Armah, Nii Ayi; Swanson, Norman R.

Working Paper

Predictive inference under model misspecification with an application to assessing the marginal predictive content of money for output

Working Paper, No. 2011-03

Provided in Cooperation with:
Department of Economics, Rutgers University

Suggested Citation: Armah, Nii Ayi; Swanson, Norman R. (2011) : Predictive inference under model misspecification with an application to assessing the marginal predictive content of money for output, Working Paper, No. 2011-03, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:
https://hdl.handle.net/10419/59496

Predictive Inference Under Model Misspecification with an Application to Assessing the Marginal Predictive Content of Money for Output *

Nii Ayi Armah and Norman R. Swanson
Rutgers University

August 2006
this version: December 2006

Abstract

In this chapter we discuss model selection and predictive accuracy tests in the context of parameter and model uncertainty under recursive and rolling estimation schemes. We begin by summarizing some recent theoretical findings, with particular emphasis on the construction of valid bootstrap procedures for calculating the impact of parameter estimation error. We then discuss the Corradi and Swanson (CS: 2002) test of (non)linear out-of-sample Granger causality. Thereafter, we carry out a series of Monte Carlo experiments examining the properties of the CS and a variety of other related predictive accuracy and model selection type tests. Finally, we present the results of an empirical investigation of the marginal predictive content of money for income, in the spirit of Stock and Watson (1989), Swanson (1998) and Amato and Swanson (2001).

JEL classification: C22, C51.
Keywords: block bootstrap, forecasting, recursive estimation scheme, rolling estimation scheme, model misspecification, nonlinear causality, parameter estimation error, prediction.

---

* corresponding author: Norman R. Swanson

Nii Ayi Armah and Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA (armah@econ.rutgers.edu and nswanson@econ.rutgers.edu). We wish to thank the editor, Mark Wohar; an anonymous referee; and Valentina Corradi, Jean-Marie Dufour, Silvia Goncalves, Stephen Gordon, Clive Granger, Oliver Linton, Brendan McCabe, Antonio Mele, Andrew Patton, Rodney Strachan, Christian Schluter, Allan Timmerman, and seminar participants at Cornell University, the London School of Economics, Laval University, Queen Mary, University of London, CIREQ-Universite' de Montreal, the University of Liverpool, Southampton University, the University of Virginia, the 2004 Winter Meetings of the Econometric Society, and the Bank of Canada for useful comments and suggestions on this research topic. Additionally, Swanson thanks the Rutgers University Research Council for financial support.

1 Introduction

In a series of recent papers, Chao *et al* (2001) and Corradi and Swanson (2002, 2004, 2006a, 2007) discuss model selection and predictive accuracy tests in the context of parameter and model uncertainty under recursive and rolling estimation schemes. In this chapter, we begin by summarizing some of the theoretical findings of these papers, with particular emphasis on the construction of valid bootstrap procedures for calculating the impact of parameter estimation error on the class of test statistics with limiting distributions that are functionals of Gaussian processes with covariance kernels that are dependent upon parameter and model uncertainty. We then provide an example of a particular test which falls in this class. Namely, we outline the so-called Corradi and Swanson (CS: 2002) test of (non)linear out-of-sample Granger causality. Thereafter, we carry out a series of Monte Carlo experiments examining the properties of the CS and a variety of other related predictive accuracy and model selection type tests, including the Deibold and Mariano (DM: 1995) and West (1996) predictive accuracy test as well as the encompassing test of Clark and McCracken (CM: 2004). This is done for both recursive and rolling window estimators, hence shedding light on the finite sample impact of using shorter rolling windows rather than recursive windows. Finally, we present the results of an empirical investigation of the marginal predictive content of money for income, in the spirit of Stock and Watson (1989), Swanson (1998), Amato and Swanson (2001), and the references cited therein. The empirical results shed new light on the importance of sample periods and estimation schemes when carrying out empirical investigations.

The main link between this chapter and the overall theme of the book is that we address the issue of model uncertainty. In particular, the tests discussed herein *do not* assume correct specification under either the null or the alternative hypothesis being tested. This is a crucial assumption to have if one believes that all models are approximations of some underlying *true* DGP. Of course, if one does not believe that all models should be viewed as approximations, then there is perhaps really no obvious need to carry out ex ante inference using forecasts (assuming no structural breaks). After all, under the assumption of correct specification under the null, why not simply carry out in-sample inference, for the sake of efficiency? Our approach differs from approaches used in many (perhaps most) currently popular prediction tests, where correct specification is assumed under the null. As a case in point, consider the predictive density testing framework discussed by the important paper of Diebold, Gunther and Tay (DGT: 1998) and in Corradi and Swanson (2006a,b,c). In

their paper, DGT use the probability integral transform (see e.g. Rosenblatt (1952)) to show that $F_t(y_t|\Im_{t-1}, \theta_0)$, is identically and independently distributed as a uniform random variable on $[0, 1]$, where $F_t(\cdot|\Im_{t-1}, \theta_0)$ is a parametric distribution with underlying parameter θ_0, y_t is the random variable of interest, and $\Im_{t-1}$ is the information set containing all "relevant" past information (see below for further discussion). They thus suggest using the difference between the empirical distribution of $F_t(y_t|\Im_{t-1}, \widehat{\theta}_T)$ and the $45°-$degree line as a measure of "goodness of fit", where $\widehat{\theta}_T$ is some estimator of θ_0. This approach has been shown to be very useful for financial risk management (see e.g. Diebold, Hahn and Tay (1998)), as well as for macroeconomic forecasting (see e.g. Diebold, Tay and Wallis (1998) and Clements and Smith (2000, 2002)). Likewise, Bai (2003) proposes a Kolmogorov type test of $F_t(u|\Im_{t-1}, \theta_0)$ based on the comparison of $F_t(y_t|\Im_{t-1}, \widehat{\theta}_T)$ with the CDF of a uniform on $[0, 1]$. As a consequence of using estimated parameters, the limiting distribution of his test reflects the contribution of parameter estimation error and is not nuisance parameter free. To overcome this problem, Bai (2003) uses a novel approach based on a martingalization argument to construct a modified Kolmogorov test which has a nuisance parameter free limiting distribution. This test has power against violations of uniformity but not against violations of independence. Now, Corradi and Swanson (2006b), allow for (dynamic) misspecification under the null hypothesis, while the others mentioned above do not. This feature allows them to obtain asymptotically valid critical values even when the conditioning information set does not contain all of the relevant past history. More precisely, if one is interested in testing for correct specification, given a particular information set which may or may not contain all of the relevant past information, then the Corradi-Swanson approach is preferrable. This is relevant when a Kolmogorov test is constructed, for example, as one is generally faced with the problem of defining $\Im_{t-1}$. If enough history is not included, then there may be dynamic misspecification. Additionally, finding out how much information (e.g. how many lags) to include may involve pre-testing, hence leading to a form of sequential test bias. By allowing for dynamic misspecification, one does not require such pre-testing. Another key feature of the Corradi-Swanson approach concerns the fact that the limiting distribution of Kolmogorov type tests is affected by dynamic misspecification. Critical values derived under correct specification given $\Im_{t-1}$ are not in general valid in the case of correct specification given a subset of $\Im_{t-1}$. Consider the following example. Assume that we are interested in testing whether the conditional distribution of $y_t|y_{t-1}$ is $N(\alpha_1^{\dagger} y_{t-1}, \sigma_1)$. Suppose also that in actual fact the "relevant" information set has $\Im_{t-1}$ including both y_{t-1} and y_{t-2}, so that the true

conditional model is $y_t|\Im_{t-1} = y_t|y_{t-1}, y_{t-2} = N(\alpha_1 y_{t-1} + \alpha_2 y_{t-2}, \sigma_2)$, where $\alpha_1^\dagger$ differs from α_1. In this case, we have correct specification with respect to the information contained in y_{t-1}; but we have dynamic misspecification with respect to y_{t-1}, y_{t-2}. Even without taking account of parameter estimation error, the critical values obtained assuming correct dynamic specification are invalid, thus leading to invalid inference. Stated differently, tests that are designed to have power against both uniformity and independence violations (i.e. tests that assume correct dynamic specification under H_0) will reject; an inference which is incorrect, at least in the sense that the "normality" assumption is *not* false. In summary, if one is interested in the particular problem of testing for correct specification for a given information set, then the Corradi-Swanson approach is appropriate. In general, these sorts of arguments apply to all varieties of prediction based testing, such as that discussed in this chapter.[1]

Parameter estimation error is a crucial component of model selection and predictive accuracy tests that is often overlooked, or more precisely is often assumed away by making the assumption that the in-sample estimation period grows more quickly than the out-of-sample predictive evaluation period. However, in some circumstances, such as when constructing DM tests for equal (pointwise) predictive accuracy of two models, limiting distributions are normal random variables, and parameter estimation error can be accounted for using the framework of West (1996). In other circumstances, such as when constructing tests which have power against generic alternatives (e.g. the CS test), statistics have limiting distributions that can be shown to be functionals of Gaussian processes with covariance kernels that reflect both (dynamic) misspecification as well as the contribution of parameter estimation error. Such limiting distributions are not nuisance parameter free, and critical values cannot be tabulated. Nevertheless, valid asymptotic critical values can be obtained via use of a bootstrap procedure that allows for the formulation of statistics which properly mimic the contribution of parameter estimation error. In the first part of the chapter we summarize block bootstrap procedures which are valid for recursive and rolling m-estimators (see e.g. Corradi and Swanson (2006a, 2007)).

