Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes*

Valentina Corradi[1] and Norman R. Swanson[2]
[1]Queen Mary, University of London and [2]Rutgers University

March 2005

Abstract

Our objectives in this paper are twofold. First, we introduce block bootstrap techniques that are (first order) valid in recursive estimation frameworks. Thereafter, we present two applications where predictive accuracy tests are made operational using our new bootstrap procedures. One of the applications outlines a consistent test for out-of-sample nonlinear Granger causality, and the other outlines a test for selecting amongst multiple alternative forecasting models, all possibly misspecified. More specifically, our examples extend the White (2000) reality check to the case of non vanishing parameter estimation error, and extend the integrated conditional moment tests of Bierens (1982, 1990) and Bierens and Ploberger (1997) to the case of out-of-sample prediction. In both of these examples, it is shown that appropriate re-centering of the bootstrap score is required in order to ensure that the tests have asymptotically correct size, and the need for such re-centering is shown to arise quite naturally when testing hypotheses of predictive accuracy. We also discuss a Monte Carlo investigation that compares the finite sample properties of our block bootstrap procedures with a parametric bootstrap due to Kilian (1999); all within the context of various encompassing and predictive accuracy tests. An empirical illustration is also discussed, in which it is found that unemployment appears to have nonlinear marginal predictive content for inflation.

JEL classification: C22, C51.
Keywords: block bootstrap, recursive estimation scheme, reality check, nonlinear causality, parameter estimation error.

---

* Valentina Corradi, Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS, UK, v.corradi@qmul.ac.uk. Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, nswanson@econ.rutgers.edu. We owe a great many thanks to the editor, Frank Schorfheide, and two referees for numerous useful suggestions, all of which we feel were instrumental in our revision of this paper. Additionally, we wish to thank Jean-Marie Dufour, Silvia Goncalves, Stephen Gordon, Clive Granger, Oliver Linton, Brendan McCabe, Antonio Mele, Andrew Patton, Rodney Strachan, Christian Schluter, Allan Timmerman, and seminar participants at Cornell University, the London School of Economics, Laval University, Queen Mary, University of London, CIREQ-Universite' de Montreal, the University of Liverpool, Southampton University, the University of Virginia, the 2004 Winter Meetings of the Econometric Society, and the Bank of Canada for useful comments and suggestions on earlier versions of this paper. Corradi gratefully acknowledges financial support via ESRC grant RES-000-23-0006, and Swanson thanks the Rutgers University Research Council for financial support.

1 Introduction

Often economic models are compared in terms of their relative predictive accuracy. Thus, it is not surprising that a large literature on the topic has developed over the last 10 years, including, for example, important papers by Diebold and Mariano (DM: 1995), West (1996), and White (2000). In this literature, it is quite common to compare multiple models (which are possibly all misspecified - i.e. they are all approximations of some unknown true model) in terms of their in or out of sample predictive ability, for given loss function. In such contexts, one often compares parametric models containing estimated parameters. Hence, it is important to take into account the contribution of parameter estimation error when carrying out inference. Though some authors make a point in favor of in-sample predictive evaluation, see e.g. Inoue and Kilian (2004, 2005), it is common practice to split samples of size T into $T = R + P$ observations, where only the last P observations are used for predictive evaluation. We consider such a setup, and assume that parameters are estimated in a recursive fashion, such that R observations are used to construct a first parameter estimator, say $\widehat{\theta}_R$, a first prediction (say a 1-step ahead prediction), and a first prediction error. Then, $R + 1$ observations are used to construct $\widehat{\theta}_{R+1}$, yielding a second ex ante prediction and prediction error. This procedure is continued until a final estimator is constructed using $T - 1$ observations, resulting in a sequence of $P = T - R$ estimators, predictions, and prediction errors. If R and P grow at the same rate as the sample size increases, the limiting distributions of predictive accuracy tests using this setup generally reflects the contribution of parameter uncertainty (i.e. the contribution of $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_t - \theta^\dagger \right)$, where $\widehat{\theta}_t$ is a recursive $m-$estimator constructed using the first t observations, and $\theta^\dagger$ is its probability limit, say).[1]

Our objectives in this paper are twofold. First, we introduce block bootstrap techniques that are (first order) valid in recursive estimation frameworks. Thereafter, we present two applications where predictive accuracy tests are made operational using our new bootstrap procedures. One of the applications outlines a consistent test for out-of-sample nonlinear Granger causality, and the other outlines a test for selecting amongst multiple alternative forecasting models, all of which may be thought of as approximations of some unknown underlying model.

In some circumstances, such as when constructing Diebold and Mariano (1995) tests for equal

[1] $m-$estimators include least squares, nonlinear least square, (quasi) maximum likelihood, and exactly identified instrumental variables and generalized method of moments estimators.

(pointwise) predictive accuracy of two models, the limiting distribution is a normal random variable. In this case, the contribution of parameter estimation error can be addressed using the framework of West (1996), and essentially involves estimating an "extra" covariance term. However, in other circumstances, such as when constructing tests which have power against generic alternatives, the statistic has a limiting distribution that can be shown to be a functional of a Gaussian process with a covariance kernel that reflects both (dynamic) misspecification as well as the contribution of (recursive) parameter estimation error. Such a limiting distribution is not nuisance parameter free, and critical values cannot be tabulated. However, valid asymptotic critical values can be obtained via appropriate application of the (block) bootstrap. This requires the formulation of a bootstrap procedure that allows for the formulation of statistics which properly mimic the contribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$ (i.e. parameter estimation error). The first objective of this paper is thus to suggest a block bootstrap procedure which is valid for recursive m-estimators, in the sense that its use suffices to mimic the limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$.

When forming the block bootstrap for recursive m-estimators, it is important to note that earlier observations are used more frequently than temporally subsequent observations when forming test statistics. On the other hand, in the standard block bootstrap, all blocks from the original sample have the same probability of being selected, regardless of the dates of the observations in the blocks. Thus, the bootstrap estimator, say $\widehat{\theta}_t^*$, which is constructed as a direct analog of $\widehat{\theta}_t$, is characterized by a location bias that can be either positive or negative, depending on the sample that we observe. In order to circumvent this problem, we suggest a re-centering of the bootstrap score which ensures that the new bootstrap estimator, which is no longer the direct analog of $\widehat{\theta}_t$, is asymptotically unbiased. It should be noted that the idea of re-centering is not new in the bootstrap literature for the case of full sample estimation. In fact, re-centering is necessary, even for first order validity, in the case of overidentified generalized method of moments (GMM) estimators (see e.g. Hall and Horowitz (1996), Andrews (2002, 2004), and Inoue and Shintani (2004)). This is due to the fact that, in the overidentified case, the bootstrap moment conditions are not equal to zero, even if the population moment conditions are. However, in the context of $m-$estimators using the full sample, re-centering is needed only for higher order asymptotics, but not for first order validity, in the sense that the bias term is of smaller order than $T^{-1/2}$ (see e.g. Andrews (2002)). However, in the case of recursive $m-$estimators the bias term is instead of order $T^{-1/2}$, and so it does contribute to the limiting distribution. This points to a need for re-centering when using recursive estimation

schemes, and such re-centering is discussed in the next section.

The block bootstrap for recursive m-estimators that is discussed in this paper can be used to provide valid critical values in a variety of interesting testing contexts, and two such leading applications are developed. As mentioned above, the first is a generalization of the reality check test of White (2000) that allows for non vanishing parameter estimation error. The second is an out-of-sample version of the integrated conditional moment (ICM) test of Bierens (1982,1990) and Bierens and Ploberger (1997) which provides out of sample tests consistent against generic (non-linear) alternatives.[2] More specifically, our first application concerns the reality check of White (2000), which extends the Diebold and Mariano (1995) and West (1996) test for the relative predictive accuracy of two models by allowing for the joint comparison of multiple misspecified models against a given benchmark. The idea of White (2000) is to compare all competing models simultaneously, thus taking into account any correlation across the various models. In this context, the null hypothesis is that no competing model can outperform the benchmark, for a given loss function. As this test is usually carried out by comparing predictions form the alternative models, and given that predictions are usually formed using recursively estimated models, the issue of parameter estimation uncertainty arises. White (2000) obtains valid asymptotic critical values for his test via use of the Politis and Romano (1994) stationary bootstrap for the case in which parameter estimation error is asymptotically negligible. This is the case in which either the same loss function is used for both estimation and model evaluation, or P grows at a slower rate than R. Using the block bootstrap for recursive m-estimators, we generalize the reality check to the case in which parameter estimation error does not vanish asymptotically.

The objective of the second application is to test the predictive accuracy of a given (non)linear model against generic (non)linear alternatives. In particular, one chooses a benchmark model, and the objective is to test whether there is an alternative model which can provide more accurate, loss function specific, out-of-sample predictions. As the test is based on a continuum of moment conditions and is consistent against generic alternatives, we call it an Integrated Conditional Moment test. The suggested ICM type test differs from those developed by Bierens (1982,1990) and Bierens and Ploberger (1997) because parameters are estimated recursively, out-of-sample prediction models are analyzed, and the null hypothesis is that the reference model is the best "loss function

[2] An application to predictive density and confidence intervals forecast evaluation is given in Corradi and Swanson (2005).

specific" predictor, for a given information set. Given that the test compares out-of-sample prediction models, it can be viewed as a test for (non)linear out-of-sample Granger causality. This application builds on previous work by Corradi and Swanson (2002), who use a conditional p-value method for constructing critical values in this context, extending earlier work by Hansen (1996) and Inoue (2001). However, the conditional p-value approach suffers from the fact that under the alternative, the simulated statistics diverges (at rate as high as $\sqrt{\widetilde{l}}$), conditional on the sample, where $\widetilde{l}$ plays a role analogous to the block length in the block bootstrap. This feature clearly leads to reduced power in finite samples, as shown in Corradi and Swanson (2002). As an alternative to the conditional p-value approach, we thus establish in our second application that the bootstrap for recursive $m-$estimators yields $\sqrt{P}$-consistent ICM tests.

In order to shed evidence on the usefulness of the recursive block bootstrap, we carry out a Monte Carlo investigation that compares the finite sample properties of our block bootstrap procedures with two alternative naive block bootstraps as well as the parametric bootstrap due to Kilian (1999); all within the context of various encompassing and predictive accuracy tests including those due to Diebold and Mariano (1995), Chao, Corradi and Swanson (2001), and Clark and McCracken (2004). Results suggest that our recursive block bootstrap outperforms alternative naive block bootstraps. Additionally, the Kilian bootstrap is shown to be robust to nonlinear misspecification, and all of the test statistics examined are found to have good finite sample properties when applied in situations where there is model misspecification.

An empirical illustration is also discussed, in which it is found that unemployment appears to have nonlinear marginal predictive content for inflation, as evidenced by use of the generic out-of-sample nonlinear test discussed here as well as the Chao, Corradi and Swanson (2001) encompassing test. It turns out that much can be learned by using *all* of the different tests in consort with one another. The picture that emerges when only a subset of the tests is used to analyze the marginal predictive content of unemployment for inflation is that of an absence of predictive ability. When all of the tests are used, on the other hand, interesting evidence arises concerning the potential nonlinear predictive content of unemployment. Thus, the tests discussed in this illustration appear to be useful complements to each other.

The rest of the paper is organized as follows. Section 2 outlines the block bootstrap for recursive $m-$estimators and contains asymptotic results. Sections 3 and 4 outline the two applications of the recursive block bootstrap: White's reality check and out-of-sample integrated conditional moment

test. Monte Carlo findings are discussed in Section 5. An empirical illustration is presented in Section 5. Finally, concluding remarks are given in Section 6. All proofs are collected in an Appendix.

Hereafter, P^* denotes the probability law governing the resampled series, conditional on the sample, E^* and Var^* are the mean and variance operators associated with P^*, $o_P^*(1)$ $\Pr -P$ denotes a term converging to zero in Q^*-probability, conditional on the sample, and for all samples except a subset with probability measure approaching zero, and $O_P^*(1)$ $\Pr -P$ denotes a term which is bounded in P^*-probability, conditional on the sample, and for all samples except a subset with probability measure approaching zero. Analogously, $O_{a.s.^*}(1)$ and $o_{a.s.^*}(1)$ denote terms that are almost surely bounded and terms that approach zero almost surely, according the the probability law P^*, and conditional on the sample. Note, that P is also used to denote the length of the prediction period, however the meaning results clear from the context.

2 The Block Bootstrap for Recursive $m-$Estimators

In this section, we establish the first order validity of a block bootstrap estimator that captures the effect of parameter estimation error in the context of *recursive* m-estimators, which are defined as follows. Let $Z^t = (y_t, ..., y_{t-s_1+1}, X_t, ..., X_{t-s_2+1})$, $t = 1, ..., T$, and let $s = \max\{s_1, s_2\}$. Additionally, assume that $i = 1, ..., n$ models are estimated (thus allowing us to establish notation that will be useful in the applications presented in subsequent sections). Now, define the *recursive* m-estimator for the parameter vector associated with model i as:[3]

$$\widehat{\theta}_{i,t} = \arg\min_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=s}^{t} q_i(y_j, Z^{j-1}, \theta_i), \quad R \leq t \leq T-1, \ i = 1, ..., n \tag{1}$$

and

$$\theta_i^\dagger = \arg\min_{\theta_i \in \Theta_i} E(q_i(y_j, Z^{j-1}, \theta_i)), \tag{2}$$

where q_i denotes the objective function for model i. As the discussion below does not depend on any specific model, we drop the subscript i. Following standard practice (such as in the real-time forecasting literature), this estimator is first computed using R observations. In our applications

[3]Within the context of full sample estimation, the first order validity of the block bootstrap for $m-$estimators has been shown by Goncalves and White (2004), for dependent and heterogeneous series.

we focus on 1-step ahead prediction (although results can be extended quite easily to multiple step ahead prediction), and so that recursive estimators are thus subsequently computed using $R+1$ observations, and then $R+2$ observations, and so on, until the last estimator is constructed using $T-1$ observations. This results in a sequence of $P = T-R$ estimators. These estimators can then used to construct sequences of P 1-step ahead forecasts and associated forecast errors, for example.

Now, consider the overlapping block resampling scheme of Künsch (1989), which can be applied in our context as follows:[4] At each replication, draw b blocks (with replacement) of length l from the sample $W_t = (y_t, Z^{t-1})$, where $bl = T-s$. Thus, the first block is equal to $W_{i+1}, ..., W_{i+l}$, for some $i = s-1, ..., T-l+1$, with probability $1/(T-s-l+1)$, the second block is equal to $W_{i+1}, ..., W_{i+l}$, again for some $i = s-1, ..., T-l+1$, with probability $1/(T-s-l+1)$, and so on, for all blocks, where the block length grows with the sample size at an appropriate rate. More formally, let I_k, $k = 1, ..., b$ be *iid* discrete uniform random variables on $[s-1, s, ..., T-l+1]$. Then, the resampled series, $W_t^* = (y_t^*, Z^{*,t-1})$, is such that $W_1^*, W_2^*, ..., W_l^*, W_{l+1}^*, ..., W_T^* = W_{I_1+1}, W_{I_1+2}, ..., W_{I_1+l}, W_{I_2}, ..., W_{I_b+l}$, and so a resampled series consists of b blocks that are discrete *iid* uniform random variables, conditional on the sample.

Suppose we define the bootstrap estimator, $\widehat{\theta}_t^*$, to be the direct analog of $\widehat{\theta}_t$. Namely,

$$\widehat{\theta}_t^* = \arg \min_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=s}^{t} q(y_j^*, Z^{*,j-1}, \theta), \;\; R \leq t \leq T-1. \tag{3}$$

By first order conditions, $\frac{1}{t}\sum_{j=s}^{t} \nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t^*) = 0$, and via a mean value expansion of

[4]The main difference between the block bootstrap and the stationary bootstrap of Politis and Romano (PR:1994) is that the former uses a deterministic block length, which may be either overlapping as in Künsch (1989) or non-overlapping as in Carlstein (1986), while the latter resamples using blocks of random length. One important feature of the PR bootstrap is that the resampled series, conditional on the sample, is stationary, while a series resampled from the (overlapping or non overlapping) block bootstrap is nonstationary, even if the original sample is strictly stationary. However, Lahiri (1999) shows that all block boostrap methods, regardless of whether the block length is deterministic or random, have a first order bias of the same magnitude, but the bootstrap with deterministic block length has a smaller first order variance. In addition, the overlapping block boostrap is more efficient than the non overlapping block bootstrap.

$\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t^*)$ around $\widehat{\theta}_t$, after a few simple manipulations, we have that

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t^* - \widehat{\theta}_t\right) \\
&= \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta^2 q(y_j^*, Z^{*,j-1}, \overline{\theta}_t^*)\right)^{-1}\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t)\right) \\
&= B^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t)\right) + o_{P^*}(1)\ \Pr -P \\
&= B^\dagger\frac{a_{R,0}}{\sqrt{P}}\sum_{t=s}^{R}\nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t) + B^\dagger\frac{1}{\sqrt{P}}\sum_{j=1}^{P-1}a_{R,j}\nabla_\theta q(y_{R+j}^*, Z^{*,R+j-1}, \widehat{\theta}_t) \\
&\quad + o_{P^*}(1)\ \Pr -P, \qquad (4)
\end{aligned}$$

where $\overline{\theta}_t^* \in \left(\widehat{\theta}_t^*, \widehat{\theta}_t\right)$, $B^\dagger = E\left(\nabla_\theta^2 q(y_j, Z^{j-1}, \theta^\dagger)\right)^{-1}$, $a_{R,j} = \frac{1}{R+j} + \frac{1}{R+j+1} + ... + \frac{1}{R+P-1}$, $j = 0, 1, ..., P-1$, and where the last equality on the right hand side of (4) follows immediately, using the same arguments as those used in Lemma A5 of West (1996). Analogously,

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right) \\
&= B^\dagger\frac{a_{R,0}}{\sqrt{P}}\sum_{t=s}^{R}\nabla_\theta q(y_j, Z^{j-1}, \theta^\dagger) + B^\dagger\frac{1}{\sqrt{P}}\sum_{j=1}^{P-1}a_{R,j}\nabla_\theta q(y_{R+j}, Z^{R+j-1}, \theta^\dagger) + o_P(1). \qquad (5)
\end{aligned}$$

Now, given (2), $E\left(\nabla_\theta q(y_j, Z^{j-1}, \theta^\dagger)\right) = 0$ for all j, and $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$ has a zero mean normal limiting distribution (see Theorem 4.1 in West (1996)). On the other hand, as any block of observations has the same chance of being drawn,

$$E^*\left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t)\right) = \frac{1}{T-s}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t) + O\left(\frac{l}{T}\right)\ \Pr -P, \qquad (6)$$

where the $O\left(\frac{l}{T}\right)$ term arises because the first and last l observations have a lesser chance of being drawn (see e.g. Fitzenberger (1997)).[5] Now, $\frac{1}{T-s}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t) \neq 0$, and is instead of order $O_P\left(T^{-1/2}\right)$. Thus, $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\frac{1}{T-s}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t) = O_P(1)$, and does not vanish in probability. This clearly contrasts with the full sample case, in which $\frac{1}{T-s}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_T) = 0$, because of the first order conditions. Thus, $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t^* - \widehat{\theta}_t\right)$ cannot have a zero mean normal

[5] In fact, the first and last observation in the sample can appear only at the beginning and end of the block, for example.

limiting distribution, but is instead characterized by a location bias that can be either positive or negative depending on the sample.

