Bootstrap Procedures for Recursive Estimation Schemes With Applications to Forecast Model Selection[*]

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

June 2004

Abstract

In recent years it has become apparent that many of the classical testing procedures used to select amongst alternative economic theories and economic models are not realistic. In particular, researchers have become more aware of the fact that parameter estimation error and data dependence play a crucial role in test statistic limiting distributions, a role which had hitherto been ignored to a large extent. Given the fact that one of the primary ways for comparing different models and theories is via use of predictive accuracy tests, it is perhaps 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 (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 out of sample predictive ability, for given loss function. 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 viewed as approximations of some unknown underlying model. 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 (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 appropriate re-centering of the bootstrap score is required in order to ensure that the tests are properly sized, and the need for such re-centering is shown to arise quite naturally when testing hypotheses of predictive accuracy. The results of a Monte Carlo investigation of the ICM test suggest that the bootstrap procedure proposed in this paper yield tests with reasonable finite sample properties for samples with as few as 300 observations.

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

----

1 Introduction

In recent years it has become apparent that many of the classical testing procedures used to select amongst alternative economic theories and economic models are not realistic. In particular, researchers have become more aware of the fact that parameter estimation error and data dependence play a crucial role in test statistic limiting distributions, a role which had hitherto been ignored to a large extent. Given the fact that one of the primary ways for comparing different models and theories is via use of predictive accuracy tests, it is perhaps 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 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. Furthermore, 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

---

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

1

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 (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,

2

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 (nonlinear) alternatives. 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. In practice, one model is chosen as a benchmark against which all others are to be compared. Typically, this is either a simple model, a model frequently used in empirical work, or a model in which one is particularly interested. The objective is to compare such a model against a set, finite but arbitrarily large, of competing models, which can all be misspecified. Of course, if the benchmark model were to be sequentially compared with each of the competitors, then the well known issue of sequential test bias would arise, and valid inference would be difficult to carry out. 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 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. The ICM test is examined via a series of Monte Carlo experiments, and it is found that the bootstrap procedure proposed in this paper yield tests with reasonable empirical level, and with power that is moderately good for samples with as few as 300 observations.

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. 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 P^*-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.

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:[2]

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

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. Following standard practice (such as in the real-time forecasting literature), this estimator is first computed using R observations. In our applications 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:[3] 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 $W_{i+1}, ..., W_{i+l}$,

---

[2]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.

[3]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-

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. 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, $\widehat{\theta}_{i,t}^*$, to be the direct analog of $\widehat{\theta}_{i,t}$. Namely,

$$\widehat{\theta}_{i,t}^* = \arg\min_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=s}^{t} q_i(y_j^*, Z^{*,j-1}, \theta_i), \ R \le t \le T - 1, \ i = 1, ..., n. \tag{3}$$

By first order conditions, $\frac{1}{t} \sum_{j=s}^{t} \nabla_\theta q_i(y_j^*, Z^{*,j-1}, \widehat{\theta}_{i,t}^*) = 0$, and via a mean value expansion of $\frac{1}{t} \sum_{j=s}^{t} \nabla_\theta q_i(y_j^*, Z^{*,j-1}, \widehat{\theta}_{i,t}^*)$ around $\widehat{\theta}_{i,t}$, after a few simple manipulations, we have that

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{i,t}^* - \widehat{\theta}_{i,t} \right)$$

$$= \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\left(\frac{1}{t} \sum_{j=s}^{t} \nabla_\theta^2 q_i(y_j^*, Z^{*,j-1}, \overline{\theta}_{i,t}^*) \right)^{-1} \frac{1}{t} \sum_{j=s}^{t} \nabla_\theta q_i(y_j^*, Z^{*,j-1}, \widehat{\theta}_{i,t}) \right)$$

$$= B_i^\dagger \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\frac{1}{t} \sum_{j=s}^{t} \nabla_\theta q_i(y_j^*, Z^{*,j-1}, \widehat{\theta}_{i,t}) \right) + o_{P^*}(1) \ \Pr - P$$

