Forecast Model Evaluation in Small-Sample Persistent Processes: A Simulation Study

by

Charles Saunders

Bachelor of Science, Queen's University, 1997

Bachelor of Arts, Queen's University, 2001

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs

in fulfillment of the requirements for the degree of

Master of Arts

in

Economics

Carleton University

Ottawa Ontario

Your file Votre référence
ISBN: 978-0-494-81623-3
Our file Notre référence
ISBN: 978-0-494-81623-3

Abstract

This paper examines the effect of highly persistent processes on methods of evaluating the out-of-sample predictions of nested models. The non-parametric bootstrap method of Hubrich and West (Journal of Applied Econometrics, 2010) is modestly undersized for stationary processes, but I find that their method becomes oversized as the process examined approaches the unit root boundary. I also find size distortions in several other leading forecast evaluation procedures [e.g., Giacomini and White (Econometrica, 2006), Hansen (Journal of Business & Economic Statistics, 2005), Harvey and Newbold (Journal of Applied Econometrics, 2000), and White (Econometrica, 2000)]. I use simulation-based techniques to demonstrate that the Maximized Monte Carlo (MMC) method of Dufour (Journal of Econometrics, 2006) corrects for the over-rejection of the null even with highly persistent processes and small sample sizes. The MMC method exhibits good power properties, although in this study, the MMC procedure becomes more conservative as sample sizes increase.

Acknowledgements

I would like to thank my thesis supervisor, Professor L. Khalaf, for her support, guidance, and enthusiasm, and my family for their understanding and patience over the past year. At Informetrica Limited I would like to thank Mike C. McCracken and Carl A. Sonnen for allowing me the flexibility in hours to continue my education, finally Nina Gormanns, Can Hakyemez, Dr. Edward Hughes, Michelle Lasota, and Abeer Reza for their review and comments on previous drafts.

Table of Contents

List of Tables

List of Illustrations

List of Appendices

1 Introduction

This paper examines the effect of highly persistent processes on encompassing tests of out-of-sample predictions. Bootstrap procedures can become inconsistent, or even invalid, under certain conditions like near the unit root boundary [cf. Andrews (2000) and Mikusheva (2007)]. This indicates that bootstrap procedures on out-of-sample prediction tests [Giacomini and White (2006), Hubrich and West (2010), and White (2000)] in the presence of highly persistent data can also be inconsistent. This paper demonstrates that the maximized Monte Carlo (MMC) method of Dufour (2006) is an option to correct for this bootstrap inconsistency.

Comparison and ranking of forecast model options is a natural extension of developing forecasting models. A general method of comparing models for their forecasting ability is to divide the sample into two distinct periods: in-sample and out-of-sample periods. The in-sample period is used to identify model parameters that are subsequently used to generate the out-of-sample predictions. Prediction errors are the difference between the predicted value and actual value, and the series of prediction errors are the basis for evaluating forecasting models. There are several methods for manipulating these prediction errors to assess predictions between models: the two most common are absolute and squared differences; this study will focus on tests related to the latter.

Two streams of tests have developed in the last 15 years. The first class of out-of-sample prediction tests are related to non-nested models, which include, but are not limited to, tests developed by Diebold and Mariano (1995), West (1996), Harvey, Leybourne and Newbold (1998), White (2000), and Hansen (2005). To generalize this

class of tests, they compare the difference of the mean squared prediction errors (*MSPE*) between two models. Commonly, the tests measure the benchmark model *MSPE* less the *MSPE* of an alternative model, so a positive difference indicates that the alternative model has better predictive ability over the benchmark, and vice versa. These tests differ from the second class of tests based on nested models, examined by West (1996), Clark and McCracken (2001) and Clark and West (2007), which is described as the *MSPE* of the benchmark model less the prediction error covariance of the benchmark and alternative models. This type of test is an encompassing test (denoted as *ENC*), meaning that no additional information is provided by the alternative model over the parsimonious benchmark model. This study presents a small set of nested models, so it is this second class of test statistics that are used primarily in this study. The test statistic of interest is a t-test (*ENC-t*) put forth by Harvey, Leybourne and Newbold (1998), examined in detail by West (1996) and Clark and West (2007).

West (1996) showed that both classes of tests, specifically the t-statistic type (*MSPE-t* and *ENC-t*), when applied to non-nested models converge asymptotically to the standard normal distribution. For nested models, the predicted errors are asymptotically the same: West (1996) and Clark and McCracken (2001) showed that both test types converge in distribution to specific functions of Brownian motion, and the mean and the variance of both tests converge at the same rate but to nonstandard distributions (West, 1996). Clark and West (2007) found that, for nested models, the critical values for the *ENC-t* statistic could be approximated by the standard normal critical values, erring on the moderately conservative side, but this was not true for *MSPE-t*. It was also observed that the

distributions were affected by the number of regression observations (R) and the number of prediction periods (P) in finite samples.

The Clark and West (2007) findings were used by Hubrich and West (2010), to develop a nonparametric bootstrap method to examine a small set of nested models. In brief, using the variance-covariance matrix of the data generated *ENC* series to randomly draw many replicated *ENC* series from the multivariate normal distribution. The replicated series are used to compute a set replicated *ENC-t* statistics from which the rank of the original *ENC-t* in the ordered replication set determines its p-value. Hubrich and West (2010) found that their bootstrap procedure was modestly conservative, consistent with Clark and West (2007), in their first order autoregressive data generation process based simulations [DGP-AR(1)].

It has been well documented that bootstrap methods can become inconsistent in the presence of unknown nuisance parameters. Andrews (2000) showed that several bootstrap methods, when a parameter is on the unit root boundary, can be inconsistent. Mikusheva (2007) also showed that certain types of bootstrap methods are not consistent in the presence of unit roots. Pre-testing for a unit root test can yield mixed results, due to low power of many of the standard unit root tests in small samples.

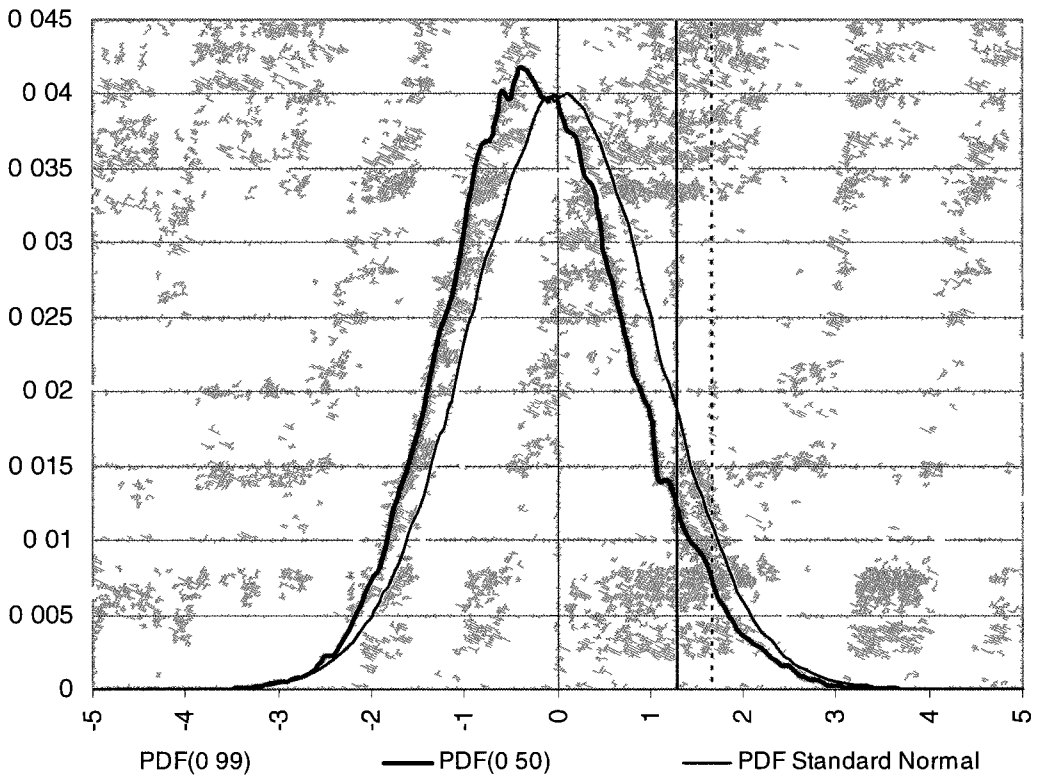
The following figures examine the simulation effects of higher persistence of the DGP-AR(1) process (50,000 replications). In figure 1,[1] the normal critical values (vertical lines) are appropriate for the level of persistence used in the Hubrich and West

---

[1] Figures 1 and 2 present the PDF curves: at the request of a reviewer, an appendix presents the complementary CDF curves.