Given (6), our objective is thus to have the bootstrap score centered around $\frac{1}{T-s}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t)$. Hence, define a new bootstrap estimator, $\widetilde{\theta}_t^*$, as:

$$\widetilde{\theta}_t^* = \arg\min_{\theta\in\Theta} \frac{1}{t}\sum_{j=s}^{t}\left(q(y_j^*, Z^{*,j-1}, \theta) - \theta'\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t)\right)\right), \tag{7}$$

$R \leq t \leq T-1$.[6]

Given first order conditions, $\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \widetilde{\theta}_t^*) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t)\right)\right) = 0$, and via a mean value expansion of $\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta q(y_j^*, Z^{*,j-1}, \widetilde{\theta}_t^*)$ around $\widehat{\theta}_t$, after a few simple manipulations, we have that

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}_t^* - \widehat{\theta}_t\right) \\
= \;& B^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t)\right)\right)\right) \\
&+ o_{P^*}(1) \;\; \Pr - P.
\end{aligned}$$

Given (6), it is immediate to see that the bias associated with $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}_t^* - \widehat{\theta}_t\right)$ is of order $O\left(lT^{-1/2}\right)$, conditional on the sample, and so it is negligible for first order asymptotics, as $l = o(T^{1/2})$.

Theorem 1, which summarizes these results, requires the following assumptions.

Assumption A1: (y_t, X_t), with y_t scalar and X_t an $R^\zeta-$valued $(0 < \zeta < \infty)$ vector, is a strictly stationary and absolutely regular $\beta-$mixing process with size $-4(4+\psi)/\psi$, $\psi > 0$.

Assumption A2: (i) $\theta^\dagger$ is uniquely identified (i.e. $E(q(y_t, Z^{t-1}, \theta)) > E(q(y_t, Z^{t-1}, \theta^\dagger))$ for any $\theta \neq \theta^\dagger$); (ii) q is twice continuously differentiable on the interior of Θ, and for Θ a compact subset of R^ϱ; (iii) the elements of $\nabla_\theta q$ and $\nabla_\theta^2 q$ are $p-$dominated on Θ, with $p > 2(2+\psi)$, where ψ is the same positive constant as defined in Assumption A1; and (iv) $E\left(-\nabla_\theta^2 q(\theta)\right)$ is negative definite

[6] More precisely, we should define

$$\widetilde{\theta}_{i,t}^* = \arg\min_{\theta_i\in\Theta_i} \frac{1}{t-s}\sum_{j=s}^{t}\left(q_i(y_j^*, Z^{*,j-1}, \theta_i) - \theta_i'\left(\frac{1}{T-s}\sum_{k=s}^{T-1}\nabla_{\theta_i} q_i(y_k, Z^{k-1}, \widehat{\theta}_{i,t})\right)\right)$$

However, for notational simplicity we approximate $\frac{1}{t-s}$ and $\frac{1}{T-s}$ with $\frac{1}{t}$ and $\frac{1}{T}$.

uniformly on Θ.[7]

Assumption A3: $T = R + P$, and as $T \to \infty$, $P/R \to \pi$, with $0 < \pi < \infty$.

Assumptions A1 and A2 are standard memory, moment, smoothness and identifiability conditions. A1 requires (y_t, X_t) to be strictly stationary and absolutely regular. The memory condition is stronger than $\alpha-$mixing, but weaker than (uniform) $\phi-$mixing. Assumption A3 requires that R and P grow at the same rate. In fact, if P grows at a slower rate than R, i.e. $P/R \to 0$, then $\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widehat{\theta}_t - \theta^\dagger \right) = o_P(1)$ and so there were no need to capture the contribution of parameter estimation error.

Theorem 1: Let A1-A3 hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

$$P\left(\omega : \sup_{v \in \Re^\varrho} \left| P_T^* \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widetilde{\theta}_t^* - \theta^\dagger \right) \leq v \right) - P \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widehat{\theta}_{i,t} - \theta_i^\dagger \right) \leq v \right) \right| > \varepsilon \right) \to 0,$$

where P_T^* denotes the probability law of the resampled series, conditional on the (entire) sample.

Broadly speaking, Theorem 1 states that $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}_t^* - \theta^\dagger \right)$ has the same limiting distribution as $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_t - \theta^\dagger \right)$, conditional on sample, and for all samples except a set with probability measure approaching zero. As outlined in the following sections, application of Theorem 1 allows us to capture the contribution of (recursive) parameter estimation error to the covariance kernel of the limiting distribution of various statistics. If Assumption 3 is violated and $P/R \to 0$, then the statement in the Theorem above is trivially satisfied, in the sense that both $\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widetilde{\theta}_t^* - \theta^\dagger \right)$ and $\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widehat{\theta}_{i,t} - \theta_i^\dagger \right)$ have a limiting distribution degenerate on zero.

3 The Reality Check

In this section, we extend the White (2000) reality check to the case in which the effect of parameter estimation error does not vanish asymptotically. In particular, we show that the block bootstrap for recursive m-estimators properly mimics the contribution of parameter estimation error to the covariance kernel of the limiting distribution of the original reality check test. Although we focus our attention in this paper on the block bootstrap, which is based on resampling blocks of deterministic

[7] We say that $\nabla_\theta q(y_t, Z^{t-1}, \theta)$ is $2r-$dominated on Θ if its $j-th$ element, $j = 1, ..., \varrho$, is such that $\left| \nabla_\theta q(y_t, Z^{t-1}, \theta) \right|_j \leq D_t$, and $E(|D_t|^{2r}) < \infty$. For more details on domination conditions, see Gallant and White (1988, pp. 33).

length, we conjecture that the same approach can be used to extend the stationary bootstrap employed by White (2000) to the case of nonvanishing parameter estimation error.

Let the forecast error be $u_{i,t+1} = y_{t+1} - \kappa_i(Z^t, \theta_i^\dagger)$, and let $\widehat{u}_{i,t+1} = y_{t+1} - \kappa_i(Z^t, \widehat{\theta}_{i,t})$, where $\kappa_i(Z^t, \widehat{\theta}_{i,t})$ is the estimated conditional mean function under model i. Also, assume that the set of regressors may vary across different models, so that Z^t is meant to denote the collection of all potential regressors. Following White (2000), define the statistic

$$S_P = \max_{k=2,...,n} S_P(1,k),$$

where

$$S_P(1,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right), \ k = 2, ..., n,$$

and where g is a given loss function (see Christoffersen and Diebold (1996,1997) and Elliott and Timmerman (2004a,b) for a detailed discussion of loss functions used in predictive evaluation). Recall that in this test, parameter estimation error need not be accounted for in the covariance kernel of the limiting distribution unless $g \neq q_i$ for some i. This follows upon examination of the results of both West (1996) and White (2000). In particular, in West (1996), the parameter estimation error components that enter into the covariance kernel of the limiting distribution of his test statistic are zero whenever the same loss function is used for both predictive evaluation and in-sample estimation. The same argument holds for the reality check test. This means that as long as $g = q_i \ \forall i$, the White test can be applied regardless of the rate of growth of P and R. When we write the covariance kernel of the limiting distribution of the statistic (see below), however, we include terms capturing the contribution of parameter estimation error, thus implicitly assuming that $g \neq q_i$ for some i. In practice, one reason why we allow for cases where $g \neq q_i$ is that least squares is sometimes better behaved in finite samples and/or easier to implement than more generic $m-$estimators, so that practitioners sometimes use least squares for estimation and more complicated (possibly asymmetric) loss functions for predictive evaluation.[8] Of course, there are also applications for which parameter estimation error does not vanish, even if the same loss function

[8]Consider linex loss, where $g(u) = e^{au} - au - 1$, so that for $a > 0$ ($a < 0$) positive (negative) errors are more (less) costly than negative (positive) errors. Here, the errors are exponentiated, so that in this particular case, laws of large numbers and central limit theorems may require a large number of observations before providing satisfactory approximations. This feature of linex loss is illustrated in the Monte Carlo findings of Corradi and Swanson (2002). (Linex loss is studied in Zellner (1986), Christoffersen and Diebold (1996, 1997) and Granger (1999), for example.)

is used for parameter estimation and predictive evaluation. One such application is discussed in the next section.

For a given loss function, the reality check tests the null hypothesis that a benchmark model (defined as model 1) performs equal to or better than all competitor models (i.e. models 2,...,n). The alternative is that at least one competitor performs better than the benchmark.[9] Formally, the hypotheses are:

$$H_0 : \max_{k=2,...,n} E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) \leq 0$$

and

$$H_A : \max_{k=2,...,n} E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) > 0.$$

In order to derive the limiting distribution of S_P we require the following additional assumption.

Assumption A4: (i) κ_i is twice continuously differentiable on the interior of Θ_i and the elements of $\nabla_{\theta_i}\kappa_i(Z^t, \theta_i)$ and $\nabla^2_{\theta_i}\kappa_i(Z^t, \theta_i)$ are $p-$dominated on Θ_i, for $i = 2, ..., n$, with $p > 2(2+\psi)$, where ψ is the same positive constant as that defined in Assumption A1; (ii) g is positive valued, twice continuously differentiable on Θ_i, and g, g' and g'' are $p-$dominated on Θ_i with p defined as in (i); and (iii) let $c_{kk} =$
$\lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, $k = 2, ..., n$, define analogous covariance terms, $c_{j,k}$, $j, k = 2, ..., n$, and assume that $[c_{j,k}]$ is positive semi-definite.

Assumptions A4(i)-(ii) are standard smoothness and domination conditions imposed on the conditional mean functions of the models. Assumption A4(iii) is standard in the literature that uses DM type tests (see e.g. West (1996) and White (2000)), and states that at least one of the competing models has to be nonnested with (and not nesting) the benchmark.

Proposition 2: Let Assumptions A1-A4 hold. Then, as $P, R \to \infty$,

$$\max_{k=2,...,n}\left(S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right) \xrightarrow{d} \max_{k=2,...,n} S(1,k),$$

where $S = (S(1,2), ..., S(1,n))$ is a zero mean Gaussian process with covariance kernel given by V, with V a $n \times n$ matrix with i, i element
$v_{i,i} = S_{g_i g_i} + 2\Pi\mu_1' B_1^\dagger C_{11} B_1^\dagger \mu_1 + 2\Pi\mu_i' B_i^\dagger C_{ii} B_i^\dagger \mu_i - 4\Pi\mu_1' B_1^\dagger C_{1i} B_i^\dagger \mu_i + 2\Pi S_{g_{iq_1}} B_1^\dagger \mu_1 - 2\Pi S_{g_{iq_i}} B_i^\dagger \mu_i$,
where $S_{g_i g_i} = \sum_{\tau=-\infty}^{\infty} E\left(\left(g(u_{1,1}) - g(u_{i,1})\right)\left(g(u_{1,1+\tau}) - g(u_{i,1+\tau})\right)\right)$,

[9] In the current context, we are interested in choosing the model which is more accurate for given loss function. An alternative approach is to combine different forecasting models in some optimal way. For very recent contributions along these lines, see Elliott and Timmermann (2004a,b).

$C_{ii} = \sum_{\tau=-\infty}^{\infty} E\left(\left(\nabla_{\theta_i} q_i(y_{1+s}, Z^s, \theta_i^\dagger)\right)\left(\nabla_{\theta i} q_i(y_{1+s+\tau}, Z^{s+\tau}, \theta_i^\dagger)\right)'\right)$,

$S_{g_{iq_i}} = \sum_{\tau=-\infty}^{\infty} E\left((g(u_{1,1}) - g(u_{i,1}))\left(\nabla_{\theta i} q_i(y_{1+s+\tau}, Z^{s+\tau}, \theta_i^\dagger)\right)'\right)$,

$B_i^\dagger = \left(E\left(-\nabla_{\theta i}^2 q_i(y_t, Z^{t-1}, \theta_i^\dagger)\right)\right)^{-1}$, $\mu_i = E\left(\nabla_{\theta_i} g(u_{i,t+1})\right)$, and $\Pi = 1 - \pi^{-1}\ln(1+\pi)$.

Just as in White (2000), note that under the null, the least favorable case arises when $E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) = 0$, $\forall\ k$. In this case, the distribution of S_P coincides with that of $\max_{k=2,\ldots,n}\left(S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, so that S_P has the above limiting distribution, which is a functional of a Gaussian process with a covariance kernel that reflects uncertainty due to parameter estimation error and dynamic misspecification. Additionally, when all competitor models are worse than the benchmark, the statistic diverges to minus infinity at rate $\sqrt{P}$. Finally, when only some competitor models are worse than the benchmark, the limiting distribution provides a conservative test, as S_P will always be smaller than $\max_{k=2,\ldots,n}\left(S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, asymptotically. Of course, when H_A holds, the statistic diverges to plus infinity at rate $\sqrt{P}$.[10]

Recall that the maximum of a Gaussian process is not Gaussian in general, so that standard critical values cannot be used to conduct inference on S_P. As pointed out by White (2000), one possibility in this case is to first estimate the covariance structure and then draw 1 realization from an $(n-1)$-dimensional normal with covariance equal to the estimated covariance structure. From this realization, pick the maximum value over $k = 2, \ldots, n$. Repeat this a large number of times, form an empirical distribution using the maximum values over $k = 2, \ldots, n$, and obtain critical values in the usual way. A drawback to this approach is that we need to rely on an estimator of the covariance structure based on the available sample of observations, which in many cases may be small relative to the number of models being compared. Furthermore, whenever the

[10] For more discussion of the properties of this variety of test, the reader is referred to Corradi and Swanson (2004a,2005), and the references cited therein. Amongst other approaches, one approach discussed in these papers is the construction of critical values based on subsampling (e.g. Politis, Romano and Wolf (1999), Ch.3). Heuristically, we construct $T - 2b_T$ statistics using subsamples of length b_T, where $b_T/T \to 0$; the empirical distribution of the statistics computed over the various subsamples, properly mimics the distribution of the statistic. Thus, it provides valid critical values even for the case of $\max_{k=2,\ldots,m} E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) = 0$, but $E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) < 0$ for some k. Needless to say, the problem is that unless the sample is very large, the empirical distribution of the subsampled statistics provides a poor approximation to the limiting distribution of the statistic. The subsampling approach has been followed by Linton, Maasoumi and Whang (2004), in the context of testing for stochastic dominance.

forecasting errors are not martingale difference sequences (as in our context, given that we wish to allow all models to be possibly misspecified), heteroskedasticity and autocorrelation consistent covariance matrices should be estimated, and thus a lag truncation parameter must be chosen. As mentioned above, another approach which avoids these problems involves using the stationary bootstrap of Politis and Romano (1994), which was done by White (2000) for the case in which parameter estimation error vanishes asymptotically. In general, bootstrap procedures have been shown to perform well in a variety of finite sample contexts (see e.g. Diebold and Chen (1996)). Our approach is to apply the block bootstrap for recursive m-estimators outlined above.

Define the bootstrap parameter estimator as:

$$\widetilde{\theta}^*_{i,t} = \arg\min_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=s}^{t} \left(q_i(y^*_j, Z^{*,j-1}, \theta_i) - \theta_i' \left(\frac{1}{T} \sum_{h=s}^{T-1} \nabla_{\theta_i} q_i(y_h, Z^{h-1}, \widehat{\theta}_{i,t}) \right) \right), \tag{8}$$

where $R \leq t \leq T-1$, $i = 1, ..., n$; and define the bootstrap statistic as:

$$S^*_P = \max_{k=2,...,n} S^*_P(1,k),$$

where

$$\begin{aligned} S^*_P(1,k) &= \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left[\left(g(y^*_{t+1} - \kappa_1(Z^{*,t}, \widetilde{\theta}^*_{1,t})) - g(y^*_{t+1} - \kappa_k(Z^{*,t}, \widetilde{\theta}^*_{k,t})) \right) \right. \\ &\quad \left. - \left\{ \frac{1}{T} \sum_{j=s}^{T-1} \left(g(y_{j+1} - \kappa_1(Z^j, \widehat{\theta}_{1,t})) - g(y_{j+1} - \kappa_k(Z^j, \widehat{\theta}_{k,t})) \right) \right\} \right]. \end{aligned} \tag{9}$$

Note that bootstrap statistic in (9) is different from the "usual" bootstrap statistic, which is defined as the difference between the statistic computed over the sample observations and over the bootstrap observations. That is, following the usual approach to bootstrap statistic construction, one might have expected that the appropriate bootstrap statistic would be:

$$\begin{aligned} \overline{S}^*_P(1,k) &= \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left[\left(g(y^*_{t+1} - \kappa_1(Z^{*,t}, \widetilde{\theta}^*_{1,t})) - g(y^*_{t+1} - \kappa_k(Z^{*,t}, \widetilde{\theta}^*_{k,t})) \right) \right. \\ &\quad \left. - \left(g(y_{t+1} - \kappa_1(Z^t, \widehat{\theta}_{1,t})) - g(y_{t+1} - \kappa_k(Z^t, \widehat{\theta}_{k,t})) \right) \right]. \end{aligned} \tag{10}$$

Instead, as can be seen by inspection of $S^*_P(1,k)$, the bootstrap (resampled) component is constructed only over the last P observations, while the sample component is constructed over all T observations. Although a formal proof is provided in the appendix, it is worthwhile to give a

heuristic explanation of the validity of the statistic in (9). For sake of simplicity, consider a single model, say model 1. Now,

$$\begin{aligned} & \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(g(y_{t+1}^{*}-\kappa_1(Z^{*,t},\widetilde{\theta}_{1,t}^{*}))-\frac{1}{T}\sum_{j=s}^{T-1}g(y_{j+1}-\kappa_1(Z^{j},\widehat{\theta}_{1,t}))\right) \\ = & \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(g(y_{t+1}^{*}-\kappa_1(Z^{*,t},\widehat{\theta}_{1,t}))-\frac{1}{T}\sum_{j=s}^{T-1}g(y_{j+1}-\kappa_1(Z^{j},\widehat{\theta}_{1,t}))\right) \\ & +\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\nabla_{\theta}g(y_{t+1}^{*}-\kappa_1(Z^{*,t},\overline{\theta}_{1,t}^{*}))\left(\widetilde{\theta}_{1,t}^{*}-\widehat{\theta}_{1,t}\right), \end{aligned} \tag{11}$$

where $\overline{\theta}_{1,t}^{*}\in\left(\widetilde{\theta}_{1,t}^{*},\widehat{\theta}_{1,t}\right)$. Notice that the first term on the RHS of (11) mimics the limiting behavior of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(g(u_{1,t+1})-E(g(u_{1,t+1}))\right)$, while the second term mimics the limiting behavior of the parameter estimation error associated with model 1. Needless to say, the same holds for any arbitrary model. This leads to the following proposition.

