a rolling window of 60 quarters ($R = 60$ and $P = 42$).

Our Monte Carlo test method fails to reject the hypothesis of a random walk, with p-values of 0.2125 and 0.225 based on the MSE*t* and ENC*t* statistics, respectively.

Our Monte Carlo test method extends naturally to a Monte Carlo inversion framework for constructing an exact confidence set for the benchmark model parameters. Our MC inversion method sets the parameters ρ, β_0, and β_1, to known values $\bar{\rho}$, $\bar{\beta}_0$, and $\bar{\beta}_1$, respectively, so that the null and alternative hypotheses are redefined as

$$\begin{cases} H_0 : \rho = \bar{\rho} \text{ and } \beta_0 = \bar{\beta}_0 \text{ and } \beta_1 = \bar{\beta}_1 \\ H_A : \rho \neq \bar{\rho} \text{ or } \beta_0 \neq \bar{\beta}_0 \text{ or } \beta_1 \neq \bar{\beta}_1. \end{cases} \tag{24}$$

The random walk model is clearly a special case, where $\rho = 1$, $\beta_0 = 0$ and $\beta_1 = 0$.

The MC test method is applied to the benchmark model or null defined by Eq. (24). The forecast evaluation statistic from the data is obtained by imposing the null for the benchmark model, and the alternative model is estimated. The Monte Carlo series and statistic are constructed by imposing the null model. Screening over a reasonable range of parameter values, the confidence set is constructed by retaining all points that are not rejected under the null at the desired significance level.

Our approach to constructing this confidence set is unique, since we are inverting a forecast evaluation statistic that is based on out-of-sample predictions, whereas

Table 6
Meese-Rogoff inversion: Deutsche Mark–US Dollar exchange rate.

	ρ_L	ρ_U	$\rho = 1$	
	$\beta_1 = 0$		$\beta_{1,L}$	$\beta_{1,U}$
Inv-MSE*t*	0.9804	1.0175	−0.0078	0.0088
Inv-ENC*t*	0.9833	1.0175	−0.0068	0.0058

most inversion methods invert an in-sample statistic. To the best of our knowledge, our approach is the first to use inversion of a statistic based on an out-of-sample series to obtain inference on model parameters. Further, the MC test method is exact in the sense of Dufour (2006), and the exactness is conferred to the confidence set constructed via our inversion.

We apply our Monte Carlo inversion to the Meese-Rogoff puzzle for the Deutsche Mark–US Dollar exchange rate, and the results are presented in Table 6. The forecast evaluation statistics that we invert are the MSE*t* and ENC*t* statistics. Our primary objective is to determine whether a model based on fundamentals (real interest rate differentials) can predict the real exchange rate better than the random walk. Table 6 presents the confidence interval for a parameter estimate conditional on a given (meaningful) value of the other parameter. Thus, the first two columns are the upper (ρ_U) and lower (ρ_L) limits of the lag-dependence parameter estimate conditional on the real interest rate differential having no predictive ability. The last two columns are the upper ($\beta_{1,U}$) and lower ($\beta_{1,L}$) limits of the predictive ability parameter where the model has a unit root. As expected, the null of a random walk is an interior point for the confidence intervals that support the *p*-values presented above, and the spread of these intervals is similar for the MSE*t* and the ENC*t* statistics.

6. Final remarks

The simulation results show that forecast evaluation methods that rely on asymptotic and bootstrap-based critical values do not achieve level control with finite and highly persistent data. In the worst case, the rejection frequencies under the alternative can be spurious. Under the same conditions, the MMC method provides both level control and good power.

The versatility of our proposed method yields level-correct inference under non-standard asymptotics, including degenerate asymptotic null distributions. Concretely, the MMC method is well-suited for parsimonious null models with a finite-dimensional set of nuisance parameters. While this paper focuses on a simple (though popular) forecasting model, the method could be extended to more complex models as long as the design of such models allows them to be simulated. Possible interesting extensions of the method using alternative null models could include time-varying-parameter models, like those of Harvey and Luati (2014), or even jump-diffusion models.

References

Alquist, R., & Kilian, L. (2010). What do we learn from the price of crude oil futures? *Journal of Applied Econometrics*, *25*(4), 539–573.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, *68*(2), 399–405.