(2010) study, where R=100 and P=100. As the lag dependent parameter approaches the boundary (0 99 in these simulations) the standard normal critical values become oversized.[2] Specifically, the standard normal 0.10 and 0.05 critical values correspond to an oversized *ENC-t* statistic of 0.152 and 0 088, respectively, in the highly persistent process simulations.

Figure 1: Simulated PDF curves for ENC-t at different levels of persistence



Notes All distributions presented are simulated The Standard Normal PDF is generated from 500,000 draws from the normal distribution The ENC-t distributions are based on 50,000 simulations of the ENC-t statistic for a single alternative model versus a benchmark model that is the null DGP The simulation design is outlined below in section 3, with 0 99 and 0 50 representing the lag dependent coefficient in the null DGP, rolling window of 100 regression observations with 100 one-step-ahead

[2] Definition and discussion of size and power are provided in an appendix

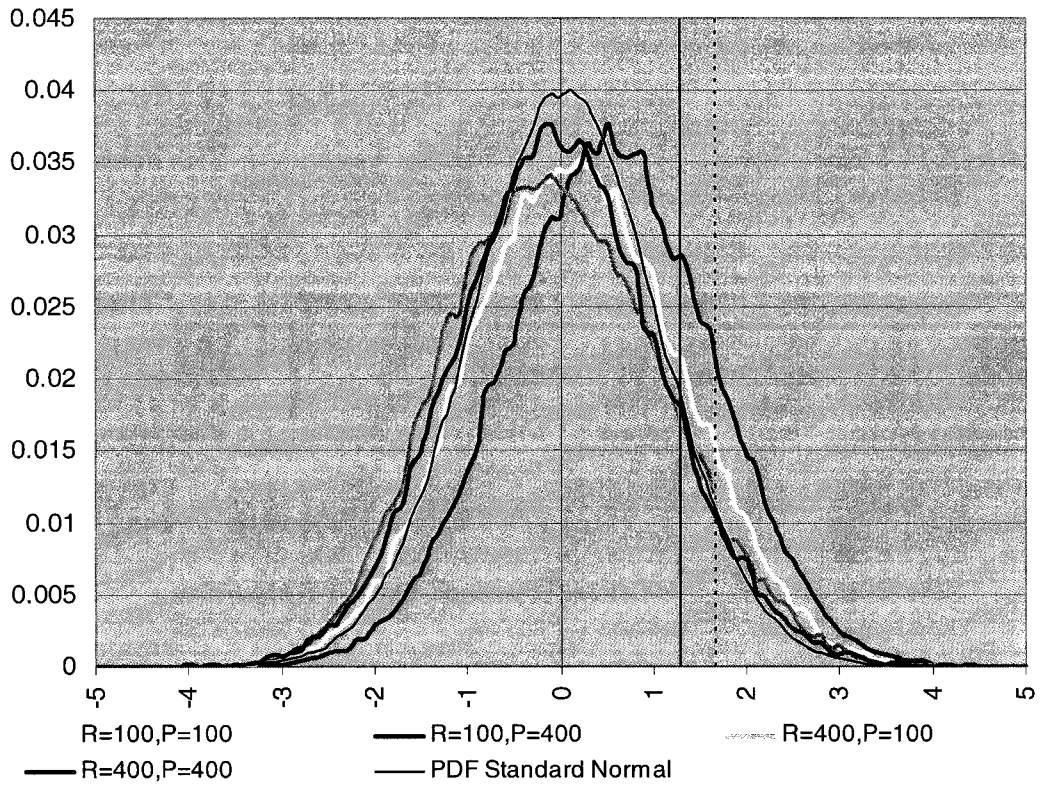
predictions. Solid and dashed vertical line are the right-tailed 90^{th} and 95^{th} percentile cut-offs for the standard normal distribution, respectively.

Figure 2 illustrates the effect of the varying P and R, both separately and individually, which are consistent with the asymptotic properties identified in West (1996). An increase in P relative to R shifts the simulated mean to the right and increases the simulated variance, which increases the number of rejections based on the standard normal critical values, again 0.10 and 0.05, for the *ENC-t* statistic to 0.231 and 0.141, respectively. As R increases relative to P the simulated mean shifts to the left, correcting for the higher simulated variance. Thus, a greater number of regression periods are required to correct for highly persistent processes. As both R and P increase, maintaining a fixed ratio (in these simulations $P/R = 1$), both the simulated mean and variance are reduced with the net result of relatively fewer rejections of the null hypothesis.

Figures 1 and 2 demonstrate that approximating the *ENC-t* critical values using the standard normal critical values in highly persistent processes will result in over rejection of the null. Bootstrap methods that rely on either draws from an assumed standard normal or mean zero probability density function will likely over reject in small sample highly persistent data.

The following simulation study examines the use of the maximized Monte Carlo (MMC) method of Dufour (2006) to correct for the over rejection of the null, so as to provide sufficient size control in forecast model assessment for finite, highly persistent data processes. The MMC method does not rely on the standard limiting distributions of the test statistic, only that all nuisance parameters of the statistic have been identified and the limiting distribution can be simulated in the presence of these nuisance parameters.

Figure 2: Simulated ENC-t PDF varying P and R, highly persistent processes



Notes: See Figure 1 notes for simulation specifics. The distributions are based on simulations where the lag dependent parameter is 0.99, and the number of regression observations (R) and one-step-ahead predictions (P) vary.

The remainder of the paper is as follows: section 2 describes the computation of the statistic *ENC-t* and *maxENC-t*, also detailing the MMC method and the comparator bootstrap methods including Hubrich and West (2010), Giacomini and White (2007), Harvey and Newbold (2000), and White (2000). Section 3 outlines the simulation design, and section 4 provides the simulation results. The final section brings together the findings of the paper and discusses opportunities for future work.

2 Econometric Procedure

The econometric procedure applied in this study is an extension of the one outlined in Hubrich and West (2010), with the addition of examining models that are nested near the boundary. The structure is as follows: there is a null, or benchmark, model and m alternative and competing models, each of which is used to predict the objective, y_t. It is assumed that the entire set of alternative models nest the null model and it is not required, or forbidden, that any individual alternative model nests any of the other alternative models. The set of all models will be denoted with i, where i=0, 1 … m, and the set of alternative models are denoted with j, where j=1, … , m. The number of alternative models will be kept at two for descriptive purposes only (m=2), but the methodologies presented below can easily be expanded to allow for more alternative models (m>2).

Note that in the Hubrich and West (2010) not-for-publication appendix, the authors present their simulated critical values for the case involving two alternative models, through the range of *ENC* covariance from -1 to 1. The p-values simulated in this paper are based on their simulation procedure, not their tabulated critical values.

The models considered are forecasting models, so a given prediction, y_{t+1}, is based on a set of historical information used to estimated parameter(s). There are two common types of information sets, rolling and recursive windows. The rolling window is based on a fixed number of observations used to estimate the model parameters, while the number of observations increases in the case of the recursive window information set. Let the information set, $\Psi_{i,t}$ at time t for model i, consist of the current and past observations of the variable of interest, y_t, and relevant independent variables known at time t, $X_{i,t}$. In the case of the recursive window, all information is used including information back to

t=0. The rolling window conditions the number of historical periods in the following manner:

$$\Psi_{i,t|R} \equiv \left\{ y_t, \ldots, y_{t-R}, X_{i,t}, \ldots, X_{i,t-R-1} \right\}. \quad (1)$$

This paper examines a small set of nested models, so appropriately focuses on *ENC*-type statistics, specifically the *ENC-t* test, examined by Clark and West (2007) and West (1996). Simulations using alternative loss functions are also examined and presented in the discussion below. One specific variant is the *maxENC-t* used to examine multiple alternative models (m>1) against a single benchmark model by comparing the maximum *ENC-t* statistic computed from the alternatives and comparing it to the critical values.