Proposition 3: Let Assumptions A1-A4 hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

$$P\left(\omega : \sup_{v \in \Re} \left| P_T^* \left(\max_{k=2,...,n} S_P^*(1,k) \leq v \right) - P\left(\max_{k=2,...n} S_P^{\mu}(1,k) \leq v \right) \right| > \varepsilon \right) \to 0,$$

and

$$S_P^{\mu}(1,k) = S_P(1,k) - \sqrt{P} E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right),$$

The above result suggests proceeding in the following manner. For any bootstrap replication, compute the bootstrap statistic, S_P^*. Perform B bootstrap replications (B large) and compute the quantiles of the empirical distribution of the B bootstrap statistics. Reject H_0, if S_P is greater than the $(1-\alpha)th$-percentile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, S_P has the same limiting distribution as the corresponding bootstrapped statistic when $E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) = 0 \; \forall \; k$, ensuring asymptotic size equal to α. On the other hand, when one or more competitor models are strictly dominated by the benchmark, the rule provides a test with asymptotic size between 0 and α (see above discussion). Under the alternative, S_P diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution, ensuring unit asymptotic power. A potential problem with the reality check is that, when the number of dominated models increases, then $P-$values go up. Hansen (2004) has provided a modified version of the White's reality check, which alleviates this problem. His idea is to use a bootstrap procedure which mimics a limiting normal distribution with an appropriate negative mean, instead of a zero mean. In such a way, also the dominated models contribute to the limiting distribution. Hansen's correction affects only the bootstrap statistics and not the actual one. We conjecture that his approach does not carry through to the case of non vanishing parameter estimation error. The intuitive is reason is that, with the Hansen's modification, the contribution of parameter estimation error in the bootstrap statistics does no longer mimic that of the original statistic. This appears from inspection of his Corollary 6.

In summary, this application shows that the block bootstrap for recursive m-estimators can be readily adapted in order to provide asymptotically valid critical values that are robust to parameter estimation error as well as model misspecification. In addition, the bootstrap statistics are very easy to construct, as no complicated adjustment terms involving possibly higher order derivatives need be included.

4 The Out-of-Sample Integrated Conditional Moment Test

Corradi and Swanson (CS: 2002) draw on both the consistent specification and predictive ability testing literatures in order to propose a test for predictive accuracy which is consistent against generic nonlinear alternatives, and which is designed for comparing nested models. The CS test is based on an out-of-sample version of the ICM test of Bierens (1982,1990) and Bierens and Ploberger (1997). This test is relevant for model selection, as it is well known that DM and reality check tests do not have a well defined limiting distributions when the benchmark is nested with all competing models and when $P/R \to 0$ (see e.g. Corradi and Swanson (2002, 2004b), McCracken (2004) and Clark and McCracken (2004)).[11] Alternative (non DM) tests for comparing the predictive ability of a fixed number of nested models have previously also been suggested. For example, Clark and McCracken (2001,2004) propose encompassing tests for comparing two nested models for one-step and multi-step ahead prediction, respectively. Chao, Corradi and Swanson (2001) propose a test which allows for dynamic misspecification under the null hypothesis. Recently, Giacomini and White (2003) introduce a test for conditional predictive ability that is valid for both nested and nonnested models. The key ingredient of their test is the fact that parameters are estimated using a fixed rolling window. Finally, Inoue and Rossi (2004) suggest a recursive test, where not only the parameters, but the statistic itself, are computed in a recursive manner.
The main difference between these tests and the CS test is that the CS test is consistent against generic (non)linear alternatives and not only against a fixed alternative.

As shown in the appendix, the limiting distribution of the ICM type test statistic proposed by CS is a functional of a Gaussian process with a covariance kernel that reflects both the time series structure of the data as well as the contribution of parameter estimation error. As a consequence, critical values are data dependent and cannot be directly tabulated. CS establish the validity of the conditional p-value method for constructing critical values in this context, thus extending earlier work by Hansen (1996) and Inoue (2001). However, the conditional p-value approach suffers from the fact that under the alternative, the simulated statistic diverges (at rate as high as $\sqrt{\widetilde{l}}$), conditional on the sample and for all samples except a set of measure zero, where $\widetilde{l}$ plays a role

[11] McCracken (2004) provides a very interesting result based on a particular version of the DM test (in which loss is quadratic and martingale difference scores are assumed - i.e. it is assumed that the model is correctly specified under the null hypothesis) has a nonstandard limiting distribution which is a functional of Brownian motions, whenever $P/R \to \pi > 0$. Clark and McCracken (2004) extend McCracken (2004) to the case of multi step ahead forecasts.

analogous to l in the block bootstrap. As this feature may lead to reduced power in finite samples, we establish in this application that the block bootstrap for recursive m-estimators can be used to provide easy to compute and asymptotically valid critical values for the CS test.

Summarizing the testing approach considered in this application, assume that the objective is to test whether there exists any unknown alternative model that has better predictive accuracy than a given benchmark model, for a given loss function. A typical example is the case in which the benchmark model is a simple autoregressive model and we want to check whether a more accurate forecasting model can be constructed by including possibly unknown (non)linear functions of the past of the process or of the past of some other process (e.g. out-of-sample (non)linear Granger causality tests can be constructed in this manner).[12] Although this is the case that we focus on, the benchmark model can in general be any (non)linear model. One important feature of this application is that the same loss function is used for in-sample estimation and out-of-sample prediction (see Granger (1993), Weiss (1996), and Schorfheide (2004) for further discussion of this issue)[13]. In contrast to the previous application, however, this does not ensure that parameter estimation error vanishes asymptotically.

Let the benchmark model be:

$$y_t = \theta_{1,1}^{\dagger} + \theta_{1,2}^{\dagger} y_{t-1} + u_{1,t}, \tag{12}$$

where $\theta_1^{\dagger} = (\theta_{1,1}^{\dagger}, \theta_{1,2}^{\dagger})' = \arg\min_{\theta_1 \in \Theta_1} E(q_1(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1}))$, $\theta_1 = (\theta_{1,1}, \theta_{1,2})'$, y_t is a scalar, and $q_1 = g$, as the same loss function is used both for in-sample estimation and out-of-sample predictive evaluation. The generic alternative model is:

$$y_t = \theta_{2,1}^{\dagger}(\gamma) + \theta_{2,2}^{\dagger}(\gamma) y_{t-1} + \theta_{2,3}^{\dagger}(\gamma) w(Z^{t-1}, \gamma) + u_{2,t}(\gamma), \tag{13}$$

where $\theta_2^{\dagger}(\gamma) = (\theta_{2,1}^{\dagger}(\gamma), \theta_{2,2}^{\dagger}(\gamma), \theta_{2,3}^{\dagger}(\gamma))' = \arg\min_{\theta_2 \in \Theta_2} E(q_1(y_t - \theta_{2,1} - \theta_{2,2} y_{t-1} - \theta_{2,3} w(Z^{t-1}, \gamma)))$, $\theta_2(\gamma) = (\theta_{2,1}(\gamma), \theta_{2,2}(\gamma), \theta_{2,3}(\gamma))'$, $\theta_2 \in \Theta_2$, Γ is a compact subset of $\Re^d$, for some finite d. The alternative model is called "generic" because of the presence of $w(Z^{t-1}, \gamma)$, which is a generically comprehensive function, such as Bierens' exponential, a logistic, or a cumulative distribution

[12] For example, Swanson and White (1997) compare the predictive accuracy of various linear models against neural network models using both in-sample and out-of-sample model selection criteria.

[13] In the context of multi-step ahead vector autoregressive prediction, Schörfheide (2004) proposes a new prediction criterion that can be used to jointly select the number of lags as well as to choose between (quasi)-maximum likelihood estimators and loss function based estimators.

function (see e.g. Stinchcombe and White (1998) for a detailed explanation of generic comprehensiveness). One example has $w(Z^{t-1},\gamma) = \exp(\sum_{i=1}^{s_2}\gamma_i\Phi(X_{t-i}))$, where Φ is a measurable one to one mapping from $\Re$ to a bounded subset of $\Re$, so that here $Z^t = (X_t, ..., X_{t-s_2+1})$, and we are thus testing for nonlinear Granger causality. The hypotheses of interest are:

$$H_0 : E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) = 0 \text{ versus } H_A : E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) > 0. \tag{14}$$

Clearly, the reference model is nested within the alternative model, and given the definitions of $\theta_1^\dagger$ and $\theta_2^\dagger(\gamma)$, the null model can never outperform the alternative.[14] For this reason, H_0 corresponds to equal predictive accuracy, while H_A corresponds to the case where the alternative model outperforms the reference model, as long as the errors above are loss function specific forecast errors. It follows that H_0 and H_A can be restated as:

$$H_0 : \theta_{2,3}^\dagger(\gamma) = 0 \text{ versus } H_A : \theta_{2,3}^\dagger(\gamma) \neq 0,$$

for $\forall\gamma \in \Gamma$, except for a subset with zero Lebesgue measure. Now, given the definition of $\theta_2^\dagger(\gamma)$, note that

$$E\left(g'(y_{t+1} - \theta_{2,1}^\dagger(\gamma) - \theta_{2,2}^\dagger(\gamma)y_t - \theta_{2,3}^\dagger(\gamma)w(Z^t,\gamma)) \times \begin{pmatrix} -1 \\ -y_t \\ -w(Z^t,\gamma) \end{pmatrix}\right) = 0,$$

where g' is the derivative of the loss function with respect to its argument. Thus, under H_0 we have that $\theta_{2,3}^\dagger(\gamma) = 0$, $\theta_{2,1}^\dagger(\gamma) = \theta_{1,1}^\dagger$, $\theta_{2,2}^\dagger(\gamma) = \theta_{1,2}^\dagger$, and $E(g'(u_{1,t+1})w(Z^t,\gamma)) = 0$. Thus, we can once again restate H_0 and H_A as:

$$H_0 : E(g'(u_{1,t+1})w(Z^t,\gamma)) = 0 \text{ versus } H_A : E(g'(u_{1,t+1})w(Z^t,\gamma)) \neq 0, \tag{15}$$

for $\forall\gamma \in \Gamma$, except for a subset with zero Lebesgue measure. Finally, define the forecast error as $\widehat{u}_{1,t+1} = y_{t+1} - \begin{pmatrix} 1 & y_t \end{pmatrix}\widehat{\theta}_{1,t}$. Following CS, the test statistic is:

$$M_P = \int_\Gamma m_P(\gamma)^2\phi(\gamma)d\gamma, \tag{16}$$

where

$$m_P(\gamma) = \frac{1}{P^{1/2}}\sum_{t=R}^{T-1} g'(\widehat{u}_{1,t+1})w(Z^t,\gamma), \tag{17}$$

[14]Needless to say, in finite sample the forecasting mean square prediction error from the small model can be lower than that associated with the larger model. Indeed, this feature appears in our empirical illustration.

and where $\int_\Gamma \phi(\gamma)d\gamma = 1$, $\phi(\gamma) \geq 0$, with $\phi(\gamma)$ absolutely continuous with respect to Lebesgue measure. In the sequel, we require the following assumptions.

Assumption A5: (i) w is a bounded, twice continuously differentiable function on the interior of Γ and $\nabla_\gamma w(Z^t, \gamma)$ is bounded uniformly in Γ; and (ii) $\nabla_\gamma \nabla_{\theta_1} q'_{1,t}(\theta_1) w(Z^{t-1}, \gamma)$ is continuous on $\Theta_1 \times \Gamma$, where $q'_{1,t}(\theta_1) = q'_1(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1})$, Γ a compact subset of R^d, and is $2r-$dominated uniformly in $\Theta_1 \times \Gamma$, with $r \geq 2(2+\psi)$, where ψ is the same positive constant as that defined in Assumption A1.

Assumption A5 requires the function w to be bounded and twice continuously differentiable; such a requirement is satisfied by logistic or exponential functions, for example.

Proposition 4: Let Assumptions A1-A3 and A5 hold. Then, the following results hold: (i) Under H_0,

$$M_P = \int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma \xrightarrow{d} \int_\Gamma Z(\gamma)^2 \phi(\gamma) d\gamma,$$

where $m_P(\gamma)$ is defined in equation (17) and Z is a Gaussian process with covariance kernel given by:

$$\begin{aligned} K(\gamma_1, \gamma_2) &= S_{gg}(\gamma_1, \gamma_2) + 2\Pi \mu'_{\gamma_1} B^\dagger S_{hh} B^\dagger \mu_{\gamma_2} + \Pi \mu'_{\gamma_1} B^\dagger S_{gh}(\gamma_2) \\ &\quad + \Pi \mu'_{\gamma_2} B^\dagger S_{gh}(\gamma_1), \end{aligned}$$

with $\mu_{\gamma_1} = E(\nabla_{\theta_1}(g'_{t+1}(u_{1,t+1}) w(Z^t, \gamma_1)))$, $B^\dagger = (-E(\nabla^2_{\theta_1} q_1(u_{1,t})))^{-1}$,
$S_{gg}(\gamma_1, \gamma_2) = \sum_{j=-\infty}^{\infty} E(g'(u_{1,s+1}) w(Z^s, \gamma_1) g'(u_{1,s+j+1}) w(Z^{s+j}, \gamma_2))$,
$S_{hh} = \sum_{j=-\infty}^{\infty} E(\nabla_{\theta_1} q_1(u_{1,s}) \nabla_{\theta_1} q_1(u_{1,s+j})')$,
$S_{gh}(\gamma_1) = \sum_{j=-\infty}^{\infty} E(g'(u_{1,s+1}) w(Z^s, \gamma_1) \nabla_{\theta_1} q_1(u_{1,s+j})')$, and γ, γ_1, and γ_2 are generic elements of Γ.

(ii) Under H_A, for $\varepsilon > 0$, $\lim_{P\to\infty} \Pr\left(\frac{1}{P} \int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma > \varepsilon\right) = 1$.

As in the previous application, the limiting distribution under H_0 is a Gaussian process with a covariance kernel that reflects both the dependence structure of the data and the effect of parameter estimation error. Hence, critical values are data dependent and cannot be tabulated.

In order to implement this statistic using the block bootstrap for recursive m-estimators, define[15]

[15]Recall that $y^*_t, Z^{*,t}$ is obtained via the resampling procedure described in Section 2

$$\widetilde{\theta}^*_{1,t} = (\widetilde{\theta}^*_{1,1,t}, \widetilde{\theta}^*_{1,2,t})' = \arg\min_{\theta_1 \in \Theta_1} \frac{1}{t}\sum_{j=2}^{t} [q_1(y^*_j - \theta_{1,1} - \theta_{1,2} y^*_{j-1})$$
$$-\theta_1' \frac{1}{T}\sum_{i=2}^{T-1} \nabla_\theta q_1(y_i - \widehat{\theta}_{1,1,t} - \widehat{\theta}_{1,2,t} y_{i-1})] \quad (18)$$

Also, define $\widetilde{u}^*_{1,t+1} = y^*_{t+1} - \begin{pmatrix} 1 & y^*_t \end{pmatrix} \widetilde{\theta}^*_{1,t}$. The bootstrap test statistic is:

$$M^*_P = \int_\Gamma m^*_P(\gamma)^2 \phi(\gamma) d\gamma,$$

where, recalling that $g = q_1$,

$$m^*_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g' \left(y^*_{t+1} - \begin{pmatrix} 1 & y^*_t \end{pmatrix} \widetilde{\theta}^*_{1,t} \right) w(Z^{*,t}, \gamma) - \frac{1}{T}\sum_{i=1}^{T-1} g' \left(y_{i+1} - \begin{pmatrix} 1 & y_i \end{pmatrix} \widehat{\theta}_{1,t} \right) w(Z^i, \gamma) \right) \quad (19)$$

As in the reality check case, the bootstrap statistic in (19) is characterized by the fact that the bootstrap (resampled) component is constructed only over the last P observations, while the sample component is constructed over all T observations. The same heuristic arguments given to justify this form of bootstrap statistic in the previous application also apply here.[16]

Proposition 5: Let Assumptions A1-A3 and A5 hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

$$P\left(\omega : \sup_{v \in \Re} \left| P^*_T \left(\int_\Gamma m^*_P(\gamma)^2 \phi(\gamma) d\gamma \leq v \right) - P\left(\int_\Gamma m^\mu_P(\gamma)^2 \phi(\gamma) d\gamma \leq v \right) \right| > \varepsilon \right) \to 0,$$

where $m^\mu_P(\gamma) = m_P(\gamma) - \sqrt{P} E\left(g'(u_{1,t+1}) w(Z^t, \gamma) \right)$.