In the second part of the chapter we review the so-called CS test, which is an out-of-sample version of the integrated conditional moment (ICM) test of Bierens (1982, 1990) and Bierens and

[1]Note that we do not address structural breaks directly, although lack of knowledge of structural breaks when specifying a model can clearly lead to misspecification under both hypotheses. This is one reason why rolling windows are sometimes used in predictive contexts.

Ploberger (1997), and which yields out-of-sample tests that are consistent against generic (non-linear) alternatives (see Corradi and Swanson (2002, 2007) and Swanson and White (1997)). The CS test can alternatively be viewed as a consistent specification test, in the spirit of Bierens, or as a nonlinear Granger causality test, as discussed in Chao *et al.* (2001). Note, however, that the CS test differs from the ICM test developed by Bierens (1982, 1990) and Bierens and Ploberger (1997) because parameters are estimated in either recursive or rolling fashion, the test is of the out-of-sample variety, and the null hypothesis is that the reference model delivers the best "loss function specific" predictor, for a given information set. Furthermore, the CS test allows for model misspecification under both hypotheses (see Corradi and Swanson (2006b)).

In order to provide evidence on the usefulness of the bootstrap methods discussed above, and in particular in order to compare bootstraps based on recursive and rolling estimators, we carry out a Monte Carlo investigation that compares the finite sample properties of our block bootstrap procedures with two alternative naive block bootstraps, all within the context of the CS test and a simpler non-generic version of the CS test due to Chao, Corradi and Swanson (CCS: 2001). In addition, various other related tests, including the standard F-test, the DM test and the CM test are included in the experiments. Results support the finding of Corradi and Swanson (2007) that the recursive block bootstrap outperforms alternative naive nonparametric block bootstraps. Additionally, we find that the rolling version of the bootstrap also outperforms the naive alternatives, Finally, we find that the finite sample properties of the other tests vary to some degree. Of note is that the Kilian (1999) bootstrap is a viable alternative to ours, although theoretical assessment thereof remains to be done (see Corradi and Swanson (2007) for further discussion).

In the last part of the chapter, an empirical illustration is presented, in which it is found that results concerning the (non)linear marginal predictive content for money for output are not only sample dependent, but also vary to some limited degree depending upon whether recursive or rolling estimation windows are used. In particular, there is evidence of predictive causation when in-sample estimation periods are ended any time during the 1980s, but little evidence of causality otherwise. Furthermore, recursive estimation windows yield better models when prediction periods begin in the 1980s, while rolling estimation windows yield better models when prediction periods begin during the 1970s and 1990s. Interestingly, these two results can be combined into a coherent picture of what is driving our empirical results. Namely, when recursive estimation windows yield lower overall predictive MSEs, then bigger prediction models that include money are preferred, while

smaller models without money are preferred when rolling models yield the lowest MSE predictors.

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. Note that P is also used to denote the length of the prediction period, and unless otherwise obvious from the context in which it is used, clarification of the meaning is given.

2 Block Bootstraps for Recursive and Rolling $m-$Estimators

In this section, we draw largely from Corradi and Swanson (2006a, 2007).

Recursive Estimation Window:

Define the block bootstrap estimator that captures the effect of parameter estimation error in the context of *recursive* m-estimators, 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]

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

Further, define

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

where q_i denotes the objective function for model i. As the discussion below does not depend on any specific model, we drop the subscript i. Following standard practice (such as in the real-time

[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.

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), 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 be used to construct sequences of P 1-step ahead forecasts and associated forecast errors, for example.

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

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

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

By first order conditions, $\frac{1}{t}\sum_{j=s}^{t} \nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t^*) = 0$, where ∇_θ denotes the denotes the derivative with respect to θ. Via a mean value expansion of $\frac{1}{t}\sum_{j=s}^{t} \nabla_\theta q(y_j^*, Z^{*,j-1}, \widehat{\theta}_t^*)$ around $\widehat{\theta}_t$, after a few simple manipulations, we have that

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

where $B^\dagger = E\left(-\nabla_\theta^2 q(y_j, Z^{j-1}, \theta^\dagger) \right)^{-1}$, $a_{R,j} = \frac{1}{R+j} + \frac{1}{R+j+1} + ... + \frac{1}{R+P-1}$, $j = 0, 1, ..., P-1$, and where the last equality on the right hand side of (4) follows immediately, using the same arguments

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

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

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

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

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)). Now, $\frac{1}{T-s} \sum_{k=s}^{T-1} \nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t) \neq 0$, and is instead of order $O_P\left(T^{-1/2}\right)$. Thus, $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \frac{1}{T-s} \sum_{k=s}^{T-1} \nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_t) = O_P(1)$, and does not vanish in probability. This clearly contrasts with the full sample case, in which $\frac{1}{T-s} \sum_{k=s}^{T-1} \nabla_\theta q(y_k, Z^{k-1}, \widehat{\theta}_T) = 0$, because of the first order conditions. Thus, $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(\widehat{\theta}_t^* - \widehat{\theta}_t \right)$ cannot have a zero mean normal limiting distribution, but is instead characterized by a location bias that can be either positive or negative depending on the sample.

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

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

$R \leq t \leq T-1$.[3]

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

[3] More precisely, we should use $\frac{1}{t-s}$ and $\frac{1}{T-s}$ to scale the summand in (7). For notational simplicity, $\frac{1}{t-s}$ and $\frac{1}{T-s}$ are approximated with $\frac{1}{t}$ and $\frac{1}{T}$.

nipulations, we have that:

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

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

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

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

Assumption A2: (i) $\theta^\dagger$ is uniquely identified (i.e. $E(q(y_t, Z^{t-1}, \theta)) > E(q(y_t, Z^{t-1}, \theta^\dagger))$ for any $\theta \neq \theta^\dagger$); (ii) q is twice continuously differentiable on the interior of Θ, and for Θ a compact subset of $\Re^\varrho$; (iii) the elements of $\nabla_\theta q$ and $\nabla_\theta^2 q$ are $p-$dominated on Θ, with $p > 2(2+\psi)$, where ψ is the same positive constant as defined in Assumption A1; and (iv) $E\left(-\nabla_\theta^2 q(\theta)\right)$ is negative definite uniformly on Θ.[4]

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

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

Theorem 1 (Corradi and Swanson (2007)): Under recursive estimation, let A1-A3 hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

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

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

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

Theorem 1 states that $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widetilde{\theta}_t^* - \widehat{\theta}_t\right)$ has the same limiting distribution as $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$, conditional on sample, and for all samples except a set with probability measure approaching zero. Of note is that if Assumption 3 is violated and $P/R \to 0$, then the statement in the theorem above is trivially satisfied, in the sense that both $\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widetilde{\theta}_t^* - \widehat{\theta}_t\right)$ and $\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widehat{\theta}_t - \theta^\dagger\right)$ have a limiting distribution degenerate on zero. Hence, the crucial impact of allowing for non-vanishing parameter estimation error is quite apparent.

Rolling Estimation Window:

In the rolling estimation scheme, one constructs a sequence of P estimators using a rolling window of R observations. The first estimator is constructed using the first R observations, the second using observations from 2 to $R+1$, and so on, with the last estimator being constructed using observations from $T-R$ to $T-1$, so that we have a sequence of P estimators, $(\widehat{\theta}_{R,R}, \widehat{\theta}_{R+1,R}, \ldots, \widehat{\theta}_{R+P-1,R})$. In general, it is common to assume that P and R grow as T grows. Giacomini and White (2003) propose using a rolling scheme with a fixed window that does not increase with the sample size, so that estimated parameters are treated as mixing variables. Pesaran and Timmerman (2004a,b) suggest rules for choosing the window of observations in order to take into account possible structure breaks.