Several distinct procedures are used to evaluate the null model: the maximized Monte Carlo method (MMC); the non-parametric bootstrap method (HW) presented by Hubrich and West (2010); the conditional predictive ability (CPA) method outlined by Giacomini and White (2007); the test of equal predictive ability (EPA) from Harvey and Newbold (2000); the Reality Check (RCMSE) from White (2000); Hansen (2005) student-adjusted Reality Check (RCMSEt), an *ENC* variant of White's Reality Check (RCENC), and finally, an ENC variant of Hansen's student-adjusted Reality Check (RCENCt).

It is important to note at this point that the tests differ in their null and alternative hypotheses. Some of the tests are t-statistic type tests (MMC, HW, RCMSE, RCMSEt, RCENC, and RCENCt) which allow for right-tailed critical regions, whereas the other tests are χ^2-distributed and F-distributed and as such are effectively two-tailed statistics. This distinction is important because a right-tailed test determines if there exists an alternative model that is significantly better than the benchmark model. The two-tailed test identifies whether there exists a model, including the benchmark model, which

surpasses all other models. The hypotheses are presented formally in subsequent sections.

2.1 *ENC-t* Test

Let T be the sample size, where $T=(R+1)+P+(\tau-1)$ and R is the rolling window size, P is the number of prediction periods, and τ is the number of steps ahead to a prediction. This paper is based entirely on one-step-ahead predictions ($\tau=1$) but this is not a limiting restriction. The in-sample OLS coefficient estimate, assuming a rolling estimation window, is given by:

$$\hat{\beta}_{i,t|R} = OLS\left\{\Psi_{i,t|R} \mid X_{i,t-1},\ldots, X_{i,t-R-1}\right\}. \tag{2}$$

The current period independent variables, $X_{i,t}$, are excluded from the estimation process because these are used to predict the one-step-ahead out-of-sample prediction given by,

$$\hat{y}_{i,t+1|t,R} = X_{i,t}\hat{\beta}_{i,t|R}. \tag{3}$$

The one-step-ahead prediction errors are:

$$\hat{e}_{i,t+1|t,R} = y_{t+1} - \hat{y}_{i,t+1|t,R}. \tag{4}$$

The previous three steps are repeated from $t=R$ to $t=T-1$ to obtain P predicted errors that are used to compute the *MSPE* for each model by:

$$\hat{\sigma}^2_{i|R} = P^{-1}\sum_{t=R}^{T}\hat{e}^2_{i,t+1|t,R}. \tag{5}$$

The *adjusted-MSPE*, described in Clark and West (2007), is computed as:

$$\tilde{\sigma}^2_{j|R} = \hat{\sigma}^2_{j|R} - P^{-1}\sum_{t=R}^{T}\left(\hat{y}_{0,t+1|t,R} - \hat{y}_{j,t+1|t,R}\right)^2, \tag{6}$$

or equivalently:

$$\tilde{\sigma}^2_{j|R} = P^{-1}\sum_{t=R}^{T}\left(2\hat{e}_{j,t+1|t,R}\hat{e}_{0,t+1|t,R} - \hat{e}^2_{0,t+1|t,R}\right). \tag{7}$$

This can be broken down into the *ENC* time series:

$$\hat{f}_{j,t+1|tR} = \hat{e}^2_{0,t+1|tR} - \hat{e}^2_{j,t+1|tR} + \left(\hat{y}_{0,t+1|tR} - \hat{y}_{j,t+1|tR}\right)^2, \tag{8}$$

simplifying to:

$$\hat{f}_{j,t+1|tR} = 2\left(\hat{e}^2_{0t+1|tR} - \hat{e}_{0t+1|tR}\hat{e}_{j,t+1|tR}\right). \tag{9}$$

The mean of the *ENC* time series can be written as:

$$\bar{f}_{j|R} = P^{-1}\sum_{t=R}^{T}\hat{f}_{j,t+1|t,R}\,. \tag{10}$$

The variance of the *ENC* time series is:

$$\hat{v}_{j|R} = P^{-1}\sum_{t=R}^{T}\left(\hat{f}_{j,t+1|tR} - \bar{f}_{j|R}\right)^2 \tag{11}$$

The previous two equations are combined to compute the t-type statistic, *ENC-t,* dependent on the choice of rolling window size, resulting in:

$$ENC-t_{j|R} = P^{1/2}\frac{\bar{f}_{j|R}}{\sqrt{\hat{v}_{j|R}}}. \tag{12}$$

This statistic is used to examine pair-wise model comparison, where Clark and West (2007) found that the critical values of the limiting distribution for *ENC-t* are similar to the standard normal distribution, when all the alternative models nest the benchmark.

Hubrich and West (2010) utilized this finding to develop a methodology that would allow the analyst to identify if at least one alternative model, in a set of alternative models, surpasses the benchmark model in predictive ability. Hubrich and West propose the use of the *maxENC-t* statistic, a right-tailed test, to evaluate multiple models at once:

$$\max ENC - t_R = \max\left(ENC - t_{j|R}\right). \tag{13}$$

The null and alternative hypotheses for the *maxENC-t* test are stated as:

$$H_0 : \sigma^2_{0|R} - \sigma^2_{j|R} = 0$$
$$H_A : \max_j\left(\sigma^2_{0|R} - \sigma^2_{j|R}\right) > 0 \quad \text{for all alternative models.}$$

The *maxENC-t* statistic is also used in the maximized Monte Carlo method presented below.

2.2 Maximized Monte Carlo Test Procedure

The Monte Carlo (MC) test procedure is summarized based on the methodology and theory presented in Dufour (2006), which is a formal presentation of Dwass (1957) and Dr. Barnard's discussion on Bartlett (1963). The test statistic of interest will be denoted as S, and is a right-tailed test.[3]

The test statistic obtained from the data is denoted as S_0, and can also be identified as the null statistic. Simulation of the hypothesized null data generation process (DGP), using N independent draws of error terms from the standard normal distribution[4], making it possible to generate N simulated test statistics, denoted S_w, where w=1, … , N. The

---

[3] For convenience in this subsection, S will represent the *maxENC-t*. The *maxENC-t* statistic is right-tailed, which is consistent with the following presentation. The MC method can be altered to accommodate both left-tailed and two-tailed test hypotheses.

[4] Standard normal draws are used in this study; this is not a limitation as it not preclude draws from other distributions.

following subsection outlines the maximized Monte Carlo (MMC) test procedure, where a MMC p-value of a test statistic is calculated that depends on the set of nuisance parameters, ξ_n. The subsequent subsection outlines special cases of the MMC p-value procedure.

2.2.1 MMC p-value

Nuisance parameters are any estimated or assumed components of the hypothesized null DGP that affect the limiting distribution of the statistic of interest. They can include the coefficients, standard deviation of the error terms and even the assumed error distribution (i.e., not standard normal errors). Identification of the set of nuisance parameters, if the theoretical set is not defined, can be accomplished with some ease by simulating the null DGP and varying each potential nuisance parameter separately, which identifies the nuisance parameters as the components that change the test statistic. The set of identified nuisance parameters are denoted as $\xi_n = \{\xi_1,...,\xi_k\}$, where k represents the total number of nuisance parameters identified, and $\xi_n \in \Omega_n$ where Ω_n is the set of all possible permutations of ξ_n, or nuisance parameter space.