The above result suggests proceeding the same way as in the first application. For any bootstrap replication, compute the bootstrap statistic, M^*_P. Perform B bootstrap replications (B large) and compute the percentiles of the empirical distribution of the B bootstrap statistics. Reject H_0 if M_P is greater than the $(1-\alpha)th$-percentile. Otherwise, do not reject. Now, for all samples except a set with probability measure approaching zero, M_P has the same limiting distribution as the corresponding bootstrap statistic under H_0, thus ensuring asymptotic size equal to α. Under the alternative, M_P diverges to (plus) infinity, while the corresponding bootstrap statistic has a well defined limiting distribution, ensuring unit asymptotic power.

[16] Note that Monte Carlo experiments reported on in Section 5 examine other functionals of $m_P(\gamma)$, including $M^{\sup}_P = \sup_{\gamma \in \Gamma} |m_P(\gamma)|$ and $|M_P| = \int_\Gamma |m_P(\gamma)| \phi(\gamma) d\gamma$.

Propositions 2-5 have been derived under the assumption that P and R grow at the same rate as the sample size increases, see Assumption 3. In practice, we just observe P and R, but not their limit. In the uncertainty, it is still worthwhile to allow for parameter estimation error. In fact, the statement of Propositions 2-5 will be still valid, simply the contribution of parameter estimation error vanish in both the original and the bootstrap statistics.

5 Monte Carlo Results

In this section we carry out a series of Monte Carlo experiments comparing the recursive block bootstrap with a variety of other bootstraps, and comparing the finite sample performance of the test discussed above with a variety of other tests. With regard to the bootstrap, we consider 4 alternatives. Namely: (i) the "Recur Block Bootstrap", which is the block bootstrap for recursive m-estimators discussed above; (ii) "Block Bootstrap, no PEE, no adjust" is a strawman block bootstrap used for comparison purposes, where it is assumed that there is no parameter estimation error (PEE), so that $\widehat{\theta}_{1,t}$ is used in place of $\widetilde{\theta}^*_{1,t}$ in the construction of M^*_P, and the term $\frac{1}{T}\sum_{i=1}^{T-1} g'\left(y_{i+1} - \begin{pmatrix} 1 & y_i \end{pmatrix}\widehat{\theta}_{1,t}\right) w(Z^i,\gamma)$ in m^*_P is replaced with $g'\left(y_{t+1} - \begin{pmatrix} 1 & y_t \end{pmatrix}\widehat{\theta}_{1,t}\right) w(Z^t,\gamma)$ (i.e. there is no bootstrap statistic adjustment, thus conforming with the usual case when the standard block bootstrap is used); (iii) "Standard Block Bootstrap" is the standard block bootstrap (i.e. this bootstrap is the same as that outlined in (ii), except that $\widehat{\theta}_{1,t}$ is replaced with $\widehat{\theta}^*_{1,t}$; and (iv) the parametric bootstrap of Kilian (1999), applied in the spirit of McCracken and Sapp (MS: 2004). As in MS, application of the parametric bootstrap begins with the estimation of a VAR model for x_t and y_t (with lags selected using the Schwarz Information Criterion (SIC)) and resample the residuals as if they were *iid*. Then the pseudo time series x^*_t and y^*_t are constructed using estimated parameters and resampled residuals, and using the original VAR structure. At this point, x^*_t and y^*_t are used to estimate parameters recursively and construct a series of one-step ahead prediction errors (see below). Finally, bootstrap statistics are constructed exactly as are the original statistics, expect that prediction errors and variables are replaced with their bootstrapped counterparts. It should be pointed out that a necessary condition for the asymptotic validity of the Kilian (1999) parametric bootstrap is that the underlying DGP is nested by the VAR used in the bootstrap procedure. Thus, it is not in general valid if the underlying DGP contains nonlinear component(s). Furthermore, in our context, validity of this bootstrap still remains to be established

even in the case for which the DGP is nested by the VAR. This is because the bootstrap statistics are formed using one-step ahead forecast errors based on recursively estimated parameters. Therefore, standard arguments used to show the validity of the parametric bootstrap no longer apply. Nevertheless, it is interesting that the parametric bootstrap clearly still performs quite well in a variety of nonlinear contexts, as shown in the experiments reported on below.

The test statistics examined in our experiments include: (i) the standard in-sample F-test; (ii) the encompassing test due to Clark and McCracken (CM: 2004) and Harvey, Leybourne and Newbold (1997); (iii) the Diebold and Mariano (DM: 1995) test; (iv) a version of the M_P encompassing test defined in (16) and discussed above (called the CS test in the sequel); and (v) a linear version of the CS test due to Chao, Corradi and Swanson (CCS: 2001).[17]

To be more specific, note that the CM test is an out-of-sample encompassing test, and is defined as follows:

$$CM = (P-h+1)^{1/2} \frac{\frac{1}{P-h+1}\sum_{t=R}^{T-h} \widehat{c}_{t+h}}{\frac{1}{P-h+1}\sum_{j=-\overline{j}}^{\overline{j}} \sum_{t=R+j}^{T-h} K\left(\frac{j}{M}\right)\left(\widehat{c}_{t+h}-\overline{c}\right)\left(\widehat{c}_{t+h-j}-\overline{c}\right)},$$

where $\widehat{c}_{t+h} = \widehat{u}_{1,t+h}\left(\widehat{u}_{1,t+h} - \widehat{u}_{2,t+h}\right)$, $\overline{c} = \frac{1}{P-h+1}\sum_{t=R}^{T-\tau} \widehat{c}_{t+h}$, $K(\cdot)$ is a kernel (such as the Bartlett kernel), and $0 \leq K\left(\frac{j}{M}\right) \leq 1$, with $K(0) = 1$, and $M = o(P^{1/2})$. Additionally, h is the forecast horizon (set equal to unity in our experiments), P is as defined above, and $\widehat{u}_{1,t+1}$ and $\widehat{u}_{2,t+1}$ are the out–of sample forecast errors associated with least squares estimation of "smaller" and "bigger" linear models, respectively (see below for further details). Note that $\overline{j}$ does not grow with the sample size. Therefore, the denominator in CM is a consistent estimator of the long run variance only when $E\left(c_t c_{t+|k|}\right) = 0$ for all $|k| > h$ (see Assumption A3 in Clark and McCracken (2004)). Thus, the statistic takes into account the moving average structure of the prediction errors, but still does not allow for dynamic misspecification under the null. This is one of the main differences between the CM and CS (CCS) tests.

Note also that the DM test is the mean square error version of the Diebold and Mariano (1995) test for predictive accuracy, and is defined as follows:

[17] The CCS statistic is defined as $m_P = \frac{1}{P^{1/2}}\sum_{t=R}^{T-1} \widehat{u}_{1,t+1} Z^t$. Thus, this test can be seen as a special case of the CS test that is designed to be powerful against linear alternatives, and is not a explicity desinged to have power against generic nonlinear alternatives as is the CS test. In this sense, the CCS test is comparable to the CM test, which is also designed to have power against linear alternatives.

$$DM = \sqrt{P}\frac{\frac{1}{P}\sum_{t=R}^{T}\widehat{d}_{t+h}}{\frac{1}{P-h+1}\sum_{j=-\overline{j}}^{\overline{j}}\sum_{t=R+j}^{T-h}K\left(\frac{j}{M}\right)\left(\widehat{d}_{t+h}-\overline{d}\right)\left(\widehat{d}_{t+h-j}-\overline{d}\right)},$$

where $\widehat{d}_{t+h} = \widehat{u}^2_{1,t+h} - \widehat{u}^2_{2,t+h}$, and $\overline{d} = \frac{1}{P-h+1}\sum_{t=R}^{T-\tau}\widehat{d}_{t+h}$. The limiting distributions of the CM and DM statistics are given in Theorems 3.1 and 3.2 in Clark and McCracken (2004), and for $h > 1$ contain nuisance parameters so that critical values cannot be directly tabulated, and hence Clark and McCracken (2004) use the Kilian parametric bootstrap to obtain critical values. In this case, as discussed above, it is not clear that the parametric bootstrap is asymptotically valid. However, again as alluded to above, the parametric bootstrap approach taken by Clark and McCracken is clearly a good approximation, at least for the DGPs and horizon considered in our experiments, given that these tests have very good finite sample properties (see discussion of results below). Complete details of all tests are given in Table 1.

Data are generated according to the following DGPs:

$x_t = a_1 + a_2 x_{t-1} + u_{1,t},\ u_{1,t} \sim iidN(0,1)$

$w_t = a_1 + a_3 w_{t-1} + u_{2,t},\ u_{2,t} \sim iidN(0,1)$

Size1: $y_t = a_1 + a_2 y_{t-1} + a_4 w_{t-1} + u_{3,t},\ u_{3,t} \sim iidN(0,1)$

Size2: $y_t = a_1 + a_2 y_{t-1} + a_4 w_{t-1} + a_3 u_{3,t-1} + u_{3,t}$

Power1 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(\tan^{-1}(x_{t-1}/2)) + a_4 w_{t-1} + u_{3,t}$

Power2 : $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1} + a_4 w_{t-1} + u_{3,t}$

Power3 : $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1}1\{x_{t-1} > a_1/(1-a_2)\} + a_4 w_{t-1} + u_{3,t}$

Power4 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(\tan^{-1}(x_{t-1}/2)) + a_4 w_{t-1} + a_3 u_{3,t-1} + u_{3,t}$

Power5: $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1} + a_4 w_{t-1} + a_3 u_{3,t-1} + u_{3,t}$

Power6: $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1}1\{x_{t-1} > a_1/(1-a_2)\} + a_4 w_{t-1} + a_3 u_{3,t-1} + u_{3,t}.$

Power7 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(x_{t-1}) + a_4 w_{t-1} + u_{3,t}$

Power8 : $y_t = a_1 + a_2 y_{t-1} + 2x^2_{t-1} + a_4 w_{t-1} + u_{3,t}$

Power9 : $y_t = a_1 + a_2 y_{t-1} + 2|x_{t-1}| + a_4 w_{t-1} + u_{3,t}$

Power10 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(x_{t-1}) + a_4 w_{t-1} + a_3 u_{3,t-1} + u_{3,t}$

Power11: $y_t = a_1 + a_2 y_{t-1} + 2x^2_{t-1} + a_4 w_{t-1} + a_3 u_{3,t-1} + u_{3,t}$

Power12: $y_t = a_1 + a_2 y_{t-1} + 2|x_{t-1}| + a_4 w_{t-1} + a_3 u_{3,t-1} + u_{3,t}.$

The benchmark models (*Size1* and *Size2*) are AR(1) and ARMA(1,1) processes. The null hypothesis is that no competing model outperforms the benchmark model. The alternative models

all include (non)linear functions of x_{t-1}. In this sense, our focus is on (non)linear out-of-sample Granger causality testing. The parameter $a_4 = \{0, 1\}$, so that the variable w_{t-1} sometimes enters into the DGP. As this regressor is never included in any regression models, it is meant to render all estimated models misspecified if $a_4 = 1$. Even in cases where $a_4 = 0$, all regression models are misspecified, as all fitted regression functions are linear in their variables (so that there is neglected nonlinear (Granger) causality). The exception to this rule is the case where data are generated according to *Power2.*[18]

The functional forms that are specified under the alternative include: (i) exponential (*Power1,Power7*); (ii) linear (*Power2*); (iii) self exciting threshold (*Power3*), squared (*Power8*), and absolute value (*Power9*). In addition, *Power4-Power6* and *Power10-Power*12 are the same as the others, except that an MA(1) term is added. Notice that *Power1* includes a nonlinear term that is similar in form to the test function, $g(\cdot)$, which is defined below. Also, *Power2* serves as linear causality benchmarks. Finally, *Power7* is an explosive "strawman" model. Test statistics are constructed by fitting what is referred to in the next section as a "small model" (i.e. a linear AR(1) in y_t) in order to construct the CS and CCS test statistics. Additionally, a "big model" (which is a linear AR(1) model in y_t, with x_{t-1} added as an additional regressor) is also fitted in order to construct the F, CM, and DM test statistics. All test statistics are formed using one-step ahead predictions (and corresponding prediction errors) from recursively estimated models.

In all experiments, we set $g(z^{t-1}, \gamma) = \exp(\sum_{i=1}^{2}(\gamma_i \tan^{-1}((z_{i,t-1} - \overline{z}_i)/2\widehat{\sigma}_{z_i})))$, with $z_{1,t-1} = x_{t-1}$, $z_{2,t-1} = y_{t-1}$, and γ_1, γ_2 scalars. Additionally, define $\Gamma = [0.0, 5.0]$x$[0.0, 5.0]$. We consider a grid that is delineated by increments of size 0.5. We consider quadratic loss, so that when the DGPs are as in *Size1 and Size2,* the best 1-step ahead predictor is the conditional mean (i.e. $a_1 + a_2 y_t$). All results are based on 500 Monte Carlo replications, and samples of T=200, T=300, and T=600 are used. For the sake of brevity, however, results for T=200 and T =300 are not included and are available upon request, although there is little additional to see in these results, as power is lower for smaller sample sizes, and as all tests have empirical rejection frequencies that are fairly close to nominal test levels.[19] The following parameterizations are used: $a_1 = 1.0$, $a_2 = \{0.3, 0.6, 0.9\}$, and

[18]Note that *Power*5 is also linear in x_{t-1}. However, corresponding fitted linear regression models with y_t, y_{t-1}, and x_{t-1} are still misspecified, as there is an AR error component in the DGP that is not accounted for in the fitted models.

[19]Of course, and as expected, block lengths must be increased as sample sizes are increased in order to retain good finite sample empirical level when using the CS and CCS tests.

$a_3 = 0.3$. Additionally, bootstrap critical values are constructed using 100 simulated statistics, the block length, l, is set equal to $\{2, 5, 10\}$, and $P = (1/2)T$.

Findings are summarized in Tables 2-3 for the CS test, and Tables 4-5 for all other tests. Note that Tables 2 and 3 differ only with regard to the value of a_4, which is equal to 1 for the former and to 0 for the latter. The same distinction applies to Tables 4 and 5. The first column in the tables states the model type (e.g. *Size1),* and all numerical entries are test rejection frequencies. Although results are only reported for the case where $P = 0.5T$, additional results for $P = 0.4T$ and $0.6T$ were also tabulated. These results are qualitatively similar to those reported, and are available upon request from the authors. Overall, results are quite clear-cut, as is evidenced by inspection of the tables.

First, by inspection of Table 3, we note immediately that, as predicted by the theoretical results stated in the previous sections, the recursive block bootstrap gives rise to tests with much better finite sample properties than the bootstrap which does not take into account parameter estimation error and the standard block bootstrap. In particular, both incorrect bootstrap procedures give rise to tests which are severely undersized, and consequently have lower power. Thus, we can henceforth restrict our attention to the recursive block bootstrap. Note that the third column, corresponding to a block length of 10, displays rejection rates which are closest to nominal rates. On the other hand, it is clear that in the presence of high dependence, i.e. $a_2 = 0.9$, one should use a block length larger than 10.[20] We now turn to inspection of Table 5. First, we find that the F test is highly oversized, thus confirming the tendency of in-sample overfitting. Second, we note that the CM encompassing test performs very well, no matter which critical values are used. (Of course, for $h > 1$ the Kilian bootstrap should dominate the other approaches, as discussed in Clark and McCracken (2004)). Thus, though designed for comparing linear models, the CM test seems to have quite good power against nonlinear alternatives. Third, we see that critical values based upon the Kilian bootstrap perform admirably when used with the DM and the CM tests, but not when used with the CS or the CCS tests. When implementing the CS and CCS tests, the use of the recursive block bootstrap gives raise to tests with rejection rates much closer to the nominal ones. Overall, by jointly comparing Table 3 and 5, we note that for not too highly dependent observations, the CS test based on the recursive block bootstrap compares adequately

[20]This is confirmed by additional findings using block lengths of 15 and 20, which are available upon request.

with the CM and DM tests based on the Kilian bootstrap. Fourth, the CCS test performs very well in finite samples, suggesting that the CCS and CM tests, which were designed with exactly the same null and alternative hypotheses in mind, perform very similarly. As might be expected, the overall conclusions based upon inspection of Tables 2 and 4 are qualitatively the same as those based on Tables 3 and 5. Furthermore, again as should be expected, finite sample performance of all tests is generally better when $a_4 = 0$, as the degree of misspecification is lesser in this case.

Finally, it should be pointed out that, although the CS test tends to be undersized and to have lower power than the CM, DM, and CCS tests in presence of highly dependent observations, it is able to reject in certain contexts where a variety of the other tests examined here fail to reject, as discussed in the next section.

6 Empirical Illustration

In this section we implement the F, CM, DM, CS and CCS tests that are described in Table 1. In particular, we use these forecast encompassing, equal forecast accuracy, and in-sample Granger causality tests in order to assess the marginal predictive content of unemployment for core CPI inflation. Recent contributions to this important literature include the papers of Bachmeier and Swanson (2005), Clark and McCracken (2001), Staiger, Stock and Watson (1997), and Stock and Watson (1999). It should be stressed that the results presented in this section are meant primarily to illustrate the uses of the different tests, and to underscore potentially important differences between the tests. Issues of nonlinear model selection and structural breaks, for example, are addressed elsewhere.

The data which we use are monthly, and span the period 1955:1-2004:12. The CPI data series is the consumer price index for all urban consumers, all items, seasonally adjusted (Bureau of Labor Statistics series id CPIAUCSL). The unemployment series is the civilian unemployment rate, seasonally adjusted (Bureau of Labor Statistics series id UNRATE).