$$= B_i^\dagger \frac{a_{R,0}}{\sqrt{P}} \sum_{t=s}^{R} \nabla_\theta q_i(y_j^*, Z^{*,j-1}, \widehat{\theta}_{i,t}) + B_i^\dagger \frac{1}{\sqrt{P}} \sum_{j=1}^{P-1} a_{R,j} \nabla_\theta q_i(y_{R+j}^*, Z^{*,R+j-1}, \widehat{\theta}_{i,t})$$

$$+ o_{P^*}(1) \ \Pr - P, \tag{4}$$

where $\overline{\theta}_{i,t}^* \in \left(\widehat{\theta}_{i,t}^*, \widehat{\theta}_{i,t} \right)$, $B_i^\dagger = E \left(\nabla_\theta^2 q_i(y_j, Z^{j-1}, \theta_i^\dagger) \right)^{-1}$, $a_{R,j} = \frac{1}{R+j} + \frac{1}{R+j+1} + ... + \frac{1}{R+P-1}$, $j = 0, 1, ..., P - 1$, and where the last equality on the right hand side of (4) follows immediately, using

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 bootstrap is more efficient than the non overlapping block bootstrap.

the same arguments as those used in Lemma A5 of West (1996). Analogously,

$$
\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{i,t} - \theta_i^\dagger \right)
$$

$$
= B_i^\dagger \frac{a_{R,0}}{\sqrt{P}} \sum_{t=s}^{R} \nabla_\theta q_i(y_j, Z^{j-1}, \theta_i^\dagger) + B_i^\dagger \frac{1}{\sqrt{P}} \sum_{j=1}^{P-1} a_{R,j} \nabla_\theta q_i(y_{R+j}, Z^{R+j-1}, \theta_i^\dagger) + o_P(1). \qquad (5)
$$

Now, given (2), $E \left(\nabla_\theta q_i(y_j, Z^{j-1}, \theta_i^\dagger) \right) = 0$ for all j, and $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{i,t} - \theta_i^\dagger \right)$ 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,

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

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)).[4] Now, $\frac{1}{T-s} \sum_{k=s}^{T-1} \nabla_\theta q_i(y_k, Z^{k-1}, \widehat{\theta}_{i,t}) \neq 0$, and is instead of order $O_P \left(T^{-1/2} \right)$. Thus, $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \frac{1}{T-s} \sum_{k=s}^{T-1} \nabla_\theta q_i(y_k, Z^{k-1}, \widehat{\theta}_{i,t}) = O_P(1)$, and does not vanish in probability. This clearly contrasts with the full sample case, in which $\frac{1}{T-s} \sum_{k=s}^{T-1} \nabla_\theta q_i(y_k, Z^{k-1}, \widehat{\theta}_{i,T}) = 0$, because of the first order conditions. Thus, $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{i,t}^* - \widehat{\theta}_{i,t} \right)$ cannot have a zero mean normal 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 $\frac{1}{T-s} \sum_{k=s}^{T-1} \nabla_\theta q_i(y_k, Z^{k-1}, \widehat{\theta}_{i,t})$. Hence, define a new bootstrap estimator, $\widetilde{\theta}_{i,t}^*$, as:

$$
\widetilde{\theta}_{i,t}^* = \arg \min_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=s}^{t} \left(q_i(y_j^*, Z^{*,j-1}, \theta_i) - \theta_i' \left(\frac{1}{T} \sum_{k=s}^{T-1} \nabla_{\theta_i} q_i(y_k, Z^{k-1}, \widehat{\theta}_{i,t}) \right) \right), \qquad (7)
$$

$R \leq t \leq T-1$, $i = 1, ..., n$.[5]

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

---

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

[5]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 $\frac{1}{T}$.

manipulations, we have that

$$
\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}_{i,t}^{*} - \widehat{\theta}_{i,t} \right)
$$

$$
= B^{\dagger} \frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\frac{1}{t} \sum_{j=s}^{t} \left(\nabla_{\theta} q(y_{j}^{*}, Z^{*,j-1}, \widehat{\theta}_{i,t}) - \left(\frac{1}{T} \sum_{k=s}^{T-1} \nabla_{\theta} q(y_{k}, Z^{k-1}, \widehat{\theta}_{i,t}) \right) \right) \right)
$$
$$
+ o_{P^{*}}(1) \ \mathrm{Pr} - P.
$$