The MMC p-value, $\hat{p}_N(S_0 \mid \xi_n)$, is computed from null statistics obtained from the data, the test statistics from the simulated cases, and maximized with respect to the nuisance parameters in the following manner:

$$\hat{p}_N(S_0 \mid \xi_n) = \sup_{\xi_n \in \Omega_n} \left(\frac{\hat{G}_N(S_0 \mid \xi_n)+1}{N+1} \right), \tag{14}$$

where the $\hat{G}_N(S_0 \mid \xi_n)$ is the number of the simulated statistics that equal or exceed the null statistic,

$$\hat{G}_N(S_0 \mid \xi_n) = \sum_{j=1}^{N} I_{[0,\infty]}(S_w - S_0 \mid \xi_n), \tag{15}$$

and $I_{[0,\infty]}(S_w - S_0 \mid \xi_n)$ is an indicator function given by[5]:

$$I_A(x) = \begin{cases} 1, if\ x \in A \\ 0, if\ x \notin A \end{cases}. \tag{16}$$

The computed *p*-value is then tested against the critical region:

$$\hat{p}_N(S_0 \mid \xi_n) \leq \alpha, \text{ where } 0 < \alpha < 1. \tag{17}$$

Normal gradient-based maximization methods are ineffective at maximizing the MMC *p*-value, because both N and $\hat{G}_N(S_0 \mid \xi_n)$ are integers resulting in non-differentiable points in $\hat{p}_N(S_0 \mid \xi_n)$. A non-gradient-based maximization routine that is effective for functions with plateaus, multiple local maxima, and non-differentiable points is simulated annealing by Tsionas (1995).

The primary benefit of using the MMC *p*-value procedure is that it is asymptotically valid even when the asymptotic null distribution is non-standard and dependent on nuisance parameters. The two main requirements of the MMC procedure are that all the nuisance parameters (and bounds) have been identified, and the null distribution of the statistic can be simulated conditional on the nuisance parameters.

One of the drawbacks of this procedure, for a simulation study, is that it is expensive in terms of the computational time, due to the requirement for a non-gradient

---

[5] Modifying the indicator function changes the MMC test to either a right- or a two-tailed tests.

optimization routine. Also, the set up of null model for simulation presents some cost in terms of time. The time required diminishes substantially for a single call to simulated annealing, which is the case when examining a real world example.

2.2.2 MC p-value Special Cases

Dufour (2006) also outlines several other variants of the MMC p-value procedure. The first of these is the case where the set of nuisance parameters equals the null set, following the notation in the previous section $\xi_n \equiv \{0\}$. Since the MMC p-value is invariant there is no requirement to maximize the function and the earlier equation simplifies to the MC p-value test:

$$\hat{p}_N(S_0) = \frac{N \cdot \hat{G}_N(S_0) + 1}{N + 1}. \tag{18}$$

If the nuisance parameter space is very large, Dufour puts forth the option of using more liberal methods whereby the nuisance parameter space is replaced with a consistent set nuisance parameter space, $\Omega_n^c \in \Omega_n$. The function from the previous section is then maximized over this consistent set to obtain the consistent set estimate MC (CSEMC) p-value. The consistent set can be reduced further by determining a consistent point estimate of nuisance parameters, which eliminates the requirement to maximize the MC p-value, to obtain the Local MC (LMC) p-value. Commonly, the nuisance parameter values obtained from the computation of S_0 are set as the consistent point estimates for the LMC method; Dufour shows that the LMC is equivalent in outcome to a parametric bootstrap. It is easy to see since the LMC and CSEMC p-value tests have nuisance

parameter estimates that are in the set of nuisance parameters in the MMC, so a rejection for LMC or CSEMC is also a rejection for MMC.[6]

2.3 Hubrich and West (2010)

Hubrich and West (2010) present a non-parametric bootstrap method which exploits the Clark and West (2007) proposition that the adjusted squared predicted errors (adjusted SPE) have critical values very close to the standard normal critical values. Their bootstrap method draws K simulated ENC, $\hat{f}_{k,P|\Omega}$, based on the multivariate normal distribution consistent with the variance covariance matrix of the stacked ENC time series matrix representation (time in rows), so:

$$\hat{V} = P^{-1}\begin{bmatrix} \hat{f}_{1,t+1|t,R} \\ \vdots \\ \hat{f}_{m,t+1|t,R} \end{bmatrix}\begin{bmatrix} \hat{f}_{1,t+1|t,R} \\ \vdots \\ \hat{f}_{m,t+1|t,R} \end{bmatrix}'. \tag{19}$$

The simulated ENC are used to produce K simulated $maxENC\text{-}t_k$ statistics.[7] The simulated statistics can then be used to identify the critical value for a specified level, or obtain an approximation of the p-value of the null $maxENC\text{-}t_0$.

The covariance of adjusted squared predicted errors affects the asymptotic limiting distribution of the $maxENC\text{-}t$ statistic. For this Hubrich and West provide the table of

---

[6] This feature is exploited as a time saving technique.

[7] The number of simulations, K, needs to be large enough. In keeping with the Hubrich and West procedure I use 50,000 simulations to approximate the asymptotic limiting $maxENC\text{-}t$ distribution.

asymptotic critical values for two alternative models in their not-for-publication appendix. For more than two alternative models, tables of asymptotic critical values are not available due to complicated covariance relationships; their simulation method must be replicated.

2.4 Conditional Predictive Ability

The Conditional Predictive Ability (CPA) test put forth by Giacomini and White (2007) is a Wald type test of the loss function of each model put forth. The CPA test statistic is computed in the following manner:

$$CPA_R = P\overline{F}'\hat{V}^{-1}\overline{F}, \tag{20}$$

where $\hat{V}$ is computed in the same manner as the previous section, and $\overline{F}$ is the stacked mean *ENC* time series matrix representation, specifically:

$$\overline{F} = \begin{bmatrix} \bar{f}_{1|R} \\ \vdots \\ \bar{f}_{m|R} \end{bmatrix}. \tag{21}$$

Giacomini and White provide evidence that $CPA_R \overset{ASY}{\sim} \chi^2(m)$ and as such the critical values can be obtained directly from the standard $\chi^2(m)$ tables. This test, unlike the *maxENC-t* statistic, is a two-tailed test, so the null hypothesis for CPA_R is that all the variance in predicted errors are equal with a two pronged alternative where either (i) at least one alternative model is a significantly better predictor than the benchmark model, or (ii) the benchmark model is a better predictor than at least one of the alternative models. Stated formally, the two hypotheses are:

$$H_0 : \sigma^2_{0|R} - \sigma^2_{j|R} = 0$$
$$H_A : \left|\sigma^2_{0|R} - \sigma^2_{j|R}\right| > 0 \quad \text{for all alternative models.}$$

Hubrich and West (2010) used this same design with the expectation of over rejection of the null, which was the case caused by the second part of the alternative hypothesis. It should also be the case near the boundary that this test is oversized for two reasons: (a) the second part of the alternative hypothesis, and (b) the assumption of *ENC-t* sharing critical values with the standard normal distribution.

2.5 Equal Predictive Ability

Harvey and Newbold (2000) put forth four related test statistics, all of which are designed to identify Equal Predictive Ability (EPA) between competing forecast models. Three of the tests are regression-based, where the predicted errors of alternative models less the benchmark model, are regressed against the benchmark model's predicted errors. The null hypothesis for these statistics is that all the estimated parameters are not significantly different from zero and, therefore, no alternative model provides any additional information over the benchmark model. These tests, however, are not appropriate because in the case of nested models, all the alternative forecast errors converge to the benchmark errors as R increases. With large enough estimation ranges, the OLS estimator may not be able to invert the regressor matrix.