Using the same notation as in the recursive case, but noting that we are now constructing a rolling estimator, define

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

In the case of in-sample model evaluation, the contribution of parameter estimation error is summarized by the limiting distribution of $\sqrt{T}(\widehat{\theta}_T - \theta^\dagger)$, where $\theta^\dagger$ is the probability limit of $\widehat{\theta}_T$. In the case of rolling estimation schemes, the contribution of parameter estimation error is summarized by the limiting distribution of $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$. Under mild conditions, because of the central limit theorem, $\left(\widehat{\theta}_t - \theta^\dagger\right)$is $O_P(R^{-1/2})$. Thus, if P grows at a slower rate than R (i.e. if $P/R \to 0$, as $T \to \infty$), then $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$ is asymptotically negligible. In other words, if the in-sample portion of the data used for estimation is "much larger" than the out-of-sample portion of the data to be used for predictive accuracy testing and generally for model evaluation, then the contribution of parameter estimation error is asymptotically negligible.

In the rolling estimation scheme, observations in the middle are used more frequently than

observations at either the beginning or the end of the sample. As in the recursive case, this introduces a location bias to the usual block bootstrap, as under standard resampling with replacement, any block from the original sample has the same probability of being selected. Also, the bias term varies across samples and can be either positive or negative, depending on the specific sample. Our objective is thus to properly recenter the objective function in order to obtain a bootstrap rolling estimator, say $\widetilde{\theta}_t^*$, such that $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\widetilde{\theta}_t^* - \widehat{\theta}_t)$ has the same limiting distribution as $\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\widehat{\theta}_t - \theta^\dagger\right)$, conditionally on the sample. The approach and result are largely as outlined above. Namely, resample b overlapping blocks of length l from $W_t = (y_t, Z^{t-1})$, and form a bootstrap sample, as in the recursive case. Then, define the rolling bootstrap estimator as

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

As in the recursive case, the following theorem can be stated.

Theorem 2 (Corradi and Swanson (2006a)): Under rolling estimation, 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 $R \to \infty$,

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

3 The CS Test

As an example of the implementation of the recursive and rolling bootstrap discussed above, we summarize the CS test discussed in different forms in Chao *et al* (2001) as well as in Corradi and Swanson (2002, 2007). The test is presented in a framework that is directly applicable to the empirical investigation discussed in a subsequent section of the chapter.

As discussed in the introduction, the test draws 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, which is designed for comparing nested models, and which allows for dynamic misspecification of all models being evaluated. The CS test is an out-of-sample version of the ICM test, as discussed in the introduction of this paper. Alternative (non DM) tests for comparing the predictive ability of a fixed number of nested models have previously also been suggested. For example, Clark and McCracken (2001, 2004) propose encompassing tests for

comparing two nested models for one-step and multi-step ahead prediction, respectively. 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. One of the main differences 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.

The CS testing approach that will be used in the Monte Carlo and empirical sections of the chapter, assumes 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. The benchmark model is:

$$y_t = \theta_{1,1}^{\dagger} + \theta_{1,2}^{\dagger} y_{t-1} + \theta_{1,3}^{\dagger} z_{t-1} + u_{1,t}, \tag{8}$$

where $\theta_1^{\dagger} = (\theta_{1,1}^{\dagger}, \theta_{1,2}^{\dagger}, \theta_{1,3}^{\dagger})' = \arg\min_{\theta_1 \in \Theta_1} E(q_1(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1} - \theta_{1,3} z_{t-1}))$, $\theta_1 = (\theta_{1,1}, \theta_{1,2}, \theta_{1,3})'$, 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.[5] The generic alternative model is:

$$y_t = \theta_{2,1}^{\dagger}(\gamma) + \theta_{2,2}^{\dagger}(\gamma) y_{t-1} + \theta_{2,3}^{\dagger}(\gamma) z_{t-1} + \theta_{2,4}^{\dagger}(\gamma) w(Z^{t-1}, \gamma) + u_{2,t}(\gamma), \tag{9}$$

where $\theta_2^{\dagger}(\gamma) = (\theta_{2,1}^{\dagger}(\gamma), \theta_{2,2}^{\dagger}(\gamma), \theta_{2,3}^{\dagger}(\gamma), \theta_{2,4}^{\dagger}(\gamma))' = \arg\min_{\theta_2 \in \Theta_2} E(q_1(y_t - \theta_{2,1} - \theta_{2,2} y_{t-1} - \theta_{2,3} z_{t-1} - \theta_{2,4} w(Z^{t-1}, \gamma)))$, $\theta_2(\gamma) = (\theta_{2,1}(\gamma), \theta_{2,2}(\gamma), \theta_{2,3}(\gamma), \theta_{2,4}(\gamma))'$, $\theta_2 \in \Theta_2$, 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. In fact, the above setup can be described within the context of our empirical example in Section 5. Namely, in Section 5 we set X_t is equal to a vector of two variables including money supply growth and a cointegration term connecting output, money and prices; y_t is set equal to output growth; and z_t is an interest rate spread. Turning back

[5] Note that z_{t-1} as used in (8) differs from Z^{t-1} used elsewhere in the chapter (see Section 5 for an empirical illustration where z_{t-1} is defined.

to our current discussion, note that the hypotheses of interest are:

$$H_0 : E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) = 0 \text{ versus } H_A : E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) > 0. \tag{10}$$

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.[6] For this reason, H_0 corresponds to equal predictive accuracy, while H_A corresponds to the case where the alternative model outperforms the reference model, as long as the errors above are loss function specific forecast errors. As discussed in Corradi and Swanson (2002), we can restate H_0 and H_A as:

$$H_0 : E(g'(u_{1,t+1})w(Z^t, \gamma)) = 0 \text{ versus } H_A : E(g'(u_{1,t+1})w(Z^t, \gamma)) \neq 0, \tag{11}$$

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

$$M_P = \int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma, \tag{12}$$

where

$$m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} g'(\widehat{u}_{1,t+1}) w(Z^t, \gamma), \tag{13}$$

and where $\int_\Gamma \phi(\gamma) d\gamma = 1$, $\phi(\gamma) \geq 0$, with $\phi(\gamma)$ absolutely continuous with respect to Lebesgue measure. Note also that " $'$ " denotes derivative with respect to the argument of the function. Elsewhere, we use "∇_x" to denote derivative with respect to x. In the sequel, we require the following assumptions.

Assumption A4: (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 $q'_{1,t}(\theta_1) = q'_1(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1} - \theta_{1,3} z_{t-1})$, Γ a compact subset of $\Re^d$, 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.

[6] Needless to say, in finite samples the forecasting mean square prediction error from the small model can be lower than that associated with the larger model.

Theorem 3 (Corradi and Swanson (2007)): Under either recursive or rolling estimation, let Assumptions A1-A4 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 (13) 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 $\varepsilon > 0$, $\lim_{P \to \infty} \Pr\left(\frac{1}{P} \int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma > \varepsilon \right) = 1$.

Clearly, the form of the covariance kernel depends upon whether recursive or rolling estimation is used (for further detailed discussion of these covariance kernels, the reader is referred to the appendices in Corradi and Swanson (2006a, 2007)). It is also clear that 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 or rolling m-estimators discussed above, we define:

$$\begin{aligned} \widetilde{\theta}^*_{1,t} &= (\widetilde{\theta}^*_{1,1,t}, \widetilde{\theta}^*_{1,2,t}, \widetilde{\theta}^*_{1,3,t})' = \arg \min_{\theta_1 \in \Theta_1} \frac{1}{t} \sum_{j=2}^{t} [q_1(y^*_j - \theta_{1,1} - \theta_{1,2} y^*_{j-1} - \theta_{1,3} z^*_{j-1}) \\ &\quad - \theta'_1 \frac{1}{T} \sum_{i=2}^{T-1} \nabla_\theta q_1(y_i - \widehat{\theta}_{1,1,t} - \widehat{\theta}_{1,2,t} y_{i-1} - \widehat{\theta}_{1,3,t} z_{i-1})] \end{aligned} \tag{14}$$

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

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

where, recalling that $g = q_1$,

$$\begin{aligned} & m_P^*(\gamma) \\ = \; & \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g' \left(y_{t+1}^* - \begin{pmatrix} 1 & y_t^* & z_t^* \end{pmatrix} \widetilde{\theta}_{1,t}^* \right) w(Z^{*,t}, \gamma) - \frac{1}{T} \sum_{i=2}^{T-1} g' \left(y_i - \begin{pmatrix} 1 & y_{i-1} & z_{i-1} \end{pmatrix} \widehat{\theta}_{1,t} \right) w(Z^{i-1}, \gamma) \right) \end{aligned} \tag{15}$$

The bootstrap statistic in (15) 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. This differs from the usual approach that would involve calculating:

$$m_P^{**}(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \left(g' \left(y_{t+1}^* - \begin{pmatrix} 1 & y_t^* & z_t^* \end{pmatrix} \widetilde{\theta}_{1,t}^* \right) w(Z^{*,t}, \gamma) - g' \left(y_{t+1} - \begin{pmatrix} 1 & y_t & z_t \end{pmatrix} \widehat{\theta}_{1,t} \right) w(Z^t, \gamma) \right) \tag{16}$$

However, the recursive (rolling) feature of the parameter estimation error in the CS test in the current context ensures that for all samples except a set with probability measure approaching zero, $m_P^{**}(\gamma)$ does not have the same limiting distribution as $m_P(\gamma)$ (see Corradi and Swanson (2007) for further details).