Given (6), it is immediate to see that the bias associated with $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}_{i,t}^{*} - \widehat{\theta}_{i,t} \right)$ 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) $\theta_i^{\dagger}$ is uniquely identified (i.e. $E(q_i(y_t, Z^{t-1}, \theta_i)) > E(q_i(y_t, Z^{t-1}, \theta_i^{\dagger}))$ for any $\theta_i \neq \theta_i^{\dagger}$); (ii) q_i is twice continuously differentiable on the interior of Θ_i, for $i = 1, ..., n$, and for Θ_i a compact subset of $R^{\varrho(i)}$; (iii) the elements of $\nabla_{\theta_i} q_i$ and $\nabla_{\theta_i}^2 q_i$ are $p-$dominated on Θ_i, with $p > 2(2+\psi)$, where ψ is the same positive constant as defined in Assumption A1; and (iv) $E\left(-\nabla_{\theta_i}^2 q_i(\theta_i)\right)$ is negative definite uniformly on Θ_i.[6]

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 $\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \left(\widehat{\theta}_{i,t} - \theta_i^{\dagger} \right) = o_P(1)$ 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 $R \to \infty$,

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

---

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

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

Broadly speaking, Theorem 1 states that $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widetilde{\theta}_{i,t}^* - \theta_i^\dagger \right)$ has the same limiting distribution as $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_{i,t} - \theta_i^\dagger \right)$, 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.

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 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 generic 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

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

where

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

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

9

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.[7] Of course, there are also applications for which parameter estimation error does not vanish, even if the same loss function 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.[8] 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 $c_{kk} =$

---

[7]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.)

[8]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 model in some optimal way. For very recent contributions along these lines, see Elliott and Timmermann (2004a,b).

$\lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=s}^{T}\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, $k = 2,...,n$, 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

$v_{i,i} = S_{g_i g_i} + 2\Pi\mu_1' B_1^\dagger C_{11} B_1^\dagger \mu_1 + 2\Pi\mu_i' B_i^\dagger C_{ii} B_i^\dagger \mu_i - 4\Pi\mu_1' B_1^\dagger C_{1i} B_i^\dagger \mu_i + 2\Pi S_{g_i q_1} B_1^\dagger \mu_1 - 2\Pi S_{g_i q_i} B_i^\dagger \mu_i,$

where $S_{g_i g_i} = \sum_{\tau=-\infty}^{\infty} E\left(\left(g(u_{1,1}) - g(u_{i,1})\right)\left(g(u_{1,1+\tau}) - g(u_{i,1+\tau})\right)\right),$

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

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

$B_i^\dagger = \left(E\left(-\nabla_{\theta i}^2 q_i(y_t, Z^{t-1}, \theta_i^\dagger)\right)\right)^{-1}$, $\mu_i = E\left(\nabla_{\theta_i}g(u_{i,t+1})\right)$, 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 $\max_{k=2,...,n}\left(S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, 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 $\max_{k=2,...,n}\left(S_P(1,k) - \sqrt{P}E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right)\right)$, asymptotically. Of course, when H_A holds, the statistic diverges to plus infinity at rate $\sqrt{P}$.[9]

---

[9]For more discussion of the properties of this variety of test, the reader is referred to Corradi and Swanson (2004a,b), 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

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 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:

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

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

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

computed over the various subsamples, properly mimics the distribution of the statistic. Thus, it provides valid critical values even for the case of $\max_{k=2,...,m} E\left(g(u_{1,t+1}) - g(u_{k,t+1})\right) = 0$, 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 (2003), in the context of testing for stochastic dominance.

where

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

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:

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

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,

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

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 $R \to \infty$,

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

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.

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 well defined limiting distributions when the benchmark is nested with all competing

models (see e.g. Corradi and Swanson (2002, 2004c) and McCracken (2004)).[10] Tests for comparing the predictive ability of a fixed number of nested models have previously also been suggested. For example, Clark and McCracken (2001,2003) propose 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 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

---

[10]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.

causality tests can be constructed in this manner).[11] 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 Schörfheide (2004) for further discussion of this issue)[12]. 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 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. For this reason, H_0 corresponds

---

[11]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.