The fourth test statistic, EPA, is very similar to the CPA statistic from the previous section, with some subtle differences: (i) a modified Diebold-Mariano (1995) statistic, (ii) adjustment for number of models, as proposed by Harvey and Newbold (2000), resulting

in an F distributed statistic with $F_{m-1,P-m+1}$ critical values. The EPA is an encompassing loss function intended for nested model examination. It is computed as:

$$\hat{d}_{j,t+1|tR} = \left(\hat{e}^2_{0,t+1|tR} - \hat{e}_{0,t+1|tR}\hat{e}_{j,t+1|tR}\right). \tag{22}$$

The mean loss for each model is given by:

$$\overline{d}_{j|R} = P^{-1}\sum_{t=R}^{T}\hat{d}_{j,t+1|t,R}\,, \tag{23}$$

with a stacked matrix representation, this is:

$$\overline{D} = \begin{bmatrix} \overline{d}_{1|R} \\ \vdots \\ \overline{d}_{m|R} \end{bmatrix}. \tag{24}$$

The sample variance is computed as:

$$\hat{V}_d = (P-1)^{-1}\begin{bmatrix} \hat{d}_{1,t+1|t,R} \\ \vdots \\ \hat{d}_{m,t+1|t,R} \end{bmatrix}\begin{bmatrix} \hat{d}_{1,t+1|t,R} \\ \vdots \\ \hat{d}_{m,t+1|t,R} \end{bmatrix}'. \tag{25}$$

The Equal Predictive Ability statistic is computed from the above components as:

$$EPA_R = \frac{(P-m+1)P}{(P-1)(m-1)}\overline{D}'\hat{V}_d^{-1}\overline{D}\,. \tag{26}$$

Harvey and Newbold (2000) found this statistic to have fairly good size control, which was robust even under non-normal predicted errors (multivariate Student's t-distribution of varying degree). It is expected that this will over reject under highly persistent data for two main reasons: (1) it is a two-tailed test, and (2) the assumption that the F distribution critical values are correct.

2.6 White's Reality Check (2000)

White's Reality Check is one of the most prominent comparator procedures in the literature; making it the benchmark procedure. As defined in White (2000), the procedure uses the *MSPE* as a basis of model comparison, which as discussed earlier is problematic when all the alternative models nest the benchmark model due to non-standard limiting distribution. It is expected that this will make it an inappropriate test procedure for this analysis, but the procedure as outlined by White has been included in the examination due to its benchmark status.[8] For this study, the Reality Check (*RCMSE*) is implemented in the process outlined in White (2000), implemented with the stationary bootstrap of Politis and Romano (1994). Hansen (2005) provides an enhanced version of the Reality Check, applying a student-type adjustment (*RCMSEt*), which improves upon White's version in terms of power, while retaining similar size properties.

To correct for the inappropriateness of *MSPE* as the basis of computing the *p*-values from the Reality Check, a variant of the Reality Check is introduced on the basis of the *ENC* loss function (*RCENC*). Intuitively, this should improve the consistency of the Reality Check due to the use of a loss function that is more appropriate when comparing nested models. This study does not examine the validity of this variant, but the Clark and West (2007) findings of lower variance in the first moment combined with near standard

---

[8] Several other procedures also utilize *MSPE*, precluding them from inclusion in this study, specifically Harvey, Leybourne and Newbold (1998), and Diebold and Mariano (1995).

normal critical values provide promise asymptotically. Smaller samples with high persistence are expected to exhibit similar bias as the other bootstrap procedures examined. A student-type adjustment is also implemented for this new test (*RCENCt*), which one would expect the standardization to once again improve power.

The following outlines the computation of White's Reality Check (*RCMSE*), Hansen's version (*RCMSEt*), and both encompassing based reality checks (RCENC and RCENCt). The predicted errors, $\hat{e}_{it+1|tR}$, are identical to ones described in equation 4. White (2000) uses the difference in the squared predicted errors as the basis of his Reality Check, generating a time series of the SPE differences:

$$\hat{f}_{j,t+1|tR} = \hat{e}^2_{0,t+1|tR} - \hat{e}^2_{j,t+1|tR}, \tag{27}$$

and the mean is given by:

$$\bar{f}_{j|R} = P^{-1} \sum_{t=R}^{T} \hat{f}_{j,t+1|t,R} \cdot \tag{28}$$

White's primary statistic is given by:

$$\overline{V}_{j|R} = \sqrt{P}\bar{f}_{j|R} \cdot \tag{29}$$

The largest of these in the set of i is the statistic of interest, computed as:

$$\overline{V}_{\max|R} = \max\left(\overline{V}_{1|R}, \overline{V}_{2|R}, ..., \overline{V}_{m|R}\right). \tag{30}$$

White uses the stationary bootstrap of Politis and Romano (1994) to generate Q replications of equation 27, denoted as $\hat{f}^*_{j,q,t+1|t,R}$, where $q=1,...,Q$. The mean of each replication is given by:

$$\bar{f}^*_{j,q|R} = P^{-1} \sum_{t=R}^{T} \hat{f}^*_{j,q,t+1|t,R}, \tag{31}$$

which is used to compute the deviation of the replicated mean from the respective model mean:

$$\bar{V}^{*}_{j,q|R} = \sqrt{P}\left(\bar{f}^{*}_{j,q|R} - \bar{f}_{j|R}\right). \tag{32}$$

These Q replications are sorted from lowest to highest for each model separately:

$$\vec{V}^{*}_{j|R} = ordered\left\{\bar{V}^{*}_{j,q|R}\right\}. \tag{33}$$

The ordered-pair-wise maximum of all replications is determined, recursively, by determining the pair-wise maximum beginning with the first two replications, sorting the maximum from lowest to highest, then compare the current maximum with the next model, and repeat until the last model (m). The recursive steps are:

$$\vec{V}^{*}_{u,q|R} = \max\left\{\bar{V}^{*}_{u,q|R}, \bar{V}^{*}_{u-1,q|R}\right\}, \text{ and} \tag{34}$$

$$\vec{V}^{*}_{u,q|R} = ordered\left\{\bar{V}^{*}_{u,q|R}\right\} \text{ where } u \text{ ranges from 1 and } m. \tag{35}$$

White's Reality Check p-value is computed by determining the number of elements in $\vec{V}^{*}_{m|R}$ that equal or exceeds the value of $\bar{V}_{\max|R}$ and dividing by Q.

Hansen's Reality Check is procedurally identical to White's, with the exception that Hansen's test statistics use different variants of equations 29 and 32, which will be denoted below as 29a and 32a. Two variances are computed, the first is:

$$\hat{w}^{2}_{j|R} = P^{-1}\sum_{t=R}^{T}\left[\sqrt{P}\left(\hat{f}_{j,t+1|tR} - \bar{f}_{j|R}\right)\right]^{2} \text{ ; and the second is:} \tag{36}$$

$$\left(\hat{w}^{*}_{j,q|R}\right)^{2} = P^{-1}\sum_{t=R}^{T}\left[\sqrt{P}\left(\hat{f}^{*}_{j,q,t+1|tR} - \bar{f}_{j|R}\right)\right]^{2}. \tag{37}$$

The student-adjustment of equation 29 is,

$$\overline{V}_{j|R} = \frac{\sqrt{P}\,\bar{f}_{j|R}}{\hat{w}_{j|R}}, \tag{29a}$$

and the set of stationary bootstrap replicated statistics are given as:

$$\overline{V}^{*}_{j,q|R} = \frac{\sqrt{P}\left(\bar{f}^{*}_{j,q|R} - \bar{f}_{j|R}\right)}{\hat{w}^{*}_{j,q|R}}. \tag{32a}$$

The remainder of the computation of Hansen's Reality Check *p*-value follows the same process as White's Reality Check.

The ENC-based Reality Checks follow the same methodology of White and Hansen, with the only difference in equation 27 in the following manner:

$$\hat{f}_{j,t+1|tR} = \hat{e}^{2}_{0,t+1|tR} - \hat{e}^{2}_{j,t+1|tR} + \left(\hat{e}_{0,t+1|tR} - \hat{e}_{j,t+1|tR}\right)^{2}. \tag{27a}$$

3 Simulation Design

The presented Hubrich and West (2010) simulations use an AR(1) data generation process (DGP) where the null model parameter is 0.5. The simulations below use the same framework as Hubrich and West (2010) but increase the null model parameter to 0.99, which introduces a much higher level of persistence into the DGP. Hubrich and West (2010) acknowledge that their examination of US inflation may exhibit high persistent conditions; they cite the findings of Hendry and Hubrich (2011) in inconsistent ADF unit root results for different time periods in inflation series. I examine the possibility that their bootstrap method may be biased, due to the highly persistent nature of inflation data in combination with their in-sample and out-of-sample settings.