Theorem 4 (Corradi and Swanson (2007)): Under either recursive or rolling estimation, let Assumptions A1-A3 and A5 hold. Also, assume that as $T \to \infty$, $l \to \infty$, and that $\frac{l}{T^{1/4}} \to 0$. Then, as T, P and $R \to \infty$,

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

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 in the following manner. For any bootstrap replication, compute the bootstrap statistic, $m_P^*(\gamma)$. Perform B bootstrap replications (B large) and compute the quantiles of the empirical distribution of the B bootstrap statistics. Reject H_0, if $m_P(\gamma)$ is greater than the $(1-\alpha)th$-percentile. Otherwise, do not reject.

4 Monte Carlo Experiments

In this section we carry out a series of Monte Carlo experiments comparing the recursive and rolling block bootstrap with a variety of other bootstraps, and comparing the finite sample performance of

the test discussed above with a variety of other tests. In addition to the fact that rolling as well as recursive estimators are used, the experiments in this section differ from those discussed in Corradi and Swanson because they estimate an AR(1) model as their benchmark model (i.e. the model used in size experiments), while our benchmark model includes an additional explanatory variable, z_t, which corresponds to the interest rate spread in our empirical implementation. Furthermore, they include in all models an omitted variable, which we do not use in our specifications. As shall be discussed below, it is in fact this omitted variable that drives much of the size distortion in Corradi and Swanson (2007) when comparing the F-test with various other tests.

With regard to the bootstrap, we consider 4 alternatives. Namely: (i) the "Recur Block Bootstrap", which is the block bootstrap for recursive m-estimators discussed above; (ii) the "Roll Block Bootstrap", which is also discussed above, (iii) the "Block Bootstrap, no PEE, no adjust", which is a strawman block bootstrap used for comparison purposes, where it is assumed that there is no parameter estimation error (PEE), so that $\widehat{\theta}_{1,t}$ is used in place of $\widetilde{\theta}^*_{1,t}$ in the construction of M^*_P, and the term $\frac{1}{T}\sum_{i=1}^{T-1} g'\left(y_{i+1} - \begin{pmatrix} 1 & y_i & z_i \end{pmatrix}\widehat{\theta}_{1,t}\right) w(Z^i,\gamma)$ in m^*_P is replaced with $g'\left(y_{t+1} - \begin{pmatrix} 1 & y_t & z_t \end{pmatrix}\widehat{\theta}_{1,t}\right) w(Z^t,\gamma)$ (i.e. there is no bootstrap statistic adjustment, thus conforming with the usual case when the standard block bootstrap is used) and (iv) the "Standard Block Bootstrap", which is the standard block bootstrap (i.e. this bootstrap is the same as that outlined in (iii), except that $\widehat{\theta}_{1,t}$ is replaced with $\widehat{\theta}^*_{1,t}$.

As discussed in Section 3, the hypotheses of interest are:

$$H_0 : E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) = 0 \text{ versus } H_A : E(g(u_{1,t+1}) - g(u_{2,t+1}(\gamma))) > 0. \tag{17}$$

where $u_{1,t}$ and $u_{2,t}$ are out-of-sample 1-step ahead prediction errors of the following models:

$$y_t = \theta^\dagger_{1,1} + \theta^\dagger_{1,2} y_{t-1} + \theta^\dagger_{1,3} z_{t-1} + u_{1,t}, \tag{18}$$

$$y_t = \theta^\dagger_{2,1}(\gamma) + \theta^\dagger_{2,2}(\gamma) y_{t-1} + \theta^\dagger_{2,3}(\gamma) z_{t-1} + \theta^\dagger_{2,4}(\gamma) w(Z^{t-1},\gamma) + u_{2,t}(\gamma), \tag{19}$$

where $\theta^\dagger_1 = (\theta^\dagger_{1,1}, \theta^\dagger_{1,2}, \theta^\dagger_{1,3})'$, and $\theta^\dagger_2 = (\theta^\dagger_{2,1}, \theta^\dagger_{2,2}, \theta^\dagger_{3,3}, \theta^\dagger_{3,4})'$ are parameter vectors, where $z_{t-1}$is an additional explanatory variable in the "small" model, and where Z^{t-1} in the "big" model includes the variable which is being tested for inclusion in the small model (denoted x_t in Table 1).

The test statistics examined in our experiments include: (i) the standard in-sample F-test; (ii) the encompassing test due to Clark and McCracken (CM: 2004) and Harvey *et al* (1997); (iii) the

Diebold and Mariano (DM: 1995) test; (iv) the CS test; and (v) the CCS test.[7] Of note in this context is that in the CS test we are implicitly testing whether any (non)linear function of Z^{t-1} would be useful for constructing a better prediction model of y_t. Alternatively, the other tests only consider inclusion of a linear function of Z^{t-1}, so that they are essentially setting w to be an affine function.

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

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

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

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

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

[7]The CCS statistic is essentially the same as the CS test, but uses Z^t instead of a generically comprehensive function thereof (recall that Z^t contains the additional variables included in the "big" model defined below). Thus, this test can be seen as a special case of the CS test that is designed to have power against linear alternatives, and it is not explicitely designed to have power against generic nonlinear alternatives as is the CS test. The theory in Section 3 of this paper thus applies to both the CS and CCS tests. Additionally, the CM test is included in our study because it is an encompassing test which is designed to have power against linear alternatives, and so it is directly comparable with the CCS test. Finally, the F and DM tests are included in our analysis because they are the most commonly applied and examined in- and out-of-sample tests used for model selection. They thus serve as a kind of benchmark against which the performance of the other tests can be measured.

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

Data are generated according to the DGPs summarized in Table 1 as : *Size1-Size2* and *Power1-Power12*.

In our setup, the benchmark model (denoted by *Size1 in Table 1*) is an ARX(1). (The benchmark model is also called the "small" model.) The null hypothesis is that no competing model outperforms the benchmark model. Twelve of our DGPs (denoted by *Power1-Power12*) include (non)linear functions of x_{t-1}. In this sense, our focus is on (non)linear out-of-sample Granger causality testing. Some regression models estimated in these experiments are misspecified not just because of neglected nonlinearity, but also because fitted regression functions ignore the MA error component that appears in some DGPs. Recall also, as discussed above, that CS and CCS tests only require estimation of the benchmark models. The CM, F, and DM tests require estimation of the benchmark models as well as the alternative models. In our context, the alternative model estimated is simply the benchmark model with x_{t-1} added as an additional regressor, regardless of which DGP is used to generate the data. The alternative is also sometimes called the "big" model.

The functional forms that are specified under the alternative include: (i) exponential (*Power1, Power7*); (ii) linear (*Power2*); (iii) self exciting threshold (*Power3*), squared (*Power8*) and absolute value (*Power9*). In addition, *Power4-Power6* and *Power10-Power12* are the same as the others, except that an MA(1) term is added. Notice that *Power1* includes a nonlinear term that is similar in form to the test function, $w(\cdot)$, which is defined below. Also, *Power2* serves as a linear causality benchmark. Test statistics are constructed by fitting what is referred to in the next section as a "small model" in order to construct the CS and CCS test statistics. Note that the "big model" (which is a linear ARX(1) model in y_{t-1}, and z_{t-1} with x_{t-1} added as an additional regressor) is only fitted in order to construct the F, CM, and DM test statistics. It is not necessary to fit this model when constructing the CS and CCS statistics. All test statistics are formed using one-

step ahead predictions (and corresponding prediction errors) from recursive and rolling window estimated models.

In all experiments, we set $w(z^{t-1},\gamma) = \exp(\sum_{i=1}^{3}(\gamma_i \tan^{-1}((z_{i,t-1} - \overline{z}_i)/2\widehat{\sigma}_{z_i})))$, with $z_{1,t-1} = x_{t-1}$, $z_{2,t-1} = y_{t-1}$, $z_{3,t-1} = w_{t-1}$ and $\gamma_1, \gamma_2, \gamma_3$ scalars. Additionally, define $\Gamma = [0.0, 5.0] \times [0.0, 5.0] \times [0.0, 5.0]$. We consider a grid that is delineated by increments of size 0.5. All results are based on 500 Monte Carlo replications, and a sample of T=540 is used. All tests are empirical rejection frequencies. The following parameterizations are used: $a_1 = 1.0$, $a_2 = \{0.3, 0.6, 0.9\}$, and $a_3 = 0.3$. Additionally, bootstrap critical values are constructed using 100 simulated statistics, the block length, l, is set equal to $\{2, 5, 10\}$, $\{4, 10, 20\}$, or $\{10, 20, 50\}$, depending upon the degree of DGP persistence, as given by the value of a_2. Finally, all results are based on $P = (1/2)T$ recursive and rolling window formed predictions.