[12]In the context of multi-step ahead vector autoregressive prediction, Schörfheide (2003) 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.

to equal predictive accuracy, while H_A corresponds to the case where the alternative model out-performs 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

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

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[13]

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

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:

$$M_P^* = \int_\Gamma m_P^*(\gamma)^2 \phi(\gamma) d\gamma,$$

---

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

where, recalling that $g = q_1$,

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

$$\left. - \frac{1}{T} \sum_{i=1}^{T-1} g' \left(y_{i+1} - \begin{pmatrix} 1 & y_i \end{pmatrix} \widehat{\theta}_{1,t} \right) w(Z^i, \gamma) \right) \tag{19}$$

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.[14]

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 $R \to \infty$,

$$P \left(\omega : \sup_{v \in \Re} \left| P_T^* \left(\int_\Gamma m_P^*(\gamma)^2 \phi(\gamma) d\gamma \le v \right) - P \left(\int_\Gamma m_P^\mu(\gamma)^2 \phi(\gamma) d\gamma \le v \right) \right| > \varepsilon \right) \to 0,$$

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

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.

5   Monte Carlo Results

In this section we carry out a series of Monte Carlo experiments using the out-of-sample ICM test that are closely related to those carried out by Corradi and Swanson (2002), although parameterizations are in some instances different, and critical values used in that paper where constructed

---

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

using the conditional p-value approach. We begin by generating data according to the following data generating processes:

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

Size1: $y_t = a_1 + a_2 y_{t-1} + u_{2,t}, \; u_{2,t} \sim iidN(0,1)$

Size2: $y_t = a_1 + a_2 y_{t-1} + a_3 u_{2,t-1} + u_{2,t}$

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

Power2 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(\tan^{-1}(x_{t-1}/2)) + u_{2,t}$

Power3 : $y_t = a_1 + a_2 y_{t-1} + x_{t-1} + u_{2,t}$

Power4 : $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1} + u_{2,t}$

Power5 : $y_t = a_1 + a_2 y_{t-1} + x_{t-1}1\{x_{t-1} > a_1/(1-a_2)\} + u_{2,t}$

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

Power7 : $y_t = a_1 + a_2 y_{t-1} + \exp(\tan^{-1}(x_{t-1}/2)) + a_3 u_{2,t-1} + u_{2,t}$

Power8 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(\tan^{-1}(x_{t-1}/2)) + a_3 u_{2,t-1} + u_{2,t}$

Power9 : $y_t = a_1 + a_2 y_{t-1} + x_{t-1} + a_3 u_{2,t-1} + u_{2,t}$

Power10: $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1} + a_3 u_{2,t-1} + u_{2,t}$

Power11: $y_t = a_1 + a_2 y_{t-1} + x_{t-1}1\{x_{t-1} > a_1/(1-a_2)\} + a_3 u_{2,t-1} + u_{2,t}$

Power12: $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1}1\{x_{t-1} > a_1/(1-a_2)\} + a_3 u_{2,t-1} + u_{2,t}.$

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 y_{t-1}. Thus, our focus is on (non)linear out-of-sample Granger causality testing. The functional forms that are specified under the alternative include: (i) exponential (*Power1,Power2*); (ii) linear (Power3,Power4); and (iii) self exciting threshold (*Power5,Power6*). In addition, *Power7-Power12* are the same as *Power1-Power6*, except that an MA(1) term is added. Notice that *Power1* and *Power2* include a nonlinear term that is similar in form to the test function, $g(\cdot)$. Also, *Power3* and *Power4* serve as linear causality benchmarks. 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 $\Gamma = [0.0, 5.0] \times [0.0, 5.0]$. We consider a grid that is delineated by increments of size 0.1, so that overall we have 10000 (100×100) evaluation points (with the point $\{0,0\}$ being omitted). The statistics M_P and $|M_P|$ have been computed as simple averages over the 10000 evaluation points, while M_P^{sup} has been computed as the maximum over the evaluations points. 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=150, T=300, T=600, and $T = 1200$ observations are used. For the sake of brevity, results for T=600 and T =1200 are not included and are available upon request, although there is little additional to see in these results, as power increases substantially as T increases, just as is the case when one moves from $T = 150$ to $T = 300$ observations, and all tests have empirical rejection frequencies that are fairly close to nominal test levels.[15] In addition, the following parameterization are used: $a_1 = 1.0$, $a_2 = \{0.3, 0.6\}$, and $a_3 = 0.3$. Finally, bootstrap critical values are constructed using 100 simulated statistics, the block length, l, is set equal to $\{2, 4, 5\}$, and $P = (2/3)T$.