We construct tests statistics using forecasts formed via three sampling schemes (see Table 1). The schemes are denoted Rtype=1,2,3. In each of the schemes, models are re-estimated (using least squares) at each point in time, before each new prediction is constructed. The different sampling schemes are employed in order to construct 93 different subsamples of the data (31 for each sample scheme). All subsamples begin with an in-sample period of 1955:1-1964:12, so that in all cases,

the first sub-sample has R=120. In the first sampling scheme (Rtype=1), subsequent samples are formed by rolling ahead 1-year, while retaining the feature that $R = 120$, so that the second subsample has an initial estimation period of 1956:1-1965:12, etc. The second scheme (Rtype=2), simply increases R by 12 observations (1 year), in order to form new subsamples, so that the second subsample has an initial estimation period of 1955:1-1965:12, etc. In both of these schemes, 1-step ahead predictions are constructed throughout the rest of each sub-sample, ending in 2004:12, so that $P = 600 - 120 - 12(s-1)$, where $s = 1, ..., 31$, and where s denotes the subsample. Thus, the last subsample in both sampling schemes has $P = 120$. The third scheme (Rtype=3) is the same as the first scheme, except that P is fixed to be 120 observations, so that $P = R$ in all sub-samples. Thus, the last sub-sample from Rtype=1 is identical to the last sub-sample from Rtype=3. All other subsamples across the three schemes are different from one another (with the exception of the first sub-sample for Rtype=1 and Rtype=2, which are also equivalent). Given that we fix $h = 1$, and given that $\pi = 1$ under Rtype=3, we could in principle use the critical values tabulated in Clark and McCracken (2001) and McCracken (2004) for the CM and DM tests. However, for the sake of brevity, and because the Kilian bootstrap was found to yield critical values just as close to actual distributional percentiles as when the tabulated values where used (see Monte Carlo results), we only report findings based on the use of the Kilian bootstrap. Complete results analogous to those reported in the Monte Carlo section of this paper have been tabulated, and are available upon request.[21] In all cases, the dependent variable in regressions and the target variable in forecasts is the first log difference of CPI. Explanatory variables include lags of inflation and lags of unemployment.

Results are gathered in Tables 5-6 and Figure 1. In Tables 5 and 6 mean square forecast errors (MSFEs) are tabulated for the "small model" which only contains lags of inflation (with lags chosen using the SIC), and the "big model" which contains lags of inflation and lags of unemployment (Table 5) or lags of differenced unemployment (Table 6).[22] A number of results emerge upon inspection of the tables. First, note that the big model which uses lags of unemployment often

[21] It should be noted that we do not use real-time data in this empirical illustration, even though both variables considered are subject to periodic revision. Extension of our results to incorporate real-time data is left to future research.

[22] Unemployment appears in our models in both differenced and undifferenced form because there is no consensus on which transformation is appropriate, both from a predictive perspective and from the perspective of valid statistical inference.

yields higher MSFEs than when differenced unemployment is used. An important exception to this finding, however, concerns the last 11 subsamples for Rtype=2. In particular, note that when comparing the last 11 entries of the fourth columns of MSFEs in the two tables, the MSFEs are always lower when unemployment is used (as opposed to differenced unemployment). More importantly, this is one of the few cases where the big model yields consistently lower MSFEs than the small model (when differences of unemployment are used for these 11 subsamples, on the other hand, the small model yields lower MSFEs). Thus, it remains unclear whether unemployment should be differenced or not. Second, there appears to be instability in the series, as evidenced by the fact that MSFEs associated with Rtype=1 are always lower than analogous MSFEs associated with Rtype=2; it appears to pay to use smaller windows of data when estimating prediction models, at least in the context of the simple linear models considered here. Third, predictions of inflation have clearly gotten substantially more accurate over our sample period, as evidenced by the fact that MSFEs are much bigger for early subsamples using Rtype=3 than Rtypes = 1 and 2, and are much smaller for the later sub-samples. This result may be in part due to the smooth nature of recent data relative to more distant data, although it is difficult to say with certainty what is causing this feature of our models. Finally, and although it is quite apparent that point MSFE results are highly dependent upon sampling scheme, there is clearly very little evidence of predictive power of unemployment for inflation. This evidence is in loose agreement with the results reported by Clark and McCracken (2001), where it was found that in-sample evidence of Granger causality was much stronger than out-of-sample evidence. It should be stressed, however, that thus far we have only compared MSFEs, and hence have focused our attention upon the comparison of purely linear models. In order to assess the potential impact of generic nonlinearity, we need to either fit a variety of nonlinear models (which may be a large undertaking, given the plethora of available models), or we need to carry out tests such as the generically comprehensive nonlinear Granger causality CS test. We turn to this issue next.

Figure 1 contains a summary of test results based on Rtypes=1,2,3 for the F, CM, DM, CS, and CCS tests. For the CS and CCS tests, critical values are constructed using the recursive block bootstrap (called the "nonpar boot" in the figure), while for the DM and the CM critical values have been constructed using Kilian's parametric bootstrap (called "param boot" in the figure). Plots in the figure report the cumulative number of rejections of the null of equal predictive ability, so that based on the first plotted observation (corresponding to a predictive period starting point

of 1965 - see horizontal axis), the maximum number of possible rejections is 1 (reported along the vertical axis). By the time all of the 31 subsamples are exhausted and 31 individual statistics have been calculated, there are 31 possible rejections. Thus, the maximum number of rejections at the point in the graph corresponding to an out-of-sample period that begins in 1995 is 31. By noting at which out-of-sample beginning date a rejection occurs, one can see which subsample MSFEs from Tables 6 and 7 lead to rejection of the null hypothesis. For example, because CCS rejects for 1995 for all Rtypes when unemployment is used in the bigger model, the MSFEs in the last row of Table 6 are all associated with rejection of the null of equal predictive ability, at least according to the CCS test. Inspection of Figure 1 leads to a number of findings.

First, there are far more rejections using the CS and CCS test than using any of the other tests. This suggests that the CS and CCS tests are capturing some feature of the data not captured by the other tests. Also, since the CS test is designed to have generic power against nonlinear alternatives while all of the other tests are designed to have power against linear alternatives, we have evidence of nonlinear Granger causality. Namely, unemployment may have substantial marginal predictive power for inflation, should the right nonlinear model be specified.

Second, the in-sample F-test also sometimes rejects, in accord with the evidence in the literature (see e.g. Clark and McCracken (2001)). However, this result is suspect given our Monte Carlo finding that this test is oversized in finite samples.

Third, as there are nearly no rejections of the null hypothesis using the CM and DM tests, and there are many rejections using the CS test, we have some evidence of acceptance of our out-of-sample linear Granger noncausality null hypothesis. Instead, potential marginal predictive power will likely arise through nonlinear interactions between inflation and unemployment, as might be expected, given the functional form of the Phillips curve, for example.

Fourth, it is interesting to note that the CS and CCS tests tended to have lower power than the CM and DM tests, in our Monte Carlo experiments. Additionally, for block lengths similar to those considered in this empirical illustration, the CS test was shown to be conservative. In stark contrast to these experimental findings, our empirical results suggest that only the CS and CCS seem to be able to reject the null hypothesis. This is perhaps not overly surprising. For example, the CS test is essentially an out of sample Bierens's test, which is known to have relatively low power against specific alternatives, and is assumed to have some power against a large spectrum of alternatives. Thus, it appears that the CS (CCS) test is detecting some sort of "hidden" nonlinearities in the

predictive content of unemployment for inflation. This in turn suggests that the Monte Carlo experiments discussed above are not illustrative of the whole picture. Namely, it appears that there may be departures against which the CS and CCS tests have more power than the DM and CM tests.

In summary, this empirical illustration is meant only to shed light on the empirical application of a variety of different tests, including the F, CM, DM, CS, and CCS tests. Much empirical work is needed before a complete picture emerges concerning the prevalence of nonlinear Granger causality in the unemployment/inflation relationship. This is left to future research. It is clear, however, that much can be learned by using *all* of the different tests in consort with one another. The picture that emerges when only a subset of the tests is used to analyze the marginal predictive content of unemployment for inflation is that of an absence of predictive ability. When all of the tests are used, on the other hand, interesting evidence arises concerning the potential nonlinear predictive content of unemployment. Thus, the tests discussed in this illustration appear to be useful complements to each other.

7 Conclusions

In many instances, test statistics based on recursive and/or rolling estimation schemes have limiting distributions which are functionals of Gaussian processes, and which have covariance kernels that reflect parameter uncertainty. In these cases, limiting distributions are thus not nuisance parameter free, and valid critical values are often obtained via bootstrap methods. In this paper, we first developed a bootstrap procedure that properly captures the contribution of parameter estimation error in recursive estimation schemes using dependent data. Intuitively, when parameters are estimated recursively, as is done in our framework, earlier observations in the sample enter into test statistics more frequently than later observations. This induces a location bias in the bootstrap distribution, which can be either positive or negative across different samples, and hence the bootstrap modification that we discuss is required in order to obtain first order validity of the bootstrap. Within this framework, we then presented two applications, both based on forecast model selection. In particular, we considered the comparison of multiple (possibly misspecified) models in terms of out-of-sample predictive accuracy. Our applications extend the White (2000) reality check to the case of non vanishing parameter estimation error, and extend the integrated conditional moment (ICM) tests of Bierens (1982, 1990) and Bierens and Ploberger (1997) to the

case of out-of-sample prediction. Of note is that in both of these examples, it is shown that we must construct bootstrap statistics that are different from the "usual" bootstrap statistics, which are defined as the difference between the statistic computed over the sample observations and over the bootstrap observations. This feature of our applications suggests that one must be careful when forming bootstrap statistics in all cases for which recursive estimation is used and predictive model selection is the objective. Finally,results of a Monte Carlo investigation of a variety of related tests, and an empirical illustration were presented

8 Appendix

As the statements below hold for $i = 1, ..., n$, and given that the proofs are the same regardless which model is considered, for notational simplicity we drop the subscript i.

Proof of Theorem 1: Given (7), by first order conditions,

$$\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \widetilde{\theta}_t^*) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t)\right)\right) = 0.$$

Thus, a Taylor expansion around $\widehat{\theta}_t$ yields:

$$\begin{aligned}\left(\widetilde{\theta}_t^* - \widehat{\theta}_t\right) &= \left(\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta^2 q(y_j^*, Z^{*,j-1}, \overline{\theta}_t^*)\right)^{-1} \\ &\quad \times\left(\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t) - \left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t)\right)\right)\right),\end{aligned}$$

where $\overline{\theta}_t^* \in \left(\widetilde{\theta}_t^*, \widehat{\theta}_t\right)$. Hereafter, let $B^\dagger = \left(E\left(\nabla_\theta q(y_j, Z^{j-1}, \theta^\dagger)\right)\right)^{-1}$. Recalling that we resample from the entire sample, regardless the value of t, it follows that:

$$\frac{1}{t}\sum_{j=s}^{t}E^*\left(\nabla_\theta^2 q(y_j^*, Z^{*,j-1}, \theta)\right) = \frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta^2 q(y_k, Z^{k-1}, \theta) + O_{P^*}\left(\frac{l}{T}\right), \ \Pr - P, \tag{20}$$

where the $O_{P^*}\left(\frac{l}{T}\right)$ term is due to the end effect (i.e. due to the contribution of the first and last l observations, as shown in Lemma A1 in Fitzenberger (1997)). Thus,

$$\begin{aligned}&\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\left(\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta^2 q(y_j^*, Z^{*,j-1}, \theta)\right)^{-1} - B^\dagger\right| \\ \leq\ &\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\left(\frac{1}{t}\sum_{j=s}^{t}\nabla_\theta^2 q(y_j^*, Z^{*,j-1}, \theta)\right)^{-1} - \left(\frac{1}{t}\sum_{j=s}^{t}E^*\left(\nabla_\theta^2 q(y_j^*, Z^{*,j-1}, \theta)\right)\right)^{-1}\right| \\ &+\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\left(\frac{1}{t}\sum_{j=s}^{t}E^*\left(\nabla_\theta^2 q(y_j^*, Z^{*,j-1}, \theta)\right)\right)^{-1} - B^\dagger\right|.\end{aligned} \tag{21}$$

Given (20), and Assumptions A1-A2, the second term on the RHS of (21) is $o_P(1)$. Recalling also that the resampled series consists of b independent and identically distributed blocks, and that $b/T^{1/2} \to \infty$, it follows that the first term on on the RHS of (21) is $o_{P^*}(1)$ $\Pr - P$, given the uniform

law of large number for *iid* random variables. Thus,

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}_t^*-\widehat{\theta}_t\right)\\
=\;& B^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta q(y_j^*,Z^{*,j-1},\widehat{\theta}_t)-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k,Z^{k-1},\widehat{\theta}_t)\right)\right)\right)\\
&+o_{P^*}(1)\ \Pr-P, \qquad (22)
\end{aligned}$$

and a first order expansion of the RHS of (22) around $\theta^\dagger$ yields:

$$\begin{aligned}
&\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}_t^*-\widehat{\theta}_t\right)\\
=\;& B^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta q(y_j^*,Z^{*,j-1},\theta^\dagger)-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta q(y_k,Z^{k-1},\theta^\dagger)\right)\right)\right)\\
&+B^\dagger\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta^2 q(y_j^*,Z^{*,j-1},\overline{\theta}_t)-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta^2 q(y_k,Z^{k-1},\overline{\theta}_t)\right)\right)\right)\right.\\
&\left.\times\left(\widehat{\theta}_t-\theta^\dagger\right)\right)+o_{P^*}(1)\ \Pr-P. \qquad (23)
\end{aligned}$$

We need to show that the second term on the RHS of (23) is $o_{P^*}(1)\ \Pr-P$. Note that this term is majorized by

$$B^\dagger\sup_{t\geq R}\sup_{\theta\in\Theta}\frac{\sqrt{P}}{t^{1+\vartheta}}\left|\sum_{j=s}^{t}\left(\nabla_\theta^2 q(y_j^*,Z^{*,j-1},\theta)-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta^2 q(y_k,Z^{k-1},\theta)\right)\right)\right|\sup_{t\geq R}t^\vartheta\left|\widehat{\theta}_t-\theta^\dagger\right|,$$

with $1/3<\vartheta<1/2$. Recalling also that $bl=T$ and $l=o(T^{1/4})$, it follows that $b/T^{3/4}\to\infty$. Thus, by the same argument used in Lemma 1(i) in Altissimo and Corradi (2002), and given (20), it follows that:

$$\sup_{t\geq R}\sup_{\theta\in\Theta}\left|\frac{1}{t}\sum_{j=s}^{t}\left(\nabla_\theta^2 q(y_j^*,Z^{*,j-1},\theta)-\left(\frac{1}{T}\sum_{k=s}^{T-1}\nabla_\theta^2 q(y_k,Z^{k-1},\theta)\right)\right)\right|=O_{a.s.^*}\left(\sqrt{\frac{\log\log b}{b}}\right),\ a.s.-P.$$

Thus,

$$\sup_{t\geq R}\sup_{\theta\in\Theta}\frac{\sqrt{P}}{t^{1+\vartheta}}\left|\sum_{j=s}^{t}\left(\nabla_\theta^2 q(y_j^*,Z^{*,j-1},\theta)-\left(\frac{1}{T}\sum_{j=s}^{T}\nabla_\theta^2 q(y_j,Z^{j-1},\theta)\right)\right)\right|=o_{P^*}(1),\ \Pr-P,$$

for $\vartheta > 1/3$. Finally, for all $\vartheta < 1/2$, $\sup_{t \geq R} t^{\vartheta} \left| \widehat{\theta}_t - \theta^{\dagger} \right| = o_P(1)$ by Lemma A3 in West (1996). Recalling that

$$\frac{1}{t} \sum_{j=s}^{t} E^* \left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) \right) = \frac{1}{T} \sum_{k=s}^{T-1} \nabla_\theta q(y_k, Z^{k-1}, \theta^{\dagger}) + O_P\left(\frac{l}{T} \right),$$

the right hand side of (23) can be written as:

$$\begin{aligned}
& \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}_t^* - \widehat{\theta}_t \right) \\
= \; & B^{\dagger} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\frac{1}{t} \sum_{j=s}^{t} \left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) - E^* \left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) \right) \right) \right) + o_{P^*}(1) \;\; \Pr - P \\
= \; & B^{\dagger} \frac{a_{R,0}}{\sqrt{P}} \sum_{j=1}^{R} \left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) - E^* \left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) \right) \right) \\
& + B^{\dagger} \frac{1}{\sqrt{P}} \sum_{j=1}^{P-1} a_{R,j} \left(\nabla_\theta q(y_{R+j}^*, Z^{*,R+j-1}, \theta^{\dagger}) - E^* \left(\nabla_\theta q(y_{R+j}^*, Z^{*,R+j-1}, \theta^{\dagger}) \right) \right) \\
& + o_{P^*}(1) \;\; \Pr - P, \qquad (24)
\end{aligned}$$

where $a_{R,j} = a_{R,i} = (R+i)^{-1} + ... + (R+P-1)^{-1}$, for $0 \leq i < P-1$. The second equality on the RHS of (24) follows directly from Lemma A5 in West (1996).

Now, $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}_t^* - \widehat{\theta}_t \right)$ satisfies a central limit theorem for triangular independent arrays (see e.g. White and Wooldridge (1988)), and thus, conditional on the sample, it converges in distribution to a zero mean normal random variable.