We summarize our findings from the Monte Carlo simulations in Tables 2-3 for the CS test and Tables 4-5 for the F, DM, CM and CCS tests. In addition, Tables 2 and 4 consider results under recursive estimation, while Tables 3 and 5 consider results under rolling window estimation. The first column in the mentioned tables states the DGP used to generate the data. The names are further defined in Table 1. *Size1-Size2* refer to empirical size experiments and *Power1-Power12* refer to empirical power experiments. All numerical entries are test rejection frequencies. Details of the mnemonics used to describe the columns in the tables and the different approaches used for critical value construction are contained in the footnotes to Table 2 and 4.

In the following discussion, we consider two broad issues. First, is the recursive/rolling bootstrap useful, or could one simply use more naive bootstraps such as the standard block bootstrap? Second, what can we say about the use of recursive as opposed to rolling window estimation schemes for estimating model parameters and in particular with respect to inference. As an ancilliary issue, we also consider the issue of in-sample versus out-of-sample testing since we include the in-sample F-test as an alternative test.

A first look at Tables 2 and 3, where the CS test is examined under the "Recur/Rolling Block Bootstrap" indicates that in general, empirical levels are larger and closer to the 10% nominal level under recursive estimation (Table 2) than under rolling window estimation (Table 3). For example, in Panel A of Tables 2 and 3, empirical rejection levels for $l = 2, 5, 10$ are $0.07, 0.07, 0.08$ (Table 2) and $0.05, 0.06, 0.07$ (Table 3) for *Size1*. However, empirical power is in general closer to 1 under rolling window estimation (Table 3) than under recursive estimation (Table 2). For example, in

Panel A of Tables 2 and 3 empirical power for $l = 2, 5, 10$ is $0.53, 0.73$,and 0.80 (Table 2) and $0.62, 0.87$, and 0.90 (Table 3) for *Power1*. This observation about empirical power also holds for the other bootstrap techniques considered. These findings are not surprising, given the that the rolling windows are fixed in length, while the recursive windows increase in length. Furthermore, it is worth stressing that both window types appear to be yield quite reasonable finite sample properties, overall, when the nonparametric bootstrap is used. Finally, notice also that in all panels of Tables 2 and 3, CS tests constructed using data generated according to *Size2* yield poorer empirical level performance than under *Size1*. This is as expected, given that *Size2* DGPs include unmodelled serial error dependence.

A closer look at Table 2 reveals that regardless of the level of dependence in the lagged endogenous variable as determined by the value of a_2, the nonparametric block bootstrap developed in this paper consistently has the empirical level closest to the nominal level. For example in Table 2, the closest empirical level to the nominal level is 0.08 and it occurs in Panel A when under "Recur Block Bootstrap" and *Size1* when $l = 10$. This same observation can be made in Table 3. However, such a blanket conclusion cannot be drawn when comparing empirical power. In Panels A and B of Table 2, for the smallest block lengths of 2 and 4 respectively, the "Block Bootstrap" in general has the highest power levels. For the medium block lengths of 5 and 10 of Panels A and B respectively, the "BB, no PEE, no adj" nonparametric bootstrap has higher power. Finally, for the highest block length, "Recur Block Bootstrap" has the highest empirical power. When there is too much persistence in the model as in Panel C, these conclusions no longer hold. The same conclusions can generally be drawn under the rolling window estimation in Table 3.

We now turn to a discussion of Tables 4 and 5, where results for the rest of the test statistics examined in the Monte Carlo experiments are reported. Relative to the Monte Carlo results in CS (2007), the F-test is not nearly as severely oversized. Indeed, judging from its empirical level and power figures, the F-test seems to have good size and power properties. The main reason for this is that the F-test is in-sample, and is carried out with a correctly specified model in the current analysis. Of course, an in-sample analysis of a correctly specified model for any test will generally yield superior performance. However, as shown in CS (2007), where there is model misspecification in the form of an omitted variable, the in-sample F-test is highly oversized. It is in such cases (i.e. model misspecification) that the argument can be made for considering alternative tests of model performance, even under an assumption of linearity, and particularly when nonlinearities may be

present in the true underlying DGPs.

In addition to this, in both Tables 4 and 5, there is a dramatic improvement in the empirical size of the DM test depending upon which critical values are used (i.e. whether we assume that $\pi = 0$ or $\pi > 0$ - see footnote to Table 4 for further explanation of π). The empirical size under $\pi > 0$ is much closer to the nominal size of 10%. This suggests that parameter estimation error is relevant in our setup, as standard normal critical values (under $\pi = 0$) are simply too big. For the CM test in both Tables 4 and 5, the assumption that $\pi > 0$ still generates some improvement in empirical size values albeit marginal. Empirical power is very high for the F, CM and DM tests under either assumption on π; and unlike the CS test in Tables 2 and 3, power is not compromised by high persistence levels. This is however not the case for the CCS test. In Panel A and B of Table 4, the CCS test is grossly oversized regardless of block length. However, for both Tables 4 and 5, as the model becomes more persistent, there is an improvement in size and a reduction in power. The fact that this sort of result arises for the CS and CCS tests and not for the F, DM or CM tests indicates that the power loss is due to the use of a block length dependent bootstrap for calculating critical values. Indeed, it is worth noting that the power reduction is also characteristic of the other naive bootstrap techniques in Tables 2 and 3. Furthermore, it is worth noting that under model misspecification of the variety looked at in CS (2007), the F, CM and DM tests are no longer dominant in the above respect. In the next section, we estimate models that are clearly approximations to the true underlying DGP and hence are probably misspecified. We use the CS test which is robust to model misspecification under both hypotheses, as well as the other tests examined above, to assess the models.

5 Empirical Illustration: The Marginal Predictive Content of Money for Output

In this section we implement the F, CM, DM, CS and CCS tests that are described in Table 1, and examined in the previous Monte Carlo section. In particular, we use these tests together with recursive and rolling window estimation schemes to assess the marginal predictive content of money for real income. Recent contributions to this important literature include the papers of Swanson (1998), Amato and Swanson (2001), and the papers cited therein.

The variables used are the same as those examined by Christiano and Ljungqvist (1988), Stock

and Watson (1989), Friedman and Kuttner (1993) and Thoma (1994). In particular, the variables used are monthly observations of industrial production (IP), the wholesale price index (P), the secondary market rate on 90-day U.S. Treasury bills (R), the interest rate on three-month prime commercial paper (C) and Divisia monetary aggregates of money supply ($M2$). The sample period is 1959:01 to 2003:12. Seasonally adjusted nominal measures of $M2$ exhibit erratic behavior after 1985, which can be accounted for by documented shifts in the public's demand for money balances. This might explain why the relationship between nominal $M2$, IP and P has been unstable in recent years. Our approach in dealing with shifting money demand is to consider the Divisia monetary aggregates of $M2$. Other approaches, such as including structural breaks and explicit nonlinearities in the models are left to future research. All data with the exception of the three-month prime commercial paper (C) were obtained from the St. Louis Federal Reserve Bank. The data on C were obtained from Stock and Watson (2005).

We define the small model as a vector error correction model with:

$$y_t = \theta_{1,1}^\dagger + \theta_{1,2}^\dagger y_{t-1} + \theta_{1,3}^\dagger z_{1,t-1} + u_{1,t}$$

where

$\theta_1^\dagger = (\theta_{1,1}^\dagger, \theta_{1,2}^\dagger, \theta_{1,3}^\dagger)' = \arg\min_{\theta_1 \in \Theta_1} E(q_1(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1} - \theta_{1,3} z_{1,t-1}))$ is defined conformably,

$$y_t = (\Delta \log IP_t, \Delta \log P_t, \Delta R_t)'$$

and

$$z_{1,t-1} = C_{t-1} - R_{t-1}.$$

We further define the generic alternative (big) model as:

$$y_t = \theta_{2,1}^\dagger(\gamma) + \theta_{2,2}^\dagger(\gamma) y_{t-1} + \theta_{2,3}^\dagger(\gamma) z_{1,t-1} + \theta_{2,4}^\dagger(\gamma) w(Z^{t-1}, \gamma) + u_{2,t}(\gamma)$$

where

$$\theta_2^\dagger(\gamma) = (\theta_{2,1}^\dagger(\gamma), \theta_{2,2}^\dagger(\gamma), \theta_{2,3}^\dagger(\gamma), \theta_{2,4}^\dagger(\gamma))' = \arg\min_{\theta_2 \in \Theta_2} E(q_1(y_t - \theta_{2,1} - \theta_{2,2} y_{t-1} - \theta_{2,3} z_{1,t-1} - \theta_{2,4} w(Z^{t-1}, \gamma)))$$

and

$$\begin{aligned} y_t &= (\Delta \log IP_t, \Delta \log P_t, \Delta \log M2_t, \Delta R_t) \\ z_{1,t-1} &= C_{t-1} - R_{t-1} \\ z_{2,t-1} &= \log M2_{t-1} - \log IP_{t-1} - \log P_{t-1}. \end{aligned}$$

Finally, $Z^{t-1} = (z_{2,t-1}, \Delta \log M2_{t-1})$. Notice that $z_{1,t-1}$ and $z_{2,t-1}$ can be interpreted as vector error correction terms, and are consistent with evidence presented in Swanson (1998) and Amato and Swanson (2001). Since we are interested in examining the (non)linear marginal predictive content of money for income, our forecasting analysis and test statistics are constructed based on estimates of the first equation in the vector error correction model specified above (i.e. the equation with $\Delta \log IP_t$ as dependent variable).