Findings are summarized in Tables 1-2. The first column in the tables states the model type (e.g. *Size1)*, and numerical entries are rejection frequencies for the three different statistics (M_P, M_P^{sup}, and $|M_P|$), as well as for various values of l and a_2. Although results are only reported for the case where $P = (2/3)T$, additional results for $P = 0.4T$, $0.5T$, and $0.6T$ were also tabulated. These results are qualitatively the same as those reported, and are available upon request from the authors. Overall, results are quite clear-cut, as is evidenced by inspection of the tables. Under H_0, the empirical level of the test is often fairly close to the nominal 10% level, regardless of whether M_P, M_P^{sup}, or $|M_P|$ is used, and the empirical level improves when the sample size is increased. Of note is that our empirical level results suggest that the block bootstrap may be preferable to the conditional $p-$value approach reported on by Corradi and Swanson (2002) for this test, as they obtained values as high as 0.15, even though the smallest sample they considered was $T = 600$.[16]

The finite sample rejection frequency is high under the alternatives denoted *Power1* to *Power 12* (often above 0.50 and sometimes as high as 0.80), even for samples as small as 300 observations (see the upper panel of Table 2). Not surprisingly, power is very poor for very small samples (see Table 1, where $T = 150$). Indeed, results for $T = 150$ are included for precisely this reason; namely to illustrate the substantial power increases when even moderately larger samples of 300 observations are used, and to illustrate the pitfalls associated with using small samples with these sorts of ICM type tests. Also, when the degree of dependence is increased (i.e. from $a_2 = 0.3$ to

---

[15]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 these tests.

[16]The results reported by Corradi and Swanson (2002) worsen considerably when samples smaller than $T = 600$ are used, particularly with regard to power. Complete results comparing the two approaches are avialable upon request from the authors.

$a_2 = 0.6$), power is much poorer (compare *Panels A* and *B* in either Table 1 or Table 2), although the rate at which power increases as the sample is increased appears to be somewhat immune to the degree of dependence. Of course, and as mentioned above, it should be stressed that power increases considerably when larger samples are used. For example, when $T = 600$, most power entries for the case where $a_2 = 0.3$ range from 0.75-0.99, while the case where $a_2 = 0.6$ has most entries ranging from 0.40-0.78, when $P = (2/3)T$ (results are available upon request).

Of final note is that the empirical test level is reasonably close to nominal, even when $T = 150$, regardless of whether there is an MA component in the error process (compare *Size1* with *Size2* in *Panel A),* as long as the data exhibits little dependence (i.e. $a_2 = 0.3$). When $a_2 = 0.6$ and $T = 150$, however, empirical level is very poor when there is an MA error component (see *Size 2* in *Panel B*). Interestingly, though, empirical level becomes robust to data dependence quite quickly as the sample is increased, as can be noted by examining empirical level figures in Table 2, *Panel B*, and comparing these with the figures reported in Table 1, *Panel B*. In summary, the block bootstrap for recursive m-estimators is not only easy to apply, but also appears to generate reasonable inference, at least when applied to the out-of-sample ICM test.

6   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, the results of a Monte Carlo investigation of the ICM test suggest that the bootstrap procedure proposed in this paper yield tests with reasonable finite sample properties for samples with as few as 300 observations.