Baumeister, C., Kilian, L., & Zhou, X. (2016). Are product spreads useful for forecasting? An empirical evaluation of the Verleger hypothesis. *Macroeconomic Dynamics*, Forthcoming.

Beaulieu, M.-C., Dufour, J.-M., & Khalaf, L. (2007). Multivariate tests of mean–variance efficiency with possibly non-Gaussian errors. *Journal of Business & Economic Statistics*, *25*(4).

Beaulieu, M.-C., Dufour, J.-M., & Khalaf, L. (2010a). Asset-pricing anomalies and spanning: Multivariate and multifactor tests with heavy-tailed distributions. *Journal of Empirical Finance*, *17*(4), 763–782.

Beaulieu, M.-C., Dufour, J.-M., & Khalaf, L. (2010b). Multivariate residual-based finite-sample tests for serial dependence and ARCH effects with applications to asset pricing models. *Journal of Applied Econometrics*, *25*(2), 263–285.

Beaulieu, M.-C., Dufour, J.-M., & Khalaf, L. (2013). Identification-robust estimation and testing of the zero-beta CAPM. *The Review of Economic Studies*, *80*(3), 892–924.

Berkowitz, J., & Giorgianni, L. (2001). Long-horizon exchange rate predictability? *Review of Economics and Statistics*, *83*(1), 81–91.

Bernard, J.-T., Dufour, J.-M., Khalaf, L., & Kichian, M. (2012). An identification-robust test for time-varying parameters in the dynamics of energy prices. *Journal of Applied Econometrics*, *27*(4), 603–624.

Bernard, J.-T., Idoudi, N., Khalaf, L., & Yélou, C. (2007). Finite sample multivariate structural change tests with application to energy demand models. *Journal of Econometrics*, *141*(2), 1219–1244.

Cavanagh, C. L., Elliott, G., & Stock, J. H. (1995). Inference in models with nearly integrated regressors. *Econometric Theory*, *11*(05), 1131–1147.

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

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., & Olivetti, C. (2001). Predictive ability with cointegrated variables. *Journal of Econometrics*, *104*(2), 315–358.

Coudin, E., & Dufour, J.-M. (2009). Finite-sample distribution-free inference in linear median regressions under heteroscedasticity and non-linear dependence of unknown form. *The Econometrics Journal*, *12*(s1), S19–S49.

Diebold, F. X., & Kilian, L. (2000). Unit-root tests are useful for selecting forecasting models. *Journal of Business & Economic Statistics*, *18*(3), 265–273.

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

Dufour, J.-M. (2006). Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, *133*(2), 443–477.

Dufour, J.-M., Khalaf, L., Bernard, J.-T., & Genest, I. (2004). Simulation-based finite-sample tests for heteroskedasticity and arch effects. *Journal of Econometrics*, *122*(2), 317–347.

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

Goffe, W. L., Ferrier, G. D., & Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, *60*(1), 65–99.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4).

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Harvey, A., & Luati, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association*,.

Harvey, D., & Newbold, P. (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics*, *15*(5), 471–482.

Hendry, D. F., & Hubrich, K. (2011). Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business & Economic Statistics*, *29*(2).

Hubrich, K., & West, K. D. (2010). Forecast evaluation of small nested model sets. *Journal of Applied Econometrics*, *25*(4), 574–594.

Kemp, G. C. (1999). The behavior of forecast errors from a nearly integrated AR (1) model as both sample size and forecast horizon become large. *Econometric Theory*, *15*(02), 238–256.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of IEEE international conference on neural networks, ICNN95*.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica*, *75*(5), 1411–1452.

Phillips, P. C. (1998). Impulse response and forecast error variance asymptotics in nonstationary VARs. *Journal of Econometrics*, *83*(1), 21–56.

Phillips, P. C. (2014). On confidence intervals for autoregressive roots and predictive regression. *Econometrica*, *82*(3), 1177–1195.

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

Rossi, B. (2005). Testing long-horizon predictive ability with high persistence, and the Meese–Rogoff puzzle*. *International Economic Review*, *46*(1), 61–92.

Stock, J. H. (1991). Confidence intervals for the largest autoregressive root in US macroeconomic time series. *Journal of Monetary Economics*, *28*(3), 435–459.

Stock, J. H., & Watson, M. W. (1988). Testing for common trends. *Journal of the American Statistical Association*, *83*(404), 1097–1107.

Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and banking*, *39*(s1), 3–33.

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