Of note is that standard F-tests or Wald-tests for Granger causality are prone to severe upward size distortions when vector error correction (VEC) models are estimated using only differenced data, without accounting for cointegrating restrictions (see e.g. Swanson (1998) and Swanson, Ozyildirim and Pisu (2003)). One of the reasons why this problem arises is that the moving average representation for a model with cointegrated regressors will not yield a finite order VAR representation. In Swanson (1998) it is noted that at a 1% significance level, trace test statistics support the presence of one cointegrating (CI) vector when the data are linearly detrended, and when an intercept or an intercept and a trend are included in the cointegrating relation. One of the two cointegrating vectors is $z_{1,t-1}$, based on a likelihood ratio test (see Johansen (1988,1991)). Of further note is that the null hypothesis that the other CI vector is $z_{2,t-1}$ almost always fails to reject, although confidence intervals are quite wide relative to those for the interest rate spread CI vector. Finally, it should be recalled (see the discussion in Section 4) that in the DM, CM, CCS, and F tests, unlike the CS test, the alternative model is explicitly estimated. In such cases, linearity is assumed, so that the bigger model includes linear functions of $z_{2,t-1}$ and $\Delta \log M2_{t-1}$. This is one of the main reasons why it should not be expected that the results of the different empirical tests "agree". Indeed, if the CS test rejects while all others fail to reject, we have direct evidence of nonlinear Granger causality coupled with evidence of an absence of linear causality, for example.[8]

We construct tests statistics using 1-steap ahead forecasts formed via recursive and rolling window estimated models. Thus, models are re-estimated (using least squares) at each point in time, before each new prediction is constructed. The beginning date for the in-sample period is 1959:1 when constructing the CS, CCS, DM, CM, and F tests, the prediction periods reported on are 1978:1-2003:12 ($\pi = 1.4$), 1981:1-2003:12 ($\pi = 1.0$) and 1987:1-2003:12 ($\pi = 0.6$), so that initial

[8]Here, we are using the notion of "causality" interchangeably with the notion of prediction, in the spirit of what Granger originally had in mind when he introduced causality to the time series profesion (see the discussion in Chao, Corradi and Swanson (2001) for further details).

estimation samples for both the recursive and rolling window schemes include data for the periods 1959:1-1977:12, 1959:1-1980:12 and 1959:1-1986:12, respectively. The block length is set equal to 6 in application of the recursive block bootstrap.[9] In all cases, the dependent variable in regressions and the target variable in forecasts is the first log difference of industrial production (output). As discussed above, all estimated models are linear, and explanatory variables include lags of the first log difference of industrial production, prices, lag first difference of interest rates as well as the CI term $C_{t-1} - R_{t-1}$ (in the benchmark or "small" model). Lags of the first log difference of $M2$ and the CI term $z_{2,t-1}$ are added for the alternative ("big") model. Lags are selected via use of the Schwarz information criterion. Again as discussed above, and given this setup, our tests can be viewed as tests of (non) linear Granger causality.[10]

Results are gathered in Tables 6-7. In Table 6, point mean square forecast errors (MSEs) are tabulated for the "small model" and the "big model" under rolling window and recursive window estimation schemes respectively. Results are given not only for the three prediction periods outlined above, but also for all prediction periods beginning with 1974:1, 1975:1, ..., 1993:1. In Tables 7, CS, CCS, F, DM and CM test results for the three prediction periods outlined above are reported.

Turning first to the MSE results in Table 6, note that in the case of recursive estimation, the "big" model consistently outperforms the "small" model, for every prediction period. However, in many instances the MSEs are very close in absolute and relative magnitude, with differences often less than 1%. Interestingly, this pattern does not emerge when viewing MSEs associated with models estimated using rolling windows. In particular, the bigger model that includes money only "wins" for prediction periods beginning in 1984, 1988, 1989, 1990, and 1991. This puzzle is further confounded by noting that the lowest MSE model across both estimation window types is sometimes associated with the recursive modelling strategy, and sometimes with the rolling estimation strategy (note that the bold figures denote the lowest MSE across all estimation strategies and model types for a given start year). Thus, it appears that choice of recursive versus rolling estimation in our exercise is quite dependent upon sample prediction period start date.

[9] It should be noted that we do not use real-time data in this empirical illustration, even though both variables considered are subject to periodic revision. Extension of our results to incorporate real-time data is left to future research. Additionally, note that various other block lengths were tried and the empirical findings where qualitatively similar regardless of block length.

[10] It should be stressed that the results presented in this section are meant primarily to illustrate the uses of the different tests, and to underscore potentially important differences between the tests.

As mentioned above, the bigger model is always preferred for recursive estimation, while the results are mixed for rolling estimation. In particular, for rolling estimation, the bigger model is preferred for only 5 start years. If the recursively estimated models always yielded the lowest overall MSE across both estimation strategies, our results would be quite straightforward. However, when one looks across estimation strategies, the rolling window approach "wins" when prediction periods begin in the 1990s or from 1974-1982. The recursive window approach "wins" for prediction periods beginning from 1983-1989. This corresponds to our ranking of the models when one looks across *both* estimation strategies. Namely, the lowest MSE model is essentially the bigger model during much of the 1980s (i.e. from 1983 through 1991), while the smaller model "wins" during the rest of the years. Thus, for prediction periods that include the more turbulent 1970s, the smaller model wins, while for prediction periods beginning after 1983, the bigger model with money "wins". This corresponds loosely with the money targeting experiment of the early 1980s. Namely, after this targeting experiment ended, one might argue that a sufficiently "stable" environment ensued for money to become a predictor for output. This is rather interesting, given that the stated goal of the Federal Reserve Board has indeed been stabilization at low levels of inflation.

A further point of interest is that the rolling 10 year estimator that we used in our analysis is indeed dominant with regard to point MSE for 13 or the 20 start years (i.e. 13 of the 20 different prediction periods). Thus, we have some evidence that there may indeed be instabilities resulting in the relatively poorer performance of recursive estimation strategies. As might be expected, this points to model misspecification in the form of structural breaks, missing variables, and omitted nonlinearity, for example.

Finally, it is worth stressing that predictions of income have clearly gotten substantially more accurate over our sample period, as evidenced by the fact that MSFEs are much bigger for early subsamples, and are much smaller for the later sub-samples. This result is clearly due in part to the smooth nature of recent data relative to more distant data, although one might also argue that the more accurate results are associated in large part with instances where models that include money yield superior point predictions, hence pointing to further evidence in favor of using money in output prediction models. It should be stressed, however, that thus far we have only compared MSEs, and hence have focused our attention upon the comparison of purely linear models. In order to assess the potential impact of generic nonlinearity, for example, we need to either fit a variety of nonlinear models (which may be a large undertaking, given the plethora of available models), or

we need to carry out tests such as the generically comprehensive nonlinear out-of-sample Granger causality CS test. We turn to this issue next.

As mentioned above, Table 7 contains CS, CCS, F, DM and CM test results for prediction periods beginning in 1978, 1981, and 1987. Three conclusions emerge based upon inspection of the results. First, the CS test fails to reject the null of no (non)linear predictive causation, regardless of prediction period, and regardless of whether recursive or rolling estimation is used. On the other hand, there are many rejections of the null hypothesis when the "linear" tests are used, particularly at the 10% level. Furthermore, these rejections, in the case of recursive estimation, correspond to the big model winning (as the MSE associated with the big model is always lower than that associated with the small model). Thus, based on our recursive results, there is clearly predictive causation from money to output, However, this causation appears to be "moderate" in magnitude, given the fact that the non rejection using the CS test coupled with rejections using the CCS test may be a result of low power associated with the CS test (i.e. the CS test is an omnibus test, and hence has lower power in any given specific alternative than a test designed with that alternative specifically in mind). Second, the number of rejections is close to twice as many when moving from the rolling to the recursive estimation schemes, suggesting that parameter estimation error is playing a significant role in our testing procedures. This finding is also indicative of further evidence in favor of predictive causation, given that in the rolling case, small model MSEs based on prediction periods beginning in 1978, 1981, and 1987 are always lower than corresponding big model MSEs. In other words, in the rolling cases, rejection would imply that the big model is significantly "better" than the small model; and hence fewer rejections supports the finding based on the recursive estimation scheme that there is predictive causation. Third, when changing the significance level from 10% to 5%, some rejections in the CCS, DM and CM tests become non-rejections, which again substantiates the claim that although there is predictive causation, it is somewhat "weak" in the sense that predictions do not change to a great extent when money is added to the output equation.