Furthermore, by Theorem 4.1 in West (1996):

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widehat{\theta}_t - \theta^{\dagger} \right) \xrightarrow{d} N\left(0, 2\Pi B^{\dagger} C_{00} B^{\dagger} \right),$$

where $C_{00} = \sum_{j=-\infty}^{\infty} E\left(\left(\nabla_\theta q(y_{1+s}, Z^s, \theta^{\dagger}) \right) \left(\nabla_\theta q(y_{1+s+j}, Z^{s+j}, \theta^{\dagger}) \right)' \right)$ and $\Pi = 1 - \pi^{-1} \ln(1+\pi)$. Therefore, the statement in the theorem will follow once we have shown that:

$$Var^* \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widetilde{\theta}_t^* - \widehat{\theta}_t \right) \right) = 2\Pi B^{\dagger} C_{00} B^{\dagger}, \;\; \Pr - P. \qquad (25)$$

For notational simplicity, let $\nabla_\theta q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) = h_j^*$, and let $\nabla_\theta q(y_j, Z^{j-1}, \theta^{\dagger}) = h_j$. Additionally, let $\overline{h}_T = \frac{1}{T} \sum_{t=s}^{T} h_t$. Then, given (24):

$$Var^* \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}_t^* - \widehat{\theta}_t \right) \right) = \frac{R}{P} Var^* \left(a_{R,0} \frac{1}{\sqrt{R}} \sum_{j=1}^{R} h_j^* \right)$$

$$+\frac{1}{P}Var^*\left(\sum_{j=1}^{P-1} a_{R,j}h^*_{R+j}\right) + \frac{1}{P}Cov^*\left(a_{R,0}\sum_{j=1}^{R} h^*_j, \sum_{j=1}^{P-1} a_{R,j}h^*_{R+j}\right).$$

As all blocks are independent, conditional on the sample, the covariance term in this expression is equal to zero. Without loss of generality, set $R = b_1 l$ and $P = b_2 l$, where $b_1 + b_2 = b$. It then follows that, up to a term of order $O(l/R^{1/2})$,

$$\begin{aligned} Var^*\left(a_{R,0}\frac{1}{\sqrt{R}}\sum_{j=1}^{R} h^*_j\right) &= a^2_{R,0}Var^*\left(\frac{1}{\sqrt{R}}\sum_{k=1}^{b_1}\sum_{i=1}^{l} h_{I_k+i}\right) \\ &= a^2_{R,0}E^*\left(\frac{1}{R}\sum_{k=1}^{b_1}\sum_{i=1}^{l}\sum_{k=1}^{l}(h_{I_k+i}-\overline{h}_T)(h_{I_k+j}-\overline{h}_T)'\right) \end{aligned}$$

$$= a^2_{R,0}\left(\frac{1}{R}\sum_{t=l}^{R-l}\sum_{j=-l}^{l}(h_t-\overline{h}_T)(h_{t+j}-\overline{h}_T)'\right) + O(l/R^{1/2})\ \Pr-P.$$

Thus,

$$\begin{aligned} &\frac{R}{P}Var^*\left(a_{R,0}\frac{1}{\sqrt{R}}\sum_{j=1}^{R} h^*_j\right) \\ = \ &\frac{Ra^2_{R,0}}{P}\sum_{j=-l}^{l}\gamma_j + \frac{Ra^2_{R,0}}{P}\left(\frac{1}{R}\sum_{t=l}^{R-l}\sum_{j=-l}^{l}((h_t-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_j)\right) + O\left(\frac{l^2}{R}\right), \end{aligned} \tag{26}$$

where $\gamma_j = Cov(h_1, h_{1+j})$. By West (1996, proof of Lemma A5), it follows that $\frac{Ra^2_{R,0}}{P}\sum_{j=-l}^{l}\gamma_j \to \pi^{-1}\ln^2(1+\pi)C_{00}$, while the second term on the RHS above goes to zero, $\Pr-P$ (see e.g. Theorem 2 in Newey and West (1987)). Now, up to a term of order $O(l/P^{1/2})$ $\Pr-P$:

$$Var^*\left(\frac{1}{\sqrt{P}}\sum_{j=1}^{P-1} a_{R,j}h^*_{R+j}\right) = Var^*\left(\frac{1}{\sqrt{P}}\sum_{k=b_1+1}^{b}\sum_{i=1}^{l} a_{R,((k-1)l+i)}h_{I_k+i}\right)$$

$$\begin{aligned} &= \frac{1}{P}E^*\left(\sum_{k=b_1+1}^{b}\sum_{i=1}^{l}\sum_{j=1}^{l} a_{R,((k-b_1-1)l+i)}a_{R,((k-b_1-1)l+j)}(h_{I_k+i}-\overline{h}_T)(h_{I_k+j}-\overline{h}_T)'\right) \\ &= \frac{1}{P}\sum_{k=b_1+1}^{b}\sum_{i=1}^{l}\sum_{j=1}^{l} a_{R,((k-b_1-1)l+i)}a_{R,((k-b_1-1)l+j)}E^*\left((h_{I_k+i}-\overline{h}_T)(h_{I_k+j}-\overline{h}_T)'\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{P}\sum_{k=b_1+1}^{b}\sum_{i=1}^{l}\sum_{j=1}^{l} a_{R,((k-b_1-1)l+i)}a_{R,((k-b_1-1)l+j)}\left(\frac{1}{T}\sum_{t=l}^{T-l}(h_{t+i}-\overline{h}_P)(h_{+t+j}-\overline{h}_P)'\right)+O(l/P^{1/2})\ \Pr-P\\
&= \frac{1}{P}\sum_{k=b_1+1}^{b}\sum_{i=1}^{l}\sum_{j=1}^{l} a_{R,((k-b_1-1)l+i)}a_{R,((k-b_1-1)l+j)}\gamma_{i-j}\\
&\quad+\frac{1}{P}\sum_{k=b_1+1}^{b}\sum_{i=1}^{l}\sum_{j=1}^{l} a_{R,((k-b_1-1)l+i)}a_{R,((k-b_1-1)l+j)}\left(\frac{1}{T}\sum_{t=l}^{T-l}\left((h_{t+i}-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_{i-j}\right)\right)\\
&\quad+O(l/P^{1/2})\ \Pr-P
\end{aligned}
\tag{27}
$$

We need to show that the last term on the last equality in (27) is $o(1)\ \Pr-P$. First note that this term is majorized by

$$
\begin{aligned}
&\left|\frac{b_2}{P}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\frac{1}{T}\sum_{t=l}^{T-l}\left((h_{t+i}-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_{i-j}\right)\right)\right|\\
&= \left|\frac{1}{T}\sum_{t=l}^{T-l}\sum_{j=-l}^{l}\left((h_t-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_j\right)\right|+O(l/P^{1/2})\ \Pr-P.
\end{aligned}
\tag{28}
$$

The first term on the RHS of (28) goes to zero in probability, by the same argument as that used in Lemma 2 in Corradi (1999).[23] With regard to the first term on the RHS of the last equality in (27), note that:

$$
\begin{aligned}
\frac{1}{P}\sum_{k=1}^{b_2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_{R,((k-1)l+i)}a_{R,((k-1)l+j)}\gamma_{i-j} &= \frac{1}{P}\sum_{t=l}^{P-l}\sum_{j=-l}^{l} a_{R,t}a_{R,t+j}\gamma_j+O(l/P^{1/2})\ \Pr-P\\
&= \frac{1}{P}\sum_{t=l}^{P-l}a_{R,t}^2\sum_{j=-l}^{l}\gamma_j+\frac{1}{P}\sum_{t=l}^{P-l}\sum_{j=-l}^{l}(a_{R,t}a_{R,t+j}-a_{R,t}^2)\gamma_j+O(l/P^{1/2})\ \Pr-P.
\end{aligned}
$$

By the same argument as that used in Lemma A5 of West (1996), the second term on the RHS above approaches zero, while:

$$
\frac{1}{T}\sum_{t=l}^{P-l}a_{R,t}^2\sum_{j=-l}^{l}\gamma_j\rightarrow\left(2[1-\pi^{-1}\ln(1+\pi)]-\pi^{-1}\ln^2(1+\pi)\right)C_{00}.
$$

As the first term on the RHS of (26) converges to $\pi^{-1}\ln^2(1+\pi)C_{00}$ (see West (1996), p.1082), the desired outcome then follows.□

[23]The domination conditions here are weaker than those in Lemma 2 in Corradi (1999), as we require only convergence to zero in probability, and not almot sure convergence.

Proof of Proposition 2: Let $\overline{u}_{i,t} = y_t - \kappa(Z^{t-1}, \overline{\theta}_{i,t})$, with $\overline{\theta}_{i,t} \in \left(\widehat{\theta}_{i,t}, \theta^{\dagger}\right)$. Via a mean value expansion, and given Assumptions A1-A2:

$$\begin{aligned}
S_P(1,k) &= \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1})\right) \\
&= \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g(u_{1,t+1}) - g(u_{k,t+1})\right) \\
&\quad + \frac{1}{P} \sum_{t=R}^{T-1} g'(\overline{u}_{1,t+1}) \nabla_{\theta_1} \kappa_1(Z^t, \overline{\theta}_{1,t}) P^{1/2} \left(\widehat{\theta}_{1,t} - \theta_1^{\dagger}\right) \\
&\quad - \frac{1}{P} \sum_{t=R}^{T-1} g'(\overline{u}_{k,t+1}) \nabla_{\theta_k} \kappa_k(Z^t, \overline{\theta}_{k,t}) P^{1/2} \left(\widehat{\theta}_{k,t} - \theta_k^{\dagger}\right) \\
&= \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g(u_{1,t+1}) - g(u_{k,t+1})\right) \\
&\quad + \mu_1 \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{1,t} - \theta_1^{\dagger}\right) - \mu_k \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{k,t} - \theta_k^{\dagger}\right) + o_P(1),
\end{aligned}$$

where $\mu_1 = E\left(g'(u_{1,t+1}) \nabla_{\theta_1} \kappa_1(Z^t, \theta_1^{\dagger})\right)$, and μ_k is defined analogously. Now, when all competitors have the same predictive accuracy as the benchmark model, by the same argument as that used in Theorem 4.1 of West (1996), it follows that:

$$\left(S_P^{\mu}(1,2), ..., S_P^{\mu}(1,n)\right) \xrightarrow{d} N(0, V),$$

where $S_P^{\mu}(1,k) = S_P(1,k) - \sqrt{P} E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)$, and where V is an $n \times n$ matrix with i,j element $v_{i,j}$ defined in the statement of the proposition. The distribution of S_P then follows as a straightforward application of the continuous mapping theorem.□

Proof of Proposition 3: Let $\widehat{u}^*_{i,t+1} = y^*_{t+1} - \kappa_i(Z^{*,t}, \widehat{\theta}_{i,t})$, $\overline{u}^*_{i,t+1} = y^*_{t+1} - \kappa_i(Z^{*,t}, \overline{\theta}^*_{i,t})$, with $\overline{\theta}^*_{i,t} \in \left(\widetilde{\theta}^*_{i,t}, \widehat{\theta}_{i,t}\right)$, Additionally, let $\widehat{u}^{(t)}_{i,j+1} = y_{j+1} - \kappa_i(Z^j, \widehat{\theta}_{i,t})$. It follows that:

$$\begin{aligned}
S^*_P(1,k) &= \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(g(\widetilde{u}^*_{1,t+1}) - g(\widetilde{u}^*_{k,t+1})\right) - \frac{1}{T} \sum_{j=s}^{T} \left(g(\widehat{u}^{(t)}_{1,j+1}) - g(\widehat{u}^{(t)}_{k,j+1})\right) \right) \\
&= \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(g(\widehat{u}^*_{1,t+1}) - g(\widehat{u}^*_{k,t+1})\right) - \frac{1}{T} \sum_{j=s}^{T} \left(g(\widehat{u}^{(t)}_{1,j+1}) - g(\widehat{u}^{(t)}_{k,j+1})\right) \right) \\
&\quad + \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(\nabla_{\theta_1} g(\overline{u}^*_{1,t+1}) \left(\widehat{\theta}^*_{1,t} - \widehat{\theta}_{1,t}\right) - \nabla_{\theta_1} g(\overline{u}^*_{k,t+1}) \left(\widehat{\theta}^*_{k,t} - \widehat{\theta}_{k,t}\right) \right). \qquad (29)
\end{aligned}$$

Now,

$$E^*\left(g(\widehat{u}^*_{1,t+1}) - g(\widehat{u}^*_{k,t+1})\right) = \frac{1}{T}\sum_{j=s}^{T}\left(g(\widehat{u}^{(t)}_{1,j+1}) - g(\widehat{u}^{(t)}_{k,j+1})\right) + O\left(\frac{l}{T}\right).$$

Thus, by Theorem 3.5 in in Künsch (1989), the first term on the second equality on the RHS of (29) converges in P^*-distribution to a zero mean normal random variable with variance equal to $\lim_{P\to\infty} Var^*\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(g(\widehat{u}^*_{1,t+1}) - g(\widehat{u}^*_{k,t+1})\right)\right)$, conditional on the sample and for all samples except a subset with probability measure approaching zero. Now, by the same argument used in the proof of Theorem 1:

$$Var^*\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(g(\widehat{u}^*_{1,t+1}) - g(\widehat{u}^*_{k,t+1})\right)\right) = Var\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1})\right)\right) + o(1)\ \Pr - P.$$

This implies that the first term in the second equality on the RHS of (29) has the same limiting distribution as $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) - E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, conditional on the sample, and for all samples except a subset with probability measure approaching zero. Finally, the last term in (29) has the same limiting distribution as $\mu_1\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\widehat{\theta}^*_{1,t} - \widehat{\theta}_{1,t}\right) - \mu_k\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\widehat{\theta}^*_{k,t} - \widehat{\theta}_{k,t}\right)$, conditional on the sample, and for all samples except a subset with probability measure approaching zero. The statement in the proposition then follows as a straightforward application of the continuous mapping theorem.□

Proof of Proposition 4: The proof follows directly from Theorem 1 in Corradi and Swanson (2002).□

Proof of Proposition 5: Recall that $g = q_1$. Additionally, let $\widetilde{u}^*_{1,t+1} = y^*_{t+1} - \begin{pmatrix} 1 & y^*_t \end{pmatrix}\widetilde{\theta}^*_{1,t}$, $\widehat{u}^*_{1,t+1} = y^*_{t+1} - \begin{pmatrix} 1 & y^*_t \end{pmatrix}\widehat{\theta}_{1,t}$, $\overline{u}^*_{1,t+1} = y^*_{t+1} - \begin{pmatrix} 1 & y^*_t \end{pmatrix}\overline{\theta}^*_{1,t}$, and $\widehat{u}^{(t)}_{1,j+1} = y_{j+1} - \begin{pmatrix} 1 & y_t \end{pmatrix}\widehat{\theta}_{1,t}$, where $\overline{\theta}^*_{1,t} \in \left(\widetilde{\theta}^*_{1,t}, \widehat{\theta}_{1,t}\right)$. It then follows that:

$$\begin{aligned}
&\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(g'(\widetilde{u}^*_{1,t+1})w(Z^{*,t},\gamma) - \frac{1}{T}\sum_{j=2}^{T-1}g'(\widehat{u}^{(t)}_{1,j+1})w(Z^j,\gamma)\right) \\
= &\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(g'(\widehat{u}^*_{1,t+1})w(Z^{*,t},\gamma) - \frac{1}{T}\sum_{j=2}^{T-1}g'(\widehat{u}^{(t)}_{1,j+1})w(Z^j,\gamma)\right) \\
&+\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\nabla_\theta g'(\overline{u}^*_{1,t+1})w(Z^{*,t},\gamma)\right)\left(\widetilde{\theta}^*_{1,t} - \widehat{\theta}_{1,t}\right). \qquad (30)
\end{aligned}$$

First, note that the first term on the RHS of the last equality in (30) has the same limiting distribution as

$\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(g'(u_{1,t+1})w(Z^t,\gamma)-E\left(g'(u_{1,t+1})w(Z^t,\gamma)\right)\right)$, pointwise in γ. Further, note that stochastic equicontinuity on Γ can be shown using the same approach as that used in the proof of Theorem 2 in Corradi and Swanson (2002). Therefore, under H_0, any continuous functional over Γ of $\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(g'(u^*_{1,t+1})w(Z^{*,t},\gamma)-\frac{1}{T}\sum_{j=2}^{T-1}g'(\widehat{u}^{(t)}_{1,j+1})w(Z^j,\gamma)\right)$ has the same limiting distribution of the same functional of $\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(g'(u_{1,t+1})w(Z^t,\gamma)-E\left(g'(u_{1,t+1})w(Z^t,\gamma)\right)\right)$. Finally, note that $\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}\left(\nabla_\theta g'(\overline{u}^*_{1,t+1})w(Z^{*,t},\gamma)\right)\left(\widetilde{\theta}^*_{1,t}-\widehat{\theta}_{1,t}\right)$ properly captures the contribution of recursive parameter estimation error to the covariance kernel.□

9 References

Altissimo, F. and V. Corradi, (2002), Bounds for Inference with Nuisance Parameters Present only under the Alternative", *Econometrics Journal*, 5, 494-518.

Andrews, D.W.K., (2002), Higher-Order Improvements of a Computationally Attractive $k-$step Bootstrap for Extremum Estimators, *Econometrica,* 70, 119-162.

Andrews, D.W.K., (2004), The Block-Block Bootstrap: Improved Asymptotic Refinements, *Econometrica,* 72, 673-700.

Bachmeier, L. and N.R. Swanson, (2005), Predicting Inflation: Does The Quantity Theory Help?, *Economic Inquiry,* forthcoming.

Bierens, H.B., (1982): Consistent model specification tests, *Journal of Econometrics*, 20, 105-134.

Bierens, H.B., (1990): A Conditional Moment Test of Functional Form, *Econometrica* , 58, 1443-1458.

Bierens, H.J. and W. Ploberger, (1997): Asymptotic theory of integrated conditional moment tests, *Econometrica*, 65, 1129-1152.

Carlstein, E. (1986), The Use of Subseries Methods for Estimating the Variance of a General Statistic from a Stationary Time Series, *Annals of Statistics*, 14, 1171-1179.

Chao, J.C., V. Corradi, and N.R. Swanson (2001), An Out of Sample Test for Granger Causality", *Macroeconomic Dynamics,* v.5, 598-620.

Christoffersen, P. and F.X. Diebold, (1996), Further Results on Forecasting and Model Selection under Asymmetric Loss, *Journal of Applied Econometrics, 11, 561-572.*

Christoffersen, P. and F.X. Diebold, (1997), Optimal Prediction Under Asymmetric Loss, *Econometric Theory, 13, 808-817.*

Clark, T.E., and M.W., McCracken, (2001), Tests of Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Econometrics* 105, 85-110.

Clark, T.E., and M.W., McCracken (2004), Evaluating Long-Horizon Forecasts, Working Paper, University of Missouri-Columbia

Corradi, V., (1999), Deciding between $I(0)$ and $I(1)$ via FLIL-Based Bounds, *Econometric Theory*, 15, 643-663.

Corradi, V. and N.R. Swanson, (2002), A Consistent Test for Out of Sample Nonlinear Predictive Ability, *Journal of Econometrics,* 110, 353-381.

Corradi, V. and N.R. Swanson, (2004a), "Predictive Density Evaluation", in: *Handbook of Economic Forecasting*, eds. Clive W.J. Granger, Graham Elliot and Allan Timmerman, Elsevier, Amsterdam, forthcoming.