The objective of the simulation is to determine whether or not disaggregate components provide additional information when forecasting highly persistent aggregate

data. Consider the aggregate series y_t as the sum of D disaggregate components, $y_{d,t}$ where d=1,…,D, specifically:

$$y_t = \sum_{j=1}^{D} y_{d,t} \,. \tag{38}$$

The benchmark model and m=2 alternative models (or m=4 alternative models) that nest the benchmark in the following manner:

the benchmark is $y_t = const. + \beta_{0,0,t|R}\, y_{t-1} + e_{0,t}$, and (39)

the alternative j is $y_t = const. + \beta_{0,j,t|R}\, y_{t-1} + \beta_{1,j,t|R}\, y_{i,t-1} + e_{j,t}$, (40)

for $j = 1$ and 2.

The DGP for the disaggregates of y_t is a VAR of order 1 with a D x D matrix of autoregressive parameters, Φ, a mean vector $\mu = (\mu_1, \mu_2, ..., \mu_D)'$ where $\mu_d = 1$ for all d, and zero mean i.i.d. normal disturbances $U_t = (u_{1,t}, u_{2,t}, ...u_{D,t})'$ so:

$$Y_t \equiv (y_{1,t}, y_{2,t}, ..., y_{D,t})' = \mu + \Phi Y_{t-1} + U_t \,. \tag{41}$$

For the determination of size for each of the test statistics with persistent data, I assume a common value of $\phi = 0.99$ as the diagonal elements of Φ, and D=3, specifically:

$$\Phi = \begin{bmatrix} 0.99 & 0 & 0 \\ 0 & 0.99 & 0 \\ 0 & 0 & 0.99 \end{bmatrix}. \tag{42}$$

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, $\phi = 0.99$.

In the power simulations, it is assumed that at least one of the disaggregate components Granger causes the aggregate in (42), and as such Φ is updated to[9]:

$$\Phi = \begin{bmatrix} 0.99 & -0.008 & 0 \\ 0.2 & 0.5 & 0 \\ 0 & 0 & 0.99 \end{bmatrix}. \tag{43}$$

This design allows for high persistence in two of the series, and weaker persistence in the other disaggregate.[10]

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 (42) are increased to 4 x 4 with $\phi = 0.99$ on the diagonal. In the power simulations, the matrix in (43) is augmented to[11]:

$$\Phi = \begin{bmatrix} 0.99 & -0.008 & 0 & 0 \\ 0.2 & 0.5 & 0 & 0 \\ 0 & 0 & 0.99 & 0 \\ 0 & 0 & 0 & 0.99 \end{bmatrix}. \tag{44}$$

---

[9] The choice of Φ for the power simulations has an effect on the power results. Power is greatly diminished if the diagonal elements remain at 0.99, because there is little leeway in the off diagonal elements that preserve stationarity of the system. So 0.5, 0.2 and -0.008 were used to allow for observable and comparable power results. Even with one stationary disaggregate, the estimation of the aggregate parameter continued to be dominated by the 0.99 parameters.

[10] The eigenvalues of the matrix in (32) are: 0.98671264, 0.50328736, and 0.99.

[11] The eigenvalues of the matrix in (33) are: 0.98671264, 0.50328736, 0.99, and 0.99

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.[12] Specifically, if the evaluation of the *p*-value exceeded the nominal size (α) then the simulated annealing procedure was halted and returned identifying to retain the null hypothesis.[13] The *maxENC-t* test statistic from the null DGP was determined through simulation to be invariant to: the standard deviation of the error term draws (U_t), constant term (μ), and initial value of the disaggregates ($y_{d,0}$). The only nuisance parameter identified is the common value ϕ.

The Hubrich and West (2010) procedure was followed closely, with the rank of the null statistic identified in relation to 50,000 replications of the adjusted squared predicted errors. The precise *p*-value was not required for the simulation study, so time saving

---

[12] N can also take on values of 19 or 199 with little change to the simulation results. To decrease simulation time N=19 was chosen for the R=400 tests. For R=400, P=100, and m=2, the empirical size was determined as 0.031 for N=99 and 0.030 for N=19, also the power results were 0.510 and 0.497, respectively. An appendix table provides the simulation comparison for R=200, for both size and power.

[13] Another time saving option not used in this study is to limit the clock time for a single call of simulated annealing; this is usually more appropriate when there are a large number of nuisance parameters.

techniques were used to avoid having to simulate all 50,000 replications every time.[14] The asymptotic tabulation of critical values for m=2, provided in their not-for-publication Appendix, were not used in this simulation study; instead the author's 50,000 replications procedure for both m=2 and m=4 was used.

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

All four reality check simulations, RCMSE, RCMSEt, RCENC and RCENCt, use 1000 stationary bootstrap replications, with the geometric mean block size of 2, which is consistent with Hubrich and West (2010) and White (2000).

4 Simulation Results

The simulation results are displayed in the following two tables for empirical size and power, respectively. The MMC method, shown in Table 1, provides a significant measure of size control under the conditions examined, whereas the other procedures tested exhibit oversized or inconsistently sized results. The size results for smaller values of R are modestly conservative, becoming more conservative as R increases. The

---

[14] The first time saving technique was to stop the replications and accept H_0 if the null statistic's rank exceeded critical number (α*50,000). The second time saver would stop the replications and reject the null if there were too few replications remaining for this decision to change.

simulations indicate that the MMC method is robust to increases in the number of alternative models (m). It also exhibits good size stability as P increases for a given R

The Hubrich and West (2010) bootstrap approach is oversized for all combinations of R, P and m examined in this study. The empirical size rises as P increases, which is expected based on the shift in the simulated distribution presented in Figure 2. It also increases in m, but the rise is modest for the small set of models examined in this study. As R increases, the severity of the oversized results diminish in their nonparametric bootstrap method, to the point that it is almost correctly sized with R=400, m=2 and for all the values of P tested. It appears that the asymptotic properties of the Hubrich and West (2010) method may be recovered with large enough values of R.

The CPA method's empirical size results are considerably oversized for all combinations of R, P and m in this examination. Unlike Hubrich and West (2010), the CPA size results do not exhibit any consistent patterns. For m=2, size results generally improve as R increases, but this is not the case when m=4 where patterns are erratic. The increase in the number of alternative models uniformly increases the empirical size, but with differing relative increases. When m=4 and R>40, an increase in P results in an improvement in the empirical size, but this pattern, in the m=2 results, is only apparent for R=100 and R=200. Hubrich and West (2010) found that the empirical size of the CPA was higher than that of the *maxENC-t*, but this is not the case for the results shown in Table 1. This is somewhat surprising since *CPA* is a two-tailed test and *maxENC-t* is one-tailed.

Table 1: Empirical size for nominal 0.10 tests for one-step-ahead predictions

		m = 2			*m* = 4		
		P = 40	*P* = 100	*P* = 200	*P* = 40	*P* = 100	*P* = 200
R = 40	MMC	0.076	0.089	0.092	0.080	0.091	0.071
	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	MMC	0.068	0.067	0.081	0.086	0.076	0.084
	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	MMC	0.067	0.061	0.063	0.087	0.079	0.065
	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
R = 400	MMC†	0.024	0.030	0.088	0.037	0.026	0.039
	HW2010	0.113	0.112	0.119	0.141	0.135	0.146
	CPA	0.165	0.170	0.171	0.297	0.270	0.248
	EPA	0.363	0.363	0.382	0.346	0.365	0.359
	RCMSE	0.111	0.096	0.067	0.166	0.162	0.141
	RCMSEt	0.125	0.113	0.077	0.187	0.174	0.153
	RCENC	0.142	0.151	0.163	0.211	0.227	0.229
	RCENCt	0.157	0.167	0.172	0.219	0.240	0.235

† Based on N=19.

The commonly utilized Reality Check (RCMSE), which normally is a very conservative test, provides inconsistent size control. Increases in P causes the empirical size to diminish, but it rises as m and R increase. As discussed earlier, the *MSE*-type statistic used in the common Reality Check is inappropriate for nested models, so in an

attempt to correct for this I introduce the use of an alternative loss function based on the *ENC*-type statistic, RCENC. This reformulated Reality Check provides more sensible responses to increases in R, P and m, but the test is now heavily oversized under almost all the conditions. It is possible that the RCENC may correct for some of the size and power problems shown in nested model studies, but this is outside the scope of this paper and left for future research.

The EPA test is heavily oversized under all conditions examined. The empirical size results have the following properties: size decreases as m increases, but increases in P and R cause inconsistent changes in empirical size.