In summary, power considerations are relevant, as should be expected, when using the CS test, as evidenced by the fact that in our illustration the CS test may be good at detecting non-linear Granger causality, but it is clearly not good at detecting moderate levels of linear predictive causation. Additionally, our evidence is clearly leaning toward a finding of predictive causation from money to output. However, much empirical work is needed before a complete picture emerges

concerning the prevalence of nonlinear Granger causality in the income/money relationship. This is left to future research. It is clear, though, that much can be learned by using *all* of the different tests in consort with one another.

6 Concluding Remarks

We have discussed bootstrap procedures valid for construction of critical values in the case of test statistics based on recursive and/or rolling estimation schemes that 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 discussed the Corradi and Swanson (2002: CS) model selection type test and carried out a series of experiments evaluating the CS as well as a variety of other tests including ones due to Diebold and Mariano (1995) and Clark and McCracken (2004). Finally, we carried out an empirical investigation using all of the tests examined in the Monte Carlo experiments. the investigation focused on predictive money-income causation. We found that sample size, prediction period, and estimator type (i.e. recursive versus rolling) play an important role in our empirical findings, although concrete evidence supporting the existence of predictive causation was found, particularly for prediction periods beginning during the 1980s.

7 References

Amato, J.D. and Swanson, N.R. (2001). The Real Time Predictive Content of Money for Output. *Journal of Monetary Economics* 48, 3-24.

Bai, J. (2003). Testing Parametric Conditional Distributions of Dynamic Models. *Review of Economics and Statistics* 85, 531-549.

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 Ploberger, W. (1997). Asymptotic Theory of Integrated Conditional Moment Tests. *Econometrica* 65, 1129-1152.

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

Christiano, L.J. and Ljungqvist, L. (1988). Money Does Granger-Cause Output in the Bivariate Money-Output Relation. *Journal of Monetary Economics* 22, 217-235.

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

Clark, T.E., and McCracken, M.W. (2005). Evaluating Direct Multistep Forecasts. *Econometric Reviews* 24, 369-404.

Clements, M.P. and Smith, J. (2000). Evaluating the Forecast Densities of Linear and Nonlinear Models: Applications to Output Growth and Unemployment. *Journal of Forecasting* 19, 255-276.

Clements, M.P. and Smith, J. (2002). Evaluating Multivariate Forecast Densities: A Comparison of Two Approaches. *International Journal of Forecasting* 18, 397-407.

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

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

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

Corradi, V. and Swanson, N.R. (2006a). Predictive Density Evaluation. In C. Granger, G.Elliot and A. Timmerman (Eds.) *Handbook of Economic Forecasting* (pp. 197-284). Amsterdam: Elsevier.

Corradi, V. and Swanson, N.R. (2006b). Bootstrap Conditional Distribution Tests In the Presence of Dynamic Misspecification. *Journal of Econometrics* 133, 779-806.

Corradi, V. and Swanson, N.R. (2006c). Predictive Density and Conditional Confidence Intervals Accuracy Tests. *Journal of Econometrics* 135, 187-228.

Corradi, V. and N.R. Swanson, N.R. (2007). Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes. *International Economic Review* forthcoming.

Diebold, F.X., T. Gunther and Tay, A.S. (1998). Evaluating Density Forecasts with Applications to Finance and Management. *International Economic Review* 39, 863-883.

Diebold, F.X., Hahn, J. and Tay, A.S. (1999). Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High Frequency Returns on Foreign Exchange. *Review of Economics and Statistics* 81, 661-673.

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

Diebold, F.X., A.S. Tay and Wallis, K.D. (1998). Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters, *in Festschrift in Honor of C.W.J. Granger*, eds. R.F. Engle and H. White, Oxford University Press, Oxford.

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

Friedman, B.M. and Kuttner, K.N. (1993). Another Look at the Evidence on Money-Income Causality. *Journal of Econometrics* 57, 189-203.

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

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

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

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

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

Johansen, S. (1988). Statistical Analysis of Cointegrating Vectors. *Journal of Economic Dynamics and Control 12, 231-254.*

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica 59, 1551-1580.*

Kilian, L. (1999). Exchange Rate and Monetary Fundamentals: What do we Learn from Long-Horizon Regressions. *Journal of Applied Econometrics* 14, 491-510.

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

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

Pesaran, M.H. and Timmerman, A. (2004a). How Costly is to Ignore Breaks when Forecasting the Direction of a Time Series? *International Journal of Forecasting* 20, 411-425.

Pesaran, M.H. and Timmerman, A. (2004b). Selection of Estimation Window for Strictly Exogenous Regressors. Working Paper, Cambridge University and University of California, San Diego.

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

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

Stock, J.H. and Watson, M.M. (1989). Interpreting the Evidence on Money-Income Causality. *Journal of Econometrics 40, 161-181.*

Stock, J.H. and Watson, M.M. (2005). Implications of Dynamic Factor Models for VAR Analysis. Working Paper, Princeton University and Harvard University.

Swanson, N.R. (1998). Money and Output Viewed Through a Rolling Window. *Journal of Monetary Economics* 41, 455-474.

Swanson, N.R., Ozyildirim, A. and Pisu, M.(2003). A Comparison of Alternative Causality and Predictive Ability Tests in the Presence of Integrated and Cointegrated Economic Variables. In D. Giles (Ed.), *Computer Aided Econometrics* (pp. 91-148). New York: Marcel Dekker.

Swanson, N.R. and White, H. (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.

Thoma, M.A. (1994). Subsample Instability and Asymmetries in Money-Income Causality. *Journal of Econometrics* 64, 279-306.

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.

Table 1: Test Statistics, Sampling Scheme, and Data Generating Processes Used in Monte Carlo Experiments

Panel A: Test Statistic Mnemonics and Definitions

F – The standard Wald version of the in-sample F-test is calculated using the entire sample of T observations. In particular, we use: $F = T\left(\sum_{t=1}^{T}\widehat{u}_{1,t}^2 - \sum_{t=1}^{T}\widehat{u}_{2,t}^2)/\sum_{t=1}^{T}\widehat{u}_{2,t}^2\right)$, where $\widehat{u}_{1,t}$ and $\widehat{u}_{2,t}$ are the in-sample residuals associated with least squares estimation of the small and big models, respectively, and where T denotes the sample size.
CM – The Clark and McCracken (2004) test outlined in Section 4.
DM – The Diebold and Mariano (1995) test outlined in Section 4.
CS – The Corradi and Swanson (2002,2007) test outlined in Section 3.
CCS – The Chao, Corradi and Swanson (2001) test discussed in Section 4.

Panel B: Data Generating Processes Used in Monte Carlo Experiments

$x_t = a_1 + a_2 x_{t-1} + u_{1,t}$, $u_{1,t} \sim iidN(0,1)$
$z_t = a_1 + a_3 z_{t-1} + u_{2,t}$, $u_{2,t} \sim iidN(0,1)$
Size1: $y_t = a_1 + a_2 y_{t-1} + a_4 z_{t-1} + u_{3,t}$, $u_{3,t} \sim iidN(0,1)$
Size2: $y_t = a_1 + a_2 y_{t-1} + a_4 z_{t-1} + a_3 u_{3,t-1} + u_{3,t}$
Power1 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(\tan^{-1}(x_{t-1}/2)) + a_4 z_{t-1} + u_{3,t}$
Power2 : $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1} + a_4 w_{t-1} + u_{3,t}$
Power3 : $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1}1\{x_{t-1} > a_1/(1-a_2)\} + a_4 z_{t-1} + u_{3,t}$
Power4 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(\tan^{-1}(x_{t-1}/2)) + a_4 z_{t-1} + a_3 u_{3,t-1} + u_{3,t}$
Power5: $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1} + a_4 z_{t-1} + a_3 u_{3,t-1} + u_{3,t}$
Power6: $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1}1\{x_{t-1} > a_1/(1-a_2)\} + a_4 z_{t-1} + a_3 u_{3,t-1} + u_{3,t}$.
Power7 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(x_{t-1}) + a_4 z_{t-1} + u_{3,t}$
Power8 : $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1}^2 + a_4 z_{t-1} + u_{3,t}$
Power9 : $y_t = a_1 + a_2 y_{t-1} + 2|x_{t-1}| + a_4 z_{t-1} + u_{3,t}$
Power10 : $y_t = a_1 + a_2 y_{t-1} + 2\exp(x_{t-1}) + a_4 z_{t-1} + a_3 u_{3,t-1} + u_{3,t}$
Power11: $y_t = a_1 + a_2 y_{t-1} + 2x_{t-1}^2 + a_4 z_{t-1} + a_3 u_{3,t-1} + u_{3,t}$
Power12: $y_t = a_1 + a_2 y_{t-1} + 2|x_{t-1}| + a_4 z_{t-1} + a_3 u_{3,t-1} + u_{3,t}$.