7 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:

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

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:

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

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,

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

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 $o_{P^*}(1)\,\mathrm{Pr}-P$, given the uniform

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

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

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

$$
+ o_{P^*}(1) \ \Pr - P, \tag{22}
$$

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

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

$$
= B^\dagger \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\frac{1}{t} \sum_{j=s}^{t} \left(\nabla_\theta q(y_j^*, Z^{*,j-1}, \theta^\dagger) - \left(\frac{1}{T} \sum_{k=s}^{T-1} \nabla_\theta q(y_k, Z^{k-1}, \theta^\dagger) \right) \right) \right)
$$

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

$$
\times \left(\widehat{\theta}_t - \theta^\dagger \right) \right) + o_{P^*}(1) \ \Pr - P. \tag{23}
$$

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:

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

$$= B^{\dagger} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\frac{1}{t} \sum_{j=s}^{t} \left(\nabla_{\theta} q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) - E^* \left(\nabla_{\theta} q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) \right) \right) \right) + o_{P^*}(1) \ \Pr - P$$

$$= B^{\dagger} \frac{a_{R,0}}{\sqrt{P}} \sum_{j=1}^{R} \left(\nabla_{\theta} q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) - E^* \left(\nabla_{\theta} q(y_j^*, Z^{*,j-1}, \theta^{\dagger}) \right) \right)$$

$$+ B^{\dagger} \frac{1}{\sqrt{P}} \sum_{j=1}^{P-1} a_{R,j} \left(\nabla_{\theta} q(y_{R+j}^*, Z^{*,R+j-1}, \theta^{\dagger}) - E^* \left(\nabla_{\theta} q(y_{R+j}^*, Z^{*,R+j-1}, \theta^{\dagger}) \right) \right)$$

$$+ o_{P^*}(1) \ \Pr - P, \tag{24}$$

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:

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

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 $R=b_1l$ and $P=b_2l$, where $b_1+b_2=b$. It then follows that, up to a term of order $O(l/R^{1/2})$,

$$Var^*\left(a_{R,0}\frac{1}{\sqrt{R}}\sum_{j=1}^{R}h^*_j\right) \;=\; a^2_{R,0}Var^*\left(\frac{1}{\sqrt{R}}\sum_{k=1}^{b_1}\sum_{i=1}^{l}h_{I_k+i}\right)$$

$$=\; a^2_{R,0}E^*\left(\frac{1}{R}\sum_{k=1}^{b_1}\sum_{i=1}^{l}\sum_{k=1}^{l}(h_{I_k+i}-\overline{h}_T)(h_{I_k+j}-\overline{h}_T)'\right)$$

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

Thus,

$$\frac{R}{P}Var^*\left(a_{R,0}\frac{1}{\sqrt{R}}\sum_{j=1}^{R}h^*_j\right)$$

$$=\;\; \frac{Ra^2_{R,0}}{P}\sum_{j=-l}^{l}\gamma_j+\frac{Ra^2_{R,0}}{P}\left(\frac{1}{R}\sum_{t=l}^{R-l}\sum_{j=-l}^{l}((h_t-\overline{h}_T)(h_{t+j}-\overline{h}_T)'-\gamma_j)\right)+O\left(\frac{l^2}{R}\right),\qquad(26)$$

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

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

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

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

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

$$= \frac{1}{P} \sum_{k=b_1+1}^{b} \sum_{i=1}^{l} \sum_{j=1}^{l} a_{R,((k-b_1-1)l+i)} a_{R,((k-b_1-1)l+j)} \gamma_{i-j}$$

$$+ \frac{1}{P} \sum_{k=b_1+1}^{b} \sum_{i=1}^{l} \sum_{j=1}^{l} a_{R,((k-b_1-1)l+i)} a_{R,((k-b_1-1)l+j)} \left(\frac{1}{T} \sum_{t=l}^{T-l} \left((h_{t+i} - \overline{h}_T)(h_{t+j} - \overline{h}_T)' - \gamma_{i-j} \right) \right)$$

$$+ O(l/P^{1/2}) \ \Pr-P \tag{27}$$

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

$$\left| \frac{b_2}{P} \sum_{i=1}^{l} \sum_{j=1}^{l} \left(\frac{1}{T} \sum_{t=l}^{T-l} \left((h_{t+i} - \overline{h}_T)(h_{t+j} - \overline{h}_T)' - \gamma_{i-j} \right) \right) \right|$$

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

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).[17] With regard to the first term on the RHS of the last equality in (27), note that:

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

$$= \frac{1}{P} \sum_{t=l}^{P-l} a_{R,t}^2 \sum_{j=-l}^{l} \gamma_j + \frac{1}{P} \sum_{t=l}^{P-l} \sum_{j=-l}^{l} (a_{R,t} a_{R,t+j} - a_{R,t}^2) \gamma_j + O(l/P^{1/2}) \ \Pr-P.$$

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

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

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. $\square$

---

[17]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 $\bar{u}_{i,t} = y_t - \kappa(Z^{t-1}, \bar{\theta}_{i,t})$, with $\bar{\theta}_{i,t} \in \left(\widehat{\theta}_{i,t}, \theta^{\dagger} \right)$. Via a mean value expansion, and given Assumptions A1-A2:

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

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.$\square$

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

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

29

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 $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left((g(u_{1,t+1}) - g(u_{k,t+1})) - E\left(g(u_{1,t+1}) - g(u_{k,t+1}) \right) \right)$, 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:

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

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 $\frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g'(u_{1,t+1}^*)w(Z^{*,t},\gamma) - \frac{1}{T}\sum_{j=2}^{T-1} g'(\widehat{u}_{1,j+1}^{(t)})w(Z^j,\gamma) \right)$ 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.$\square$

8 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.

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 (2003), 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), "Predictive Density Accuracy Tests", Working Paper, Rutgers University and Queen Mary, University of London.

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

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.

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.

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, (2003), Consistent Testing for Stochastic Dominance Under General Sampling Schemes, Manuscript, LSE, Southern Methodist University and Ewha University.

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

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.

Schörfheide, F., (2003), 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: Rejection Frequencies of Various M_P Tests Using the Bootstrap for Recursive m-Estimators: T=150, P=2/3 *

Model	l=2			l=4			l=5		
	M_P	M_P^{sup}	$|M_P|$	M_P	M_P^{sup}	$|M_P|$	M_P	M_P^{sup}	$|M_P|$
Panel A: $a_2 = 0.3$									
Size1	0.10	0.10	0.10	0.06	0.06	0.05	0.06	0.07	0.06
Size2	0.08	0.09	0.08	0.06	0.06	0.05	0.05	0.05	0.04
Power1	0.33	0.30	0.37	0.22	0.16	0.23	0.19	0.16	0.23
Power2	0.46	0.42	0.54	0.26	0.25	0.34	0.26	0.23	0.29
Power3	0.48	0.42	0.51	0.29	0.23	0.35	0.18	0.12	0.25
Power4	0.54	0.49	0.63	0.32	0.27	0.37	0.31	0.26	0.36
Power5	0.53	0.44	0.56	0.29	0.26	0.35	0.27	0.27	0.29
Power6	0.61	0.54	0.67	0.32	0.28	0.40	0.29	0.26	0.35
Power7	0.25	0.20	0.27	0.16	0.12	0.19	0.11	0.11	0.16
Power8	0.43	0.39	0.46	0.26	0.21	0.33	0.26	0.20	0.27
Power9	0.42	0.36	0.45	0.31	0.29	0.38	0.21	0.19	0.27
Power10	0.56	0.48	0.63	0.30	0.21	0.35	0.33	0.22	0.37
Power11	0.45	0.38	0.54	0.26	0.21	0.30	0.23	0.19	0.29
Power12	0.63	0.58	0.69	0.33	0.28	0.36	0.30	0.27	0.33
Panel B: $a_2 = 0.6$									
Size1	0.09	0.10	0.10	0.06	0.06	0.04	0.03	0.06	0.02
Size2	0.04	0.05	0.03	0.03	0.04	0.02	0.04	0.06	0.03
Power1	0.16	0.14	0.24	0.11	0.08	0.12	0.08	0.06	0.07
Power2	0.23	0.20	0.27	0.13	0.10	0.16	0.11	0.08	0.13
Power3	0.34	0.31	0.36	0.16	0.14	0.17	0.14	0.11	0.17
Power4	0.28	0.26	0.35	0.18	0.17	0.19	0.15	0.11	0.19
Power5	0.33	0.28	0.42	0.24	0.20	0.25	0.13	0.12	0.18
Power6	0.35	0.32	0.42	0.17	0.16	0.21	0.13	0.12	0.12
Power7	0.11	0.10	0.15	0.06	0.08	0.08	0.06	0.03	0.07
Power8	0.19	0.16	0.24	0.12	0.12	0.16	0.11	0.10	0.15
Power9	0.28	0.25	0.32	0.13	0.12	0.21	0.11	0.08	0.13
Power10	0.30	0.28	0.36	0.16	0.14	0.17	0.17	0.12	0.18
Power11	0.33	0.25	0.35	0.17	0.15	0.17	0.16	0.15	0.16
Power12	0.34	0.30	0.37	0.16	0.16	0.20	0.16	0.15	0.17