Corradi, V. and N.R. Swanson, (2004b), Some Recent Developments in Predictive Accuracy Testing with Nested Models and (Generic) Nonlinear Alternatives", *International Journal of Forecasting*, 20, 185-199.

Corradi, V. and N.R. Swanson, (2005), "Predictive Density and Confidence Intervals Accuracy Tests", *Journal of Econometrics,* forthcoming.

Diebold, F.X., and R.S. Mariano, (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

Diebold, F.X. and C. Chen, (1996), Testing Structural Stability with Endogenous Breakpoint: A Size Comparison of Analytic and Bootstrap Procedures, *Journal of Econometrics*, 70, 221-241.

Elliott, G. and A. Timmerman, (2004a), Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions, *Journal of Econometrics,* forthcoming.

Elliott, G. and A. Timmerman, (2004b), Optimal Forecast Combination Under Regime Switching, *International Economic Review,* forthcoming.

Fitzenberger, B. (1997), The Moving Block Bootstrap and Robust Inference for Linear Least Square and Quantile Regressions, *Journal of Econometrics*, 82, 235-287.

Gallant, A.R. and H. White, (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models,* Blackwell, Oxford.

Giacomini, R., and H. White (2003), Conditional Tests for Predictive Ability, manuscript, University of California, San Diego.

Goncalves, S., and H. White, (2004), Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models, *Journal of Econometrics,* 119, 199-219.

Granger, C.W.J., (1993), On the Limitations of Comparing Mean Squared Forecast Errors: Comment, *Journal of Forecasting,* 12, 651-652.

Granger, C.W.J., (1999), Outline of Forecast Theory Using Generalized Cost Functions, *Spanish Economic Review,* 1, 161-173.

Hall, P., and J.L. Horowitz, (1996), Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators, *Econometrica,* 64, 891-916.

Hansen, B.E., (1996), Inference When a Nuisance Parameter is Not Identified Under the Null Hypothesis, *Econometrica,* 64, 413-430.

Harvey, D.I., S.J. Leybourne and P. Newbold, (1997), Tests for Forecast Encompassing, *Journal of Business and Economic Statistics*, 16, 254-259.

Inoue, A., (2001), Testing for Distributional Change in Time Series, *Econometric Theory*, 17, 156-187.

Inoue, A., and M. Shintani, (2004), Bootstrapping GMM Estimators for Time Series, *Journal of Econometrics,* forthcoming.

Inoue, A., and B. Rossi, (2004), Recursive Predictive Ability Tests for Real Time Data, Working Paper, Duke University and NC State.

Inoue, A. and L. Kilian, (2004), In-Sample and Out-of-Sample Tests of Predictability: Which One Should We Use?, *Econometric Reviews,* 23, 371-402.

Inoue, A. and L. Kilian, (2005), On the Selection of Forecasting Models, *Journal of Econometrics,* forthcoming.

Kilian, L., (1999), Finite Sample Properties of Percentile and Percentile t-Bootstrap Confidence Intervals for Impulse Responses, *Review of Economics and Statistics*, 81, 652-660.

Künsch H.R., (1989), The Jackknife and the Bootstrap for General Stationary Observations, *Annals of Statistics,* 17, 1217-1241.

Lahiri, S.N., (1999), Theoretical Comparisons of Block Bootstrap Methods, *Annals of Statistics,* 27, 386-404.

Linton, O., E. Maasoumi and Y.J. Whang, (2004), Consistent Testing for Stochastic Dominance Under General Sampling Schemes, *Review of Economic Studies,* forthcoming.

McCracken, M.W., (2004), Asymptotics for Out of Sample Tests of Causality, Working Paper, University of Missouri-Columbia.

McCracken. M.W., and S. Sapp, (2004), Evaluating the Predictive Ability of Exchange Rates Using Long Horizon Regressions: Mind your p's and q's. *Journal of Money, Credit and Banking*, forthcoming.

Newey, W.K. and K.D. West, (1987), A Simple Positive-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica,* 55, 703-708.

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

Politis, D.N., J.P. Romano and M. Wolf, (1999), *Subsampling,* Springer and Verlag, New York.

Schorfheide, F., (2004), VAR Forecasting under Misspecification, *Journal of Econometrics,* forthcoming.

Stinchcombe, M.B. and H. White, (1998), Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative, *Econometric Theory,* 14, 3, 295-325.

Swanson, N.R. and H. White, (1997), A Model Selection Approach to Real-Time Macroeconomic Forecasting using Linear Models and Artificial Neural Networks, *Review of Economics and Statistics*, 79, 540-550.

Weiss, A., (1996) Estimating Time Series Models Using the Relevant Cost Function, *Journal of Applied Econometrics,* 11, 539-560.

West, K., (1996), Asymptotic Inference About Predictive Ability, *Econometrica*, 64, 1067-1084.

White, H., (2000), A Reality Check for Data Snooping, *Econometrica*, 68, 1097-1126.

Wooldridge, J.M. and H. White, (1988), Some Invariance Principles and Central Limit Theorems for Dependent and Heterogeneous Processes, *Econometric Theory*, 4, 210-230.

Zellner, A., (1986), Bayesian Estimation and Prediction Using Asymmetric Loss Function, *Journal of the American Statistical Association,* 81, 446-451.

Table 1: Test Statistic and Sampling Scheme Mnemonics

Panel A: Test Statistic Mnemonics

F – The standard Wald version of the in-sample F-test is calculated using the entire sample of T observations. In particular, we use: $F = T\left(\sum_{t=1}^{T} \widehat{u}_{1,t}^2 - \sum_{t=1}^{T} \widehat{u}_{2,t}^2)/\sum_{t=1}^{T} \widehat{u}_{2,t}^2\right)$, where $\widehat{u}_{1,t}$ and $\widehat{u}_{2,t}$ are the in-sample residuals associated with least squares estimation of the smaller and bigger models, respectively, and where T denotes the sample size.

CM – The Clark and McCracken (2004) test is an out-of-sample encompassing test (see also Harvey, Leybourne and Newbold (1997), and is defined as follows: $CM = (P-h+1)^{1/2} \frac{\frac{1}{P-h+1}\sum_{t=R}^{T-h} \widehat{c}_{t+h}}{\frac{1}{P-h+1}\sum_{j=-\overline{j}}^{\overline{j}} \sum_{t=R+j}^{T-h} K\left(\frac{j}{M}\right)(\widehat{c}_{t+h}-\overline{c})(\widehat{c}_{t+h-j}-\overline{c})}$, where $\widehat{c}_{t+h} = \widehat{u}_{1,t+h}\left(\widehat{u}_{1,t+h} - \widehat{u}_{2,t+h}\right)$, $\overline{c} = \frac{1}{P-h+1}\sum_{t=R}^{T-\tau} \widehat{c}_{t+h}$, $K(\cdot)$ is a kernel (such as the Bartlett kernel), and $0 \leq K\left(\frac{j}{M}\right) \leq 1$, with $K(0) = 1$, and $M = o(P^{1/2})$. Additionally, h is the forecast horizon, P is the out-of-sample prediction period, and $\widehat{u}_{1,t+1}$ and $\widehat{u}_{2,t+1}$ are the out-of-sample residuals associated with least squares estimation of the smaller and bigger models, respectively. Note finally, that $\overline{j}$ does not grow with the sample size.

DM – The mean square error version of the Diebold and Mariano (1995) test is a predictive accuracy test, and is defined as follows: $DM = \sqrt{P}\frac{\frac{1}{P}\sum_{t=R}^{T} \widehat{d}_{t+h}}{\frac{1}{P-h+1}\sum_{j=-\overline{j}}^{\overline{j}} \sum_{t=R+j}^{T-h} K\left(\frac{j}{M}\right)(\widehat{d}_{t+h}-\overline{d})(\widehat{d}_{t+h-j}-\overline{d})}$, where $\widehat{d}_{t+h} = \widehat{\epsilon}_{t+h}^2 - \widehat{u}_{t+h}^2$, and $\overline{d} = \frac{1}{P-h+1}\sum_{t=R}^{T-\tau} \widehat{d}_{t+h}$.

CS – The so-called Corradi and Swanson (2002) test is a generically comprehensive out-of-sample encompassing test, and is defined as follows: $M_P = \int_\Gamma |m_P(\gamma)|\phi(\gamma)d\gamma$, where $m_P(\gamma) = \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1} g'(\widehat{u}_{1,t+1})w(Z^t,\gamma)$, and where $\int_\Gamma \phi(\gamma)d\gamma = 1$, $\phi(\gamma) \geq 0$, with $\phi(\gamma)$ absolutely continuous with respect to Lebesgue measure, Γ is a compact subset of $\Re^d$, for some finite d, and g' is the derivative of the loss function used for predictive evaluation, with respect to its argument. Additionally, $w(Z^t,\gamma)$ is a generically comprehensive function as discussed above and in Corradi and Swanson (2002), and Z^t is a vector of variables of interest.

CCS – The so-called Chao, Corradi and Swanson (2001) test is a simplified version of CS which is not designed to have power against generic nonlinear alternatives, and is defined as follows: $CCS = \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1} \widehat{u}_{1,t+1}Z^t$.

Panel B: Sampling Schemes

Rtype = 1 – In this sampling scheme, the initial in-sample estimation period is fixed to be 10 years in length, so that R=120 in all cases. Additionally, P is set to be the remainder of the sample, up until 2004:12. For the first of the 31 calculated statistics, the initial in-sample period is thus 1955:1-1964:12, and the out-of-sample period is 1965:1-2004:12. Now, the sample used for the next statistic rolls the initial estimation period forward 1 year. Thus, the in-sample period is 1956:1-1965:12, and the out-of-sample period is 1966:1-2004:12. This scheme is continued until the last statistic is calculated, using an initial in-sample period is 1985:1-1994:12, and the out-of-sample period is 1995:1-2004:12. For each statistic, a recursive estimation scheme is applied, so that the model is re-estimated P times, and P ex-ante 1-step ahead forecast errors are constructed. In the case of F-tests, we simply use $T = R + P$ observations for statistic calculation.

Rtype = 2 – In this sampling scheme, the initial in-sample estimation period is only 10 years in length for the first statistic calculations. Thereafter, the initial in-sample period is increased by 1 year at a time. P is again set to be the remainder of the sample, up until 2004:12. For the first of the 31 calculated statistics, the initial in-sample period is thus 1955:1-1964:12, and the out-of-sample period is 1965:1-2004:12. Now, the sample used for the next statistic adds 1 year of observations to the initial estimation period. Thus, the in-sample period is 1955:1-1965:12, and the out-of-sample period is 1966:1-2004:12. This scheme is continued until the last statistic is calculated, using an initial in-sample period is 1955:1-1994:12, and the out-of-sample period is 1995:1-2004:12.

Rtype = 3 – In this sampling scheme, the initial in-sample estimation period is fixed to be 10 years in length, so that R=120 in all cases. Additionally, P is fixed to be 10 years in length, so that P=120 in all cases. For the first of the 31 calculated statistics, the initial in-sample period is thus 1955:1-1964:12, and the out-of-sample period is 1965:1-1974:12. Now, the sample used for the next statistic rolls the initial estimation period forward 1 year. Thus, the in-sample period is 1956:1-1965:12, and the out-of-sample period is 1966:1-1975:12. This scheme is continued until the last statistic is calculated, using an initial in-sample period is 1985:1-1994:12, and the out-of-sample period is 1995:1-2004:12.

Table 2: Rejection Frequencies of CS Test $-\ a_4 = 0\ -\ T = 600$**,** $P = 0.5T$ *

Model	Recur Block Bootstrap			BB, no PEE, no adj			Standard Block Bootstrap		
	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$
Panel A: $a_2 = 0.3$									
Size1	0.09	0.12	0.12	0.06	0.01	0.01	0.00	0.00	0.00
Size2	0.04	0.07	0.11	0.00	0.04	0.03	0.00	0.00	0.00
Power1	0.75	0.97	0.98	0.01	0.99	0.96	0.98	0.94	0.88
Power2	0.82	0.99	0.99	0.01	1.00	0.98	1.00	0.96	0.94
Power3	0.81	0.99	1.00	0.02	1.00	0.98	0.99	0.94	0.93
Power4	0.75	0.97	0.98	0.00	1.00	0.97	0.98	0.94	0.89
Power5	0.81	0.99	0.99	0.00	1.00	0.98	1.00	0.95	0.93
Power6	0.81	0.98	0.99	0.01	1.00	0.98	1.00	0.94	0.92
Power7	0.57	0.70	0.74	0.00	0.77	0.74	0.76	0.70	0.69
Power8	0.70	0.87	0.91	0.01	0.95	0.90	0.96	0.83	0.81
Power9	0.78	0.98	0.99	0.01	1.00	0.96	1.00	0.94	0.90
Power10	0.57	0.71	0.73	0.00	0.71	0.73	0.73	0.70	0.68
Power11	0.71	0.91	0.93	0.01	0.95	0.90	0.95	0.83	0.80
Power12	0.76	0.97	1.00	0.01	1.00	0.96	1.00	0.94	0.90
Panel B: $a_2 = 0.6$									
Size1	0.07	0.07	0.09	0.00	0.06	0.03	0.00	0.00	0.00
Size2	0.01	0.03	0.04	0.00	0.01	0.04	0.00	0.00	0.00
Power1	0.66	0.94	0.97	0.00	0.69	0.98	0.92	0.90	0.88
Power2	0.72	0.98	0.99	0.00	0.94	0.99	0.98	0.94	0.91
Power3	0.73	0.97	0.99	0.00	0.96	0.99	0.98	0.93	0.90
Power4	0.62	0.93	0.96	0.00	0.42	0.99	0.84	0.90	0.88
Power5	0.67	0.95	0.97	0.00	0.55	0.98	0.92	0.91	0.88
Power6	0.69	0.95	0.96	0.00	0.69	0.99	0.97	0.91	0.87
Power7	0.52	0.65	0.68	0.00	0.25	0.68	0.53	0.64	0.66
Power8	0.65	0.85	0.88	0.00	0.78	0.92	0.96	0.83	0.82
Power9	0.68	0.97	0.98	0.00	0.95	0.99	0.99	0.92	0.86
Power10	0.52	0.64	0.66	0.00	0.23	0.68	0.54	0.62	0.65
Power11	0.60	0.80	0.84	0.00	0.31	0.91	0.87	0.79	0.75
Power12	0.65	0.91	0.97	0.00	0.67	0.99	0.98	0.88	0.84
Panel C: $a_2 = 0.9$									
Size1	0.00	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.00
Size2	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Power1	0.14	0.49	0.71	0.00	0.00	0.08	0.00	0.41	0.64
Power2	0.20	0.63	0.76	0.00	0.00	0.20	0.00	0.64	0.71
Power3	0.24	0.65	0.78	0.00	0.00	0.20	0.03	0.66	0.73
Power4	0.12	0.46	0.67	0.00	0.00	0.04	0.00	0.34	0.67
Power5	0.20	0.62	0.76	0.00	0.00	0.17	0.01	0.59	0.71
Power6	0.23	0.64	0.79	0.00	0.00	0.17	0.01	0.61	0.72
Power7	0.34	0.47	0.46	0.00	0.00	0.05	0.00	0.28	0.40
Power8	0.31	0.63	0.69	0.00	0.00	0.20	0.00	0.65	0.67
Power9	0.22	0.63	0.75	0.00	0.00	0.21	0.01	0.67	0.69
Power10	0.33	0.44	0.46	0.00	0.00	0.04	0.00	0.29	0.41
Power11	0.31	0.61	0.71	0.00	0.00	0.21	0.01	0.64	0.67
Power12	0.25	0.64	0.75	0.00	0.00	0.17	0.01	0.67	0.72

* Notes: All entries are rejection frequencies of the null hypothesis of equal predictive accuracy based on 10% nominal size critical values constructed using the bootstrap approaches discussed above, where l denotes the block length, and empirical bootstrap distributions are constructed using 100 bootstrap statistics. In particular, "Recur Block Bootstrap" is the bootstrap developed in this paper, "BB, no PEE, no adj" is a naive block bootstrap where no parameter estimation error is assumed, and no recentering (i.e. adjustment) is done in parameter estimation or bootstrap statistic construction, and "Standard Block Bootstrap" is the usual block bootstrap that allows for parameter estimation error, but does not recenter parameter estimates or bootstrap statistics. For all models denoted Power$i$, $i = 1, ..., 6$, data are generated with (non) linear Granger causality (see above for further discussion of DGPs. In all experiments, the ex ante forecast period is of length P, which is set equal to $(1/2)T$, where T is the sample size. All models are estimated recursively, so that parameter estimates are updated before each new prediction is constructed. All reported results are based on 500 Monte Carlo simulations. See Table 1 and Section 5 for further details.