The simulated raw power results for the various methods are presented in Table 2. Lloyd (2005) demonstrates that the examination of raw power for competing tests can lead to incorrect conclusions. He presents a size-adjustment method which takes into account the empirical size to power relationship for a given statistic, receiver operating characteristic curve (ROC), to provide a means of size-adjusting power. His methodology assumes the comparison of different test statistics: this is not the case for many of my tests, which may lead to an inappropriate size-adjustment methodology.

Based on simulated size and power distributions, it was determined that an increase in R and P should result in an increase in power, and increases in m will cause a drop in power. Consistent with the simulated expectations raw power increases in P and decreases in m for all methods; except that both *RCMS*E and *RCMSEt* display opposite relationships. Contrary to expectations, increases in R result in a drop in raw power. Empirical size decreases significantly with increases in R, which may be dominating these raw power results.

Table 2: Empirical raw power for nominal 0.10 tests for one-step-ahead predictions

		m = 2			*m* = 4		
		P = 40	*P* = 100	*P* = 200	*P* = 40	*P* = 100	*P* = 200
R = 40	MMC	0.362	0.588	0.787	0.192	0.310	0.465
	HW2010	0.531	0.806	0.963	0.358	0.599	0.831
	CPA	0.500	0.747	0.946	0.391	0.553	0.766
	EPA	0.658	0.869	0.979	0.449	0.636	0.845
	RCMSE	0.149	0.106	0.089	0.161	0.117	0.080
	RCMSEt	0.151	0.108	0.088	0.167	0.118	0.076
	RCENC	0.395	0.554	0.720	0.394	0.553	0.729
	RCENCt	0.409	0.576	0.752	0.395	0.546	0.722
R = 100	MMC	0.337	0.645	0.851	0.194	0.344	0.543
	HW2010	0.551	0.807	0.948	0.397	0.569	0.785
	CPA	0.500	0.784	0.941	0.474	0.625	0.801
	EPA	0.708	0.888	0.976	0.531	0.719	0.866
	RCMSE	0.165	0.144	0.102	0.231	0.157	0.138
	RCMSEt	0.181	0.149	0.103	0.241	0.154	0.129
	RCENC	0.322	0.408	0.511	0.386	0.464	0.607
	RCENCt	0.340	0.428	0.530	0.375	0.444	0.588
R = 200	MMC	0.332	0.601	0.834	0.177	0.323	0.504
	HW2010	0.516	0.777	0.919	0.352	0.526	0.724
	CPA	0.527	0.808	0.942	0.495	0.681	0.817
	EPA	0.695	0.919	0.983	0.544	0.775	0.875
	RCMSE	0.152	0.112	0.083	0.228	0.212	0.158
	RCMSEt	0.176	0.132	0.094	0.236	0.208	0.160
	RCENC	0.244	0.290	0.350	0.352	0.417	0.477
	RCENCt	0.275	0.305	0.368	0.343	0.405	0.460
R = 400	MMC†	0.191	0.497	0.747	0.147	0.312	0.462
	HW2010	0.478	0.755	0.886	0.279	0.504	0.705
	CPA	0.501	0.837	0.958	0.495	0.731	0.908
	EPA	0.714	0.924	0.977	0.560	0.820	0.942
	RCMSE	0.118	0.122	0.098	0.221	0.211	0.185
	RCMSEt	0.151	0.149	0.126	0.233	0.218	0.180
	RCENC	0.164	0.207	0.248	0.301	0.349	0.375
	RCENCt	0.199	0.234	0.267	0.285	0.338	0.350

† Based on N=19.

5 Conclusion

The bootstrap procedures and the use of tabulated asymptotic critical values are oversized when the null model is highly persistent and small finite sample sizes. This simulation study demonstrates that the maximized Monte Carlo method provides good size control, even under the smallest sample size tested. Raw power for the MMC procedure is in line with expectations based sample and prediction sizes and the number of competing models. Although the raw power of the MMC procedure is lower than competing methods, it is due in a large part to the oversized nature of the competing procedures.

This study introduces Reality Checks altered to incorporate the conceptual design of the encompassing test statistics, which intuitively should correct for the inappropriate use of either White's or Hansen's Reality Check when all competing models nest the benchmark model. However, these novel statistics perform poorly both in terms of size control and raw power in the simulation results.

6 Future research

I show through simulation that the MMC procedure provides good size control, compared to several other procedures, in the case of nested models of a highly persistent process. The next logical extension is to determine if there is a size problem in the case where at least one of the alternative models does not nest the benchmark model (non-nested). This has two implications: the appropriate test statistic changes from an ENC-type to MSE-type, and identifying if any of the leading MSE-based procedures have a size problem near the boundary.

Another natural extension is to use the MMC procedure to examine models of known highly persistent processes, like exchange rates, oil price, and stock price returns. Test for unit roots on these series generally yield inconsistent identification of a unit root, due to the weak power properties of these tests. This poses a problem for many of the bootstrap methods examined in this study. However, the MMC procedure retains it size control with good power, even when testing multiple models that are all close to the unit root boundary, making it an ideal option.

7 References

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

Andrews D.W.K. (2000), Notes and comments: Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space, *Econometrica*, 68(2), pp. 399-405

Bartlett M.S. (1963), The spectral analysis of point processes, *Journal of the Royal Statistical Society, Series B (Methodological)*, 25(2), pp. 264-296

Bernard J-T., Dufour J-M., Khalaf L. and Kichian M. (2010), An identification-robust test for time-varying parameters in the dynamics of energy prices, *Journal of Applied Econometrics*, doi: 10.1002/jae.1213

Clark T.E. and McCracken M.W. (2001), Test of equal forecast accuracy and encompassing for nested models, *Journal of Econometrics*, 105, pp. 85 - 110

Clark T.E. and West K.D. (2007), Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics*, 138, pp. 291- 311

Diebold F.X. and Mariano R.S. (1995), Comparing predictive accuracy, *Journal of Business & Economic Statistics*, 20(1), pp. 134 - 144

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

Dwass M. (1957), Modified randomization tests for nonparametric hypotheses, *The Annals of Mathematical Statistics*, 28(1), pp. 181-187

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

Hansen P.R. (2005), A test for superior predictive ability, *Journal of Business & Economic Statistics*, 23(4), pp. 365-379

Harvey D., Leybourne S.J. and Newbold P. (1998), Tests for forecast encompassing, *Journal of Business & Economic Statistics*, 16(2), pp. 254-259

Harvey D. and Newbold P. (2000), Tests for multiple forecast encompassing, *Journal of Applied Econometrics*, 15, pp. 471-482

Hendry D.F. and Hubrich K. (2011), Combining disaggregate forecast or combining disaggregate information to forecast an aggregate, *Journal of Business and Economic Statistics*, 29(2), pp. 216-227

Hubrich K. and West K.D. (2010), Forecast evaluation of small nested model sets, *Journal of Applied Econometrics*, 25, pp. 574-594

Lloyd C.J. (2005), Estimating test power adjusted for size, *Journal of Statistical Computation and Simulation*, 75(11) pp. 921-934

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

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

Tsionas E.G. (1995), Simulated annealing program.

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

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

8 Appendices

Appendix A: Effect of the number of draws on the MMC method

This appendix details the effect of the choice of N on the simulation results. This is important because the results presented in table 2 for R=400 are based on N=19 rather than N=99, as was the case for the other choices of R.

It is apparent from the table below that the MMC method becomes more conservative as N decreases, concurrently the power of the method is also diminished. The computing time, which is proportional to the number of calls to simulated annealing, decreases substantially for this simulation study as N is reduced from 99 to 19.