* 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 approach discussed above, where l denotes the block length, and empirical bootstrap distributions are constructed using 100 bootstrap statistics. For all models denoted Power$i$, $i = 1, ..., 12$, data are generated with (non) linear Granger causality. In all experiments, the ex ante forecast period is of length P, which is set equal to $(2/3)$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 2000 Monte Carlo simulations. See above for further details.

Table 2: Rejection Frequencies of Various M_P Tests Using the Bootstrap for Recursive

m-Estimators: T=300, P=2/3 *

Model	l=2			l=4			l=5		
	M_P	M_P^{sup}	$|M_P|$	M_P	M_P^{sup}	$|M_P|$	M_P	M_P^{sup}	$|M_P|$
Panel A: $a_2 = 0.3$									
Size1	0.10	0.09	0.11	0.06	0.06	0.08	0.07	0.07	0.07
Size2	0.08	0.08	0.08	0.07	0.07	0.07	0.05	0.05	0.09
Power1	0.54	0.44	0.61	0.42	0.36	0.47	0.39	0.35	0.45
Power2	0.78	0.71	0.84	0.57	0.46	0.69	0.56	0.47	0.64
Power3	0.75	0.65	0.83	0.60	0.46	0.74	0.49	0.48	0.54
Power4	0.86	0.72	0.88	0.64	0.54	0.66	0.57	0.51	0.64
Power5	0.80	0.73	0.88	0.65	0.59	0.70	0.60	0.50	0.66
Power6	0.87	0.82	0.90	0.67	0.64	0.76	0.67	0.58	0.71
Power7	0.40	0.34	0.47	0.36	0.26	0.40	0.30	0.24	0.39
Power8	0.71	0.61	0.79	0.55	0.43	0.66	0.49	0.37	0.57
Power9	0.69	0.59	0.74	0.50	0.42	0.63	0.43	0.36	0.54
Power10	0.80	0.72	0.87	0.63	0.54	0.68	0.60	0.50	0.69
Power11	0.74	0.67	0.82	0.56	0.51	0.64	0.59	0.51	0.67
Power12	0.85	0.79	0.90	0.68	0.62	0.74	0.62	0.56	0.69
Panel B: $a_2 = 0.6$									
Size1	0.11	0.10	0.12	0.09	0.09	0.10	0.09	0.09	0.08
Size2	0.07	0.06	0.07	0.07	0.07	0.08	0.07	0.07	0.07
Power1	0.26	0.21	0.34	0.14	0.13	0.19	0.12	0.10	0.17
Power2	0.35	0.27	0.47	0.23	0.18	0.33	0.21	0.15	0.28
Power3	0.48	0.41	0.57	0.27	0.25	0.35	0.28	0.23	0.32
Power4	0.50	0.44	0.62	0.31	0.24	0.39	0.27	0.25	0.34
Power5	0.59	0.52	0.68	0.32	0.28	0.49	0.29	0.24	0.36
Power6	0.61	0.53	0.71	0.35	0.28	0.46	0.32	0.26	0.43
Power7	0.17	0.12	0.19	0.10	0.09	0.14	0.08	0.08	0.13
Power8	0.37	0.21	0.45	0.19	0.14	0.27	0.17	0.10	0.22
Power9	0.48	0.38	0.53	0.27	0.21	0.36	0.20	0.17	0.28
Power10	0.50	0.41	0.59	0.31	0.25	0.37	0.26	0.23	0.34
Power11	0.57	0.50	0.63	0.31	0.29	0.39	0.32	0.22	0.39
Power12	0.59	0.52	0.68	0.41	0.35	0.48	0.27	0.24	0.35

* Notes: See notes to Table 1.