Table 3: Rejection Frequencies of CS Test $- a_4 = 1 - T = 600$**,** $P = 0.5T$ *

Model	Recur Block Bootstrap			BB, no PEE, no adj			Standard Block Bootstrap		
	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$
Panel A: $a_2 = 0.3$									
Size1	0.07	0.06	0.07	0.01	0.04	0.01	0.00	0.00	0.00
Size2	0.04	0.05	0.09	0.00	0.04	0.03	0.00	0.00	0.00
Power1	0.68	0.94	0.97	0.00	0.97	0.97	0.91	0.92	0.88
Power2	0.79	0.98	1.00	0.00	1.00	0.99	1.00	0.95	0.94
Power3	0.79	0.99	0.98	0.01	1.00	0.98	0.99	0.94	0.92
Power4	0.63	0.92	0.94	0.00	0.96	0.97	0.91	0.90	0.87
Power5	0.79	0.99	1.00	0.00	1.00	0.99	1.00	0.96	0.93
Power6	0.78	0.99	1.00	0.01	1.00	0.98	1.00	0.95	0.91
Power7	0.52	0.67	0.73	0.00	0.79	0.71	0.77	0.69	0.65
Power8	0.67	0.88	0.92	0.00	0.95	0.89	0.97	0.85	0.80
Power9	0.76	0.98	0.99	0.00	0.99	0.97	0.99	0.95	0.92
Power10	0.53	0.68	0.72	0.00	0.75	0.72	0.77	0.69	0.66
Power11	0.67	0.90	0.93	0.00	0.94	0.90	0.97	0.86	0.79
Power12	0.77	0.97	1.00	0.00	0.99	0.97	1.00	0.95	0.91
Panel B: $a_2 = 0.6$									
Size1	0.01	0.04	0.04	0.00	0.02	0.03	0.00	0.00	0.00
Size2	0.01	0.02	0.04	0.00	0.00	0.02	0.00	0.00	0.00
Power1	0.51	0.87	0.94	0.00	0.17	0.98	0.53	0.88	0.82
Power2	0.69	0.94	0.98	0.00	0.83	1.00	0.99	0.95	0.91
Power3	0.68	0.96	0.99	0.00	0.92	1.00	0.99	0.93	0.90
Power4	0.47	0.86	0.93	0.00	0.07	0.99	0.42	0.86	0.85
Power5	0.62	0.93	0.97	0.00	0.39	1.00	0.93	0.91	0.90
Power6	0.67	0.94	0.95	0.00	0.56	0.99	0.97	0.92	0.86
Power7	0.45	0.57	0.62	0.00	0.20	0.68	0.59	0.63	0.59
Power8	0.59	0.89	0.93	0.00	0.81	0.95	0.96	0.86	0.80
Power9	0.68	0.96	0.98	0.00	0.92	0.99	0.98	0.94	0.92
Power10	0.47	0.56	0.63	0.00	0.19	0.68	0.54	0.59	0.59
Power11	0.58	0.81	0.87	0.00	0.28	0.94	0.93	0.81	0.78
Power12	0.65	0.94	0.96	0.00	0.54	0.99	0.95	0.93	0.89
Panel C: $a_2 = 0.9$									
Size1	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Size2	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Power1	0.09	0.37	0.56	0.00	0.00	0.00	0.00	0.20	0.51
Power2	0.18	0.59	0.76	0.00	0.00	0.17	0.01	0.59	0.73
Power3	0.22	0.60	0.75	0.00	0.00	0.21	0.00	0.65	0.69
Power4	0.08	0.33	0.49	0.00	0.00	0.00	0.00	0.16	0.47
Power5	0.17	0.58	0.75	0.00	0.00	0.12	0.00	0.56	0.71
Power6	0.20	0.62	0.74	0.00	0.00	0.17	0.00	0.63	0.71
Power7	0.27	0.38	0.41	0.00	0.00	0.02	0.01	0.25	0.38
Power8	0.24	0.58	0.69	0.00	0.00	0.26	0.01	0.59	0.65
Power9	0.20	0.61	0.78	0.00	0.00	0.23	0.01	0.64	0.73
Power10	0.28	0.36	0.41	0.00	0.00	0.02	0.01	0.23	0.38
Power11	0.24	0.58	0.67	0.00	0.00	0.29	0.01	0.64	0.66
Power12	0.20	0.61	0.76	0.00	0.00	0.24	0.01	0.65	0.70

* Notes: See notes to Table 2.

Table 4: Rejection Frequencies of Various Tests $-\ a_4 = 0\ -\ T = 600,\ P = 0.5T$ *

Model		Assume $\pi = 0$		Assume $\pi > 0$		Kilian Bootstrap				Recur Block Bootstrap		
	F	DM	CM	DM	CM	DM	CM	CS	CCS	CCS-$l1$	CCS-$l2$	CCS-$l3$
Panel A: $a_2 = 0.3$												
Size1	0.26	0.01	0.04	0.12	0.10	0.13	0.14	0.08	0.09	0.25	0.23	0.20
Size2	0.27	0.01	0.06	0.12	0.12	0.12	0.14	0.01	0.03	0.19	0.21	0.21
Power1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.98
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98	0.83	0.76
Power8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.93
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Power10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	0.98	0.82	0.75
Power11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.93
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Panel B: $a_2 = 0.6$												
Size1	0.26	0.01	0.04	0.11	0.10	0.13	0.10	0.00	0.01	0.19	0.19	0.20
Size2	0.28	0.01	0.07	0.12	0.13	0.13	0.14	0.00	0.01	0.13	0.14	0.16
Power1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
Power7	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.74	1.00	0.96	0.80	0.70
Power8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.96
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00
Power10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73	1.00	0.98	0.82	0.72
Power11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.98	0.92
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
Panel C: $a_2 = 0.9$												
Size1	0.27	0.01	0.06	0.12	0.11	0.09	0.09	0.00	0.01	0.04	0.12	0.15
Size2	0.30	0.02	0.08	0.16	0.14	0.15	0.15	0.00	0.01	0.01	0.04	0.07
Power1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.86	0.98	1.00	0.95	0.95
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.96
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	0.97	0.95
Power4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	0.98	0.98	0.94	0.93
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.98	0.95
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.97	0.94
Power7	1.00	0.97	1.00	0.99	1.00	0.94	0.99	0.55	0.95	0.75	0.60	0.52
Power8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.91	0.85
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.94	0.92
Power10	1.00	0.97	1.00	0.99	1.00	0.93	0.98	0.53	0.95	0.72	0.57	0.52
Power11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	0.89	0.84
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.95	0.92

* Notes: See notes to Table 2. The "Kilian Bootstrap" is the parametric bootstrap outlined in Kilian (1999), and implemented in Clark and McCracken (2004) and McCracken and Saap (2004). Test statistics, denoted by F, DM, CM, CS, and CCS are summarized in Table 1.

Table 5: Rejection Frequencies of Various Tests $-\ a_4 = 1 - T = 600,\ P = 0.5T$ *

Model		Assume $\pi = 0$		Assume $\pi > 0$		Kilian Bootstrap				Recur Block Bootstrap		
	F	DM	CM	DM	CM	DM	CM	CS	CCS	CCS-$l1$	CCS-$l2$	CCS-$l3$
Panel A: $a_2 = 0.3$												
Size1	0.28	0.02	0.07	0.12	0.11	0.11	0.11	0.01	0.04	0.14	0.13	0.12
Size2	0.30	0.03	0.07	0.12	0.11	0.12	0.12	0.00	0.01	0.14	0.14	0.16
Power1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.99
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.88	0.77
Power8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.94
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Power10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.88	0.75
Power11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.95
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel B: $a_2 = 0.6$												
Size1	0.30	0.01	0.07	0.12	0.11	0.11	0.09	0.00	0.01	0.09	0.11	0.12
Size2	0.30	0.03	0.07	0.12	0.10	0.10	0.09	0.00	0.01	0.07	0.09	0.11
Power1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
Power4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	1.00
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
Power7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	0.98	0.82	0.70
Power8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	0.98	0.96
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00
Power10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	0.99	0.82	0.70
Power11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.97	0.92
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98
Panel C: $a_2 = 0.9$												
Size1	0.29	0.02	0.08	0.10	0.10	0.08	0.08	0.00	0.01	0.02	0.02	0.04
Size2	0.32	0.03	0.07	0.11	0.10	0.10	0.09	0.00	0.01	0.01	0.02	0.04
Power1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.78	0.97	0.96	0.89	0.90
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	0.98	0.96
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	0.96	0.92
Power4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.79	0.97	0.95	0.87	0.88
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	0.97	0.95
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	0.93	0.95
Power7	1.00	0.97	1.00	0.99	1.00	0.94	0.98	0.62	0.96	0.77	0.57	0.50
Power8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	0.93	0.89
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.99	0.96	0.94
Power10	1.00	0.97	1.00	0.99	1.00	0.94	0.99	0.61	0.96	0.79	0.56	0.51
Power11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	0.92	0.89
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.96	0.96

* Notes: See notes to Table 4.

Table 6: Predictive Mean Square Errors Based on Inflation and Unemployment Models - Various Subsamples and Sampling Schemes*

Start Year	Rtype=1		Rtype=2		Rtype=3	
	small model	*big model*	*small model*	*big model*	*small model*	*big model*
1965	0.6239	0.6316	0.6239	0.6316	1.0167	1.0149
1966	0.5978	0.6103	0.6329	0.6413	0.9533	0.9920
1967	0.5944	0.6059	0.6335	0.6449	0.9084	0.9453
1968	0.5977	0.6116	0.6363	0.6511	0.9157	0.9613
1969	0.5953	0.6178	0.6363	0.6591	0.8949	0.9699
1970	0.5885	0.6178	0.6267	0.6599	0.8952	0.9909
1971	0.5916	0.6119	0.6295	0.6596	0.9586	1.0178
1972	0.6047	0.6176	0.6438	0.6703	0.9885	1.0145
1973	0.6164	0.6269	0.6564	0.6816	1.0808	1.1127
1974	0.5255	0.5319	0.5603	0.5854	0.7689	0.7891
1975	0.4835	0.4889	0.4929	0.5038	0.6061	0.6221
1976	0.4886	0.4941	0.4849	0.4864	0.5824	0.5957
1977	0.5023	0.5053	0.4953	0.4963	0.6618	0.6607
1978	0.5067	0.5097	0.4933	0.4944	0.6397	0.6332
1979	0.5141	0.5185	0.4882	0.4900	0.6270	0.6181
1980	0.4993	0.5066	0.4598	0.4619	0.5695	0.5690
1981	0.4794	0.4886	0.4226	0.4267	0.5532	0.5512
1982	0.4938	0.5093	0.4158	0.4213	0.5624	0.5698
1983	0.4553	0.5206	0.3810	0.3813	0.4478	0.5545
1984	0.4295	0.4745	0.3799	0.3799	0.3772	0.4581
1985	0.4270	0.4559	0.3873	0.3870	0.3606	0.4146
1986	0.4395	0.4625	0.3982	0.3977	0.3524	0.3926
1987	0.4140	0.4287	0.3751	0.3740	0.2866	0.3099
1988	0.4268	0.4360	0.3881	0.3870	0.2784	0.2924
1989	0.4321	0.4370	0.4008	0.3996	0.2636	0.2713
1990	0.4071	0.4052	0.4102	0.4089	0.2829	0.2831
1991	0.3452	0.3450	0.3796	0.3781	0.2501	0.2522
1992	0.3345	0.3378	0.3915	0.3897	0.2907	0.2952
1993	0.3492	0.3524	0.4117	0.4093	0.2935	0.2974
1994	0.3654	0.3846	0.4329	0.4301	0.3423	0.3634
1995	0.3856	0.4057	0.4559	0.4528	0.3856	0.4057

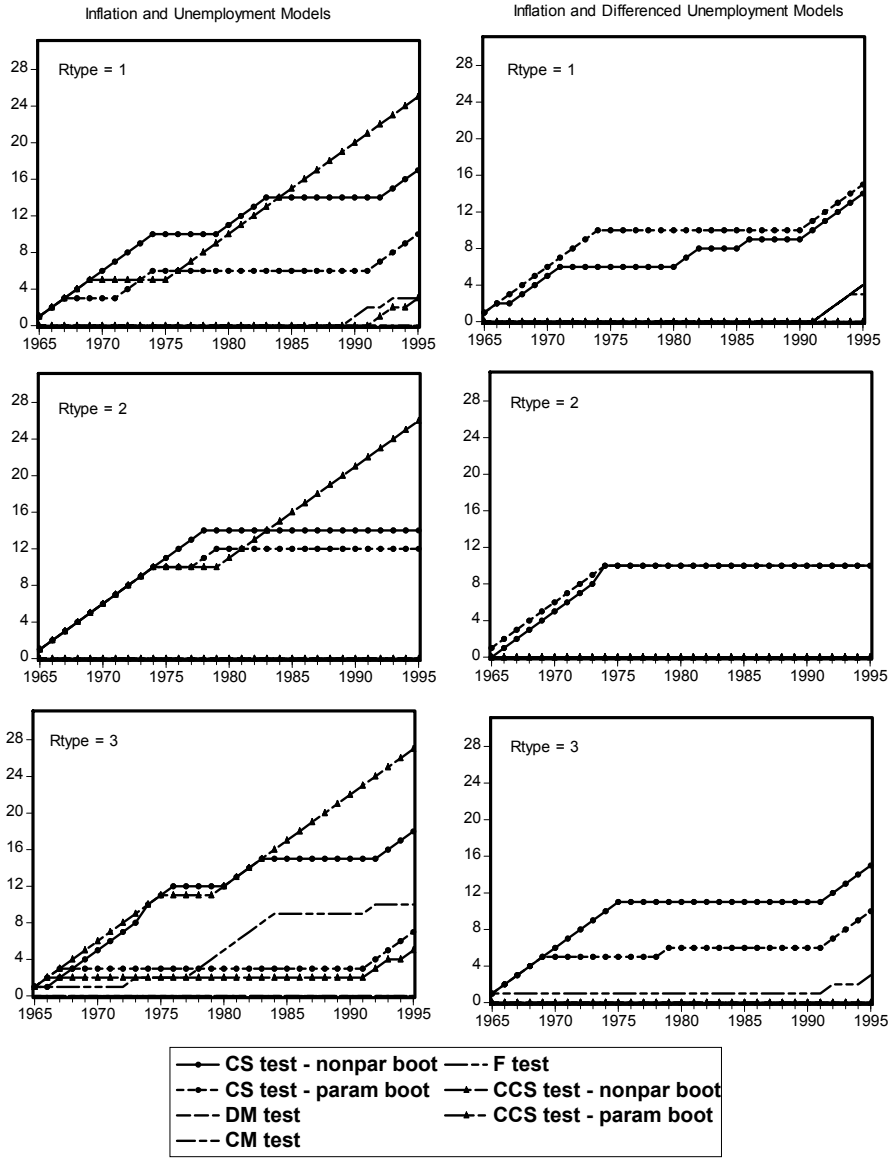
* Notes: Mean square prediction errors (multiplied by 100000) are reported for two models (*small* which contains only lags of inflation, and *big* which contains lags of inflation and unemployment. All predictions are of 1-step ahead inflation, and predictive periods begin in the year given in the first column of entries in the table. The last period of the predictive period varies according to whether the sampling scheme used is Rtype=1, 2, or 3 (see Table 1 for complete details of sampling schemes). Initial in-sample estimation periods vary by Rtype, again as outlined in Table 1. Overall, there are 31 different subsamples considered, corresponding to each row in the table, and each entry in the table summarizes 1-step ahead predictions constructed using P recursively estimated prediction models, where P is the out-of-sample period. The test statistics constructed in the empirical illustration are all based upon the predictive errors for which mean squared values are reported in this table.

Table 7: Predictive Mean Square Errors Based on Inflation and Differenced Unemployment Models - Various Subsamples and Sampling Schemes*

Start Year	Rtype=1		Rtype=2		Rtype=3	
	small model	*big model*	*small model*	*big model*	*small model*	*big model*
1965	0.6239	0.6289	0.6239	0.6289	1.0167	1.0176
1966	0.5978	0.5994	0.6329	0.6368	0.9533	0.9528
1967	0.5944	0.5959	0.6335	0.6370	0.9084	0.9070
1968	0.5977	0.5991	0.6363	0.6403	0.9157	0.9177
1969	0.5953	0.5970	0.6363	0.6399	0.8949	0.8979
1970	0.5885	0.5906	0.6267	0.6305	0.8952	0.8977
1971	0.5916	0.5946	0.6295	0.6363	0.9586	0.9657
1972	0.6047	0.6095	0.6438	0.6507	0.9885	1.0006
1973	0.6164	0.6231	0.6564	0.6636	1.0808	1.0998
1974	0.5255	0.5282	0.5603	0.5650	0.7689	0.7753
1975	0.4835	0.4883	0.4929	0.4993	0.6061	0.6182
1976	0.4886	0.4924	0.4849	0.4916	0.5824	0.5912
1977	0.5023	0.5074	0.4953	0.5026	0.6618	0.6745
1978	0.5067	0.5094	0.4933	0.4979	0.6397	0.6458
1979	0.5141	0.5163	0.4882	0.4905	0.6270	0.6319
1980	0.4993	0.5021	0.4598	0.4610	0.5695	0.5758
1981	0.4794	0.4816	0.4226	0.4246	0.5532	0.5579
1982	0.4938	0.4960	0.4158	0.4178	0.5624	0.5672
1983	0.4553	0.4613	0.3810	0.3820	0.4478	0.4607
1984	0.4295	0.4311	0.3799	0.3815	0.3772	0.3804
1985	0.4270	0.4279	0.3873	0.3887	0.3606	0.3633
1986	0.4395	0.4406	0.3982	0.3995	0.3524	0.3553
1987	0.4140	0.4155	0.3751	0.3755	0.2866	0.2901
1988	0.4268	0.4278	0.3881	0.3887	0.2784	0.2796
1989	0.4321	0.4333	0.4008	0.4018	0.2636	0.2644
1990	0.4071	0.4083	0.4102	0.4110	0.2829	0.2836
1991	0.3452	0.3454	0.3796	0.3814	0.2501	0.2507
1992	0.3345	0.3328	0.3915	0.3930	0.2907	0.2908
1993	0.3492	0.3471	0.4117	0.4131	0.2935	0.2929
1994	0.3654	0.3641	0.4329	0.4346	0.3423	0.3426
1995	0.3856	0.3848	0.4559	0.4586	0.3856	0.3848

* Notes: See notes to Table 6.

Figure 1: Out-of-Sample Tests -- Predictability of Unemployment for Inflation



Notes: Empirical results based on CS, DM, CM, F, and CCS tests are summarized (see paper for a detailed explanation of all mnemonics used) for a variety of data sub-samples. Results in the first column depict the cumulative number of rejections for each of the different sampling procedures (Rtype=1 in the first graph, Rtype=2 in the second graph, and Rtype=3 in the third graph), for the case where Inflation is modelled using Inflation and/or Unemployment. Graphs in the second column are analogous, except that Unemployment is modelled in differences. In all cases, 31 different versions of each test are constructed for each Rtype. Each of the 31 statistics is based on different in sample and out-of sample periods, where the out-of-sample period begins in 1965 for the first statistic, 1966 for the second statistic, and so on, until the last statistic, for which the out-of-sample period begins in 1995. In Rtype=1 and Rtype=2, the end of the out-of-sample period is always 2004:12, while in Rtype=3, the length of the out-of-sample period is fixed at 10 years (the in-sample period is also fixed at 10 years in this case). See paper for further details about Rtype=1,2,3. Each test statistic is calculated using the forecast errors from sequences of 1-step ahead predictions constructed using recursively estimated forecasting models both with and without Unemployment. Rejection denote cases where Unemployment was found to be relevant, from a forecasting perspective.