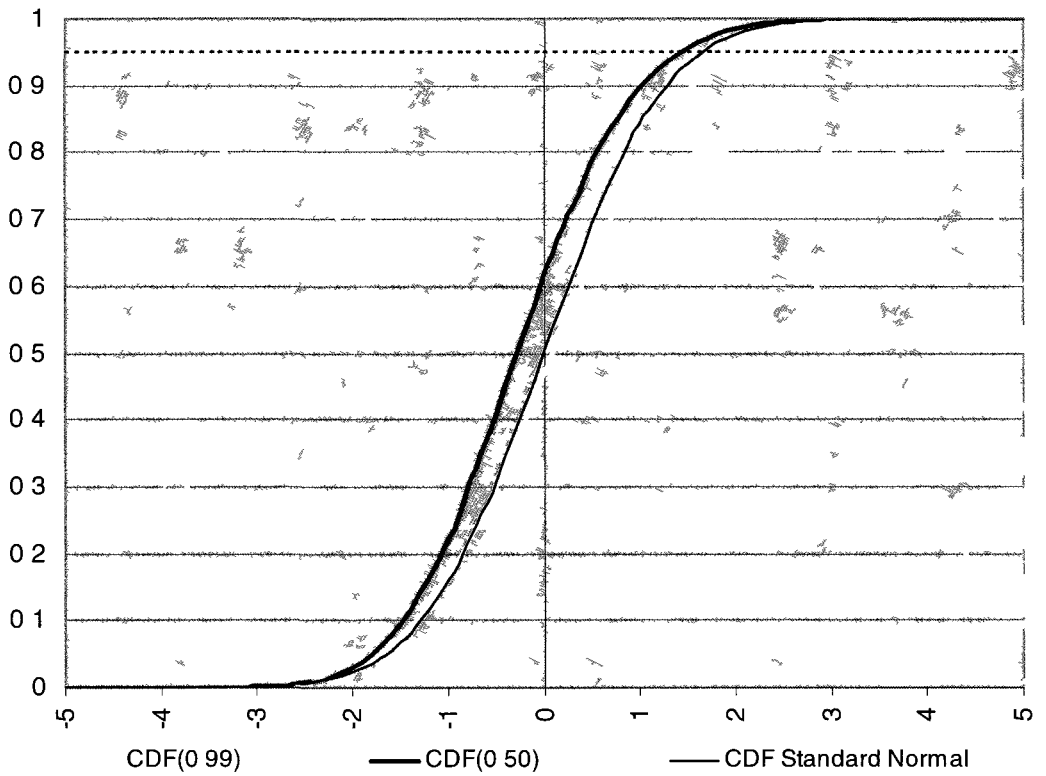
Table 3: Empirical Size and Power Properties of the MMC Method as N varies

		Size			Power		
		N = 99	N = 19	ξ	N = 99	N = 19	ξ
m = 2	P = 40	0.067	0.012	0.055	0.332	0.197	0.135
	P = 100	0.061	0.029	0.032	0.601	0.452	0.149
	P = 200	0.063	0.048	0.015	0.834	0.732	0.102
m = 4	P = 40	0.087	0.063	0.024	0.177	0.102	0.075
	P = 100	0.079	0.029	0.050	0.323	0.229	0.094
	P = 200	0.065	0.066	-0.001	0.504	0.424	0.080

Appendix B: Cumulative distribution curves for the ENC-t statistic for a highly persistent process

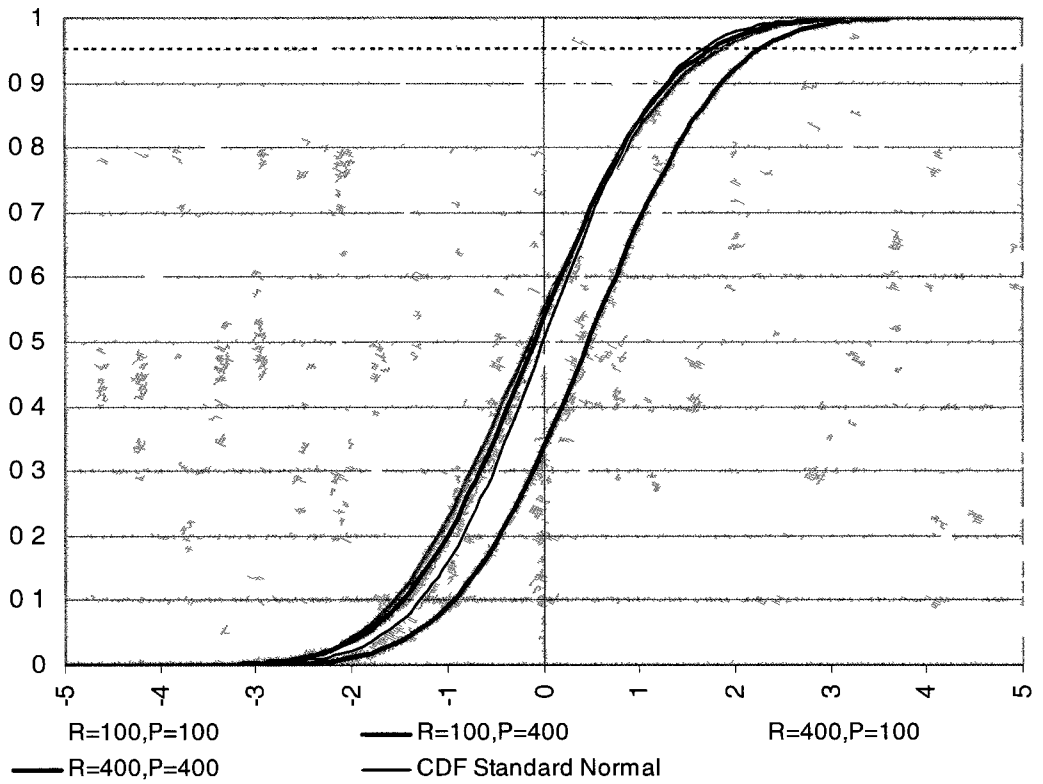
Critical values for a statistic are generally identified based on the CDF The following charts are the complementary CDF for the simulated PDF curves in figures 1 and 2, respectively

Figure 3: Simulated CDF curves for ENC-t at different levels of persistence



Notes All distributions presented are simulated The Standard Normal CDF is generated from 500,000 draws from the normal distribution The ENC-t distributions are based on 50,000 simulations of the ENC-t statistic for a single alternative model versus a benchmark model that is the null DGP The simulation design is outlined in section 3, with 0 99 and 0 50 representing the lag dependent coefficient in the null DGP, rolling window of 100 regression observations with 100 one-step-ahead predictions The dashed line is the right-tailed 95th percentile cut-off for the standard normal distribution

Figure 4: Simulated ENC-t CDF varying P and R, highly persistent processes



Notes See Figure 3 notes for simulation specifics The distributions are based on simulations where the lag dependent parameter is 0 99, and the number of regression observations (R) and one-step-ahead predictions (P) vary

Appendix C: Size and power of statistical test procedures

There are two ways in which a test procedure can result in erroneous conclusions:

- type I error: the null hypothesis is true and the test procedure rejects, and
- type II error: the null hypothesis is false and the test procedure fails to reject.

The size of a test procedure (or significance level) is the probability of type I error, usually denoted with α. In general, the size of the test, often 10 or 5 per cent, is chosen at the discretion of the analyst and critical values are taken from asymptotic tables (in many cases generated via simulation). However, if the conditions of the test procedure are not consistent with the asymptotic conditions, then the asymptotic critical values can be inappropriate resulting in either over- or under-reject the null hypothesis: termed oversized and undersized, respectively.

The probability of type II error is usually denoted with β, and the power of the test is $1-\beta$. So for a given α, it is desirable to have the highest power possible, or lowest β, which minimized the possibility of committing either type of error.

In the body of the text, I use the term "raw power" which is used to describe the power obtained from the simulation study, based on an analyst-specified α which pairs with a simulated size. Lloyd (2005) provides a means of obtaining a "size-adjusted power" from the test procedure's ROC curve, which is developed from the raw power and simulated size as α ranges from zero to one. The size-adjusted power is then simply the raw power whose simulated size matches the analyst-specified α. So for oversized test procedures the α used to obtain the size-adjusted power will be lower than the desired α, and vice versa for undersized test procedures. This method is useful when comparing test procedures that use different statistics, but can become irrelevant for the

same statistic since it could have the identical distribution, and size-adjusted power, with competing test procedures.

An ad hoc, but commonly used, method of computing a size-adjusted power is to use raw power less simulated size plus predetermined α. This is generally appropriate to obtain inferences on each procedure and statistic.