Note that the benchmark or "small" model in our test statistic calculations is always $y_t = \alpha_1 + \alpha_2 y_{t-1} + \alpha_3 z_{t-1} + \epsilon_t$; and the "big" model is the same, but with x_{t-1} or generic functions of x_{t-1} added as an additional regressor.

Table 2: Recursive Estimation Scheme - Rejection Frequencies of CS Test with $T = 540$, $P = 0.5T$ *

Model	Recur Block Bootstrap			BB, no PEE, no adj			Block Bootstrap		
	Panel A: $a_2 = 0.3$								
	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$
Size1	0.07	0.07	0.08	0.01	0.01	0.02	0.00	0.00	0.01
Size2	0.04	0.05	0.07	0.01	0.01	0.01	0.00	0.00	0.01
Power1	0.53	0.73	0.80	0.00	0.59	0.83	0.54	0.77	0.74
Power2	0.68	0.90	0.93	0.00	0.94	0.92	0.91	0.86	0.81
Power3	0.68	0.90	0.92	0.01	0.95	0.93	0.94	0.87	0.82
Power4	0.53	0.76	0.81	0.00	0.54	0.84	0.42	0.76	0.76
Power5	0.69	0.88	0.93	0.00	0.94	0.93	0.91	0.84	0.83
Power6	0.68	0.88	0.92	0.01	0.96	0.91	0.94	0.86	0.83
Power7	0.57	0.75	0.77	0.02	0.76	0.77	0.77	0.73	0.70
Power8	0.66	0.88	0.90	0.03	0.91	0.85	0.93	0.82	0.81
Power9	0.68	0.93	0.96	0.00	0.97	0.94	0.97	0.90	0.86
Power10	0.57	0.73	0.77	0.02	0.76	0.76	0.79	0.73	0.71
Power11	0.68	0.88	0.89	0.01	0.92	0.88	0.92	0.85	0.80
Power12	0.71	0.91	0.95	0.00	0.97	0.94	0.97	0.90	0.90
	Panel B: $a_2 = 0.6$								
	$l = 4$	$l = 10$	$l = 20$	$l = 4$	$l = 10$	$l = 20$	$l = 4$	$l = 10$	$l = 20$
Size1	0.05	0.07	0.08	0.01	0.01	0.03	0.01	0.01	0.01
Size2	0.03	0.07	0.07	0.00	0.02	0.01	0.00	0.01	0.01
Power1	0.59	0.69	0.75	0.00	0.65	0.80	0.56	0.64	0.68
Power2	0.71	0.84	0.86	0.01	0.91	0.84	0.79	0.75	0.76
Power3	0.78	0.86	0.89	0.07	0.92	0.86	0.80	0.78	0.78
Power4	0.57	0.69	0.78	0.00	0.61	0.82	0.56	0.66	0.69
Power5	0.73	0.85	0.86	0.01	0.92	0.86	0.80	0.78	0.77
Power6	0.77	0.87	0.91	0.05	0.92	0.88	0.81	0.78	0.77
Power7	0.56	0.64	0.68	0.05	0.64	0.67	0.53	0.60	0.63
Power8	0.72	0.83	0.86	0.14	0.87	0.82	0.77	0.75	0.73
Power9	0.82	0.92	0.93	0.04	0.95	0.88	0.83	0.80	0.80
Power10	0.57	0.62	0.67	0.07	0.67	0.67	0.62	0.61	0.63
Power11	0.76	0.83	0.87	0.15	0.86	0.82	0.79	0.72	0.73
Power12	0.80	0.90	0.93	0.04	0.94	0.88	0.86	0.81	0.79
	Panel C: $a_2 = 0.9$								
	$l = 10$	$l = 20$	$l = 50$	$l = 10$	$l = 20$	$l = 50$	$l = 10$	$l = 20$	$l = 50$
Size1	0.01	0.03	0.07	0.00	0.01	0.03	0.00	0.00	0.01
Size2	0.01	0.03	0.06	0.00	0.01	0.02	0.00	0.00	0.01
Power1	0.41	0.56	0.64	0.00	0.47	0.75	0.29	0.53	0.61
Power2	0.59	0.71	0.77	0.06	0.78	0.79	0.58	0.65	0.72
Power3	0.61	0.72	0.79	0.09	0.82	0.77	0.61	0.68	0.71
Power4	0.42	0.53	0.64	0.00	0.43	0.73	0.27	0.50	0.61
Power5	0.61	0.69	0.77	0.03	0.81	0.78	0.57	0.67	0.70
Power6	0.62	0.72	0.80	0.12	0.81	0.79	0.59	0.68	0.72
Power7	0.41	0.47	0.54	0.07	0.47	0.54	0.34	0.46	0.53
Power8	0.57	0.67	0.75	0.23	0.76	0.69	0.56	0.62	0.66
Power9	0.61	0.72	0.82	0.13	0.83	0.81	0.62	0.67	0.72
Power10	0.42	0.47	0.53	0.06	0.50	0.54	0.35	0.47	0.52
Power11	0.60	0.66	0.75	0.27	0.76	0.73	0.59	0.63	0.66
Power12	0.64	0.76	0.83	0.17	0.85	0.81	0.59	0.66	0.71

* Notes: All entries are rejection frequencies of the null hypothesis of equal predictive accuracy based on 10% nominal size critical values constructed using the bootstrap approaches discussed above, where l denotes the block length, and empirical bootstrap distributions are constructed using 100 bootstrap statistics. In particular, "Recur Block Bootstrap" is the bootstrap developed in this paper, "BB, no PEE, no adj" is a naive block bootstrap where no parameter estimation error is assumed, and no recentering (i.e. adjustment) is done in parameter estimation or bootstrap statistic construction, "Block Bootstrap" is the usual block bootstrap that allows for parameter estimation error, but does not recenter parameter estimates or bootstrap statistics. For all models denoted Power$i$, $i = 1, ..., 12$, data are generated with (non) linear Granger causality (see above for further discussion of DGPs. In all experiments, the ex ante forecast period is of length P, which is set equal to $(1/2)T$, where T is the sample size. All models are estimated recursively, so that parameter estimates are updated before each new prediction is constructed. All reported results are based on 500 Monte Carlo simulations. See Table 1 and Section 4 for further details.

Table 3: Rolling Estimation Scheme - Rejection Frequencies of CS Test with $T = 540$, $P = 0.5T$ *

Model	Rolling Block Bootstrap			BB, no PEE, no adj			Block Bootstrap		
	Panel A: $a_2 = 0.3$								
	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$	$l = 2$	$l = 5$	$l = 10$
Size1	0.05	0.06	0.07	0.01	0.02	0.01	0.00	0.00	0.00
Size2	0.03	0.06	0.07	0.01	0.02	0.02	0.00	0.00	0.00
Power1	0.62	0.87	0.90	0.00	0.67	0.88	0.80	0.86	0.83
Power2	0.74	0.95	0.96	0.01	0.97	0.94	0.98	0.94	0.91
Power3	0.77	0.95	0.98	0.01	0.97	0.94	0.98	0.94	0.92
Power4	0.59	0.89	0.92	0.00	0.55	0.89	0.71	0.87	0.82
Power5	0.75	0.96	0.97	0.00	0.97	0.95	0.98	0.94	0.92
Power6	0.76	0.95	0.97	0.01	0.98	0.94	0.98	0.95	0.92
Power7	0.67	0.83	0.84	0.03	0.82	0.82	0.86	0.82	0.80
Power8	0.78	0.90	0.91	0.04	0.92	0.88	0.96	0.90	0.86
Power9	0.79	0.96	0.96	0.01	0.97	0.94	0.98	0.94	0.92
Power10	0.68	0.82	0.84	0.03	0.81	0.81	0.86	0.82	0.82
Power11	0.80	0.90	0.92	0.04	0.91	0.89	0.95	0.89	0.86
Power12	0.78	0.94	0.96	0.00	0.97	0.94	0.98	0.93	0.92
	Panel B: $a_2 = 0.6$								
	$l = 4$	$l = 10$	$l = 20$	$l = 4$	$l = 10$	$l = 20$	$l = 4$	$l = 10$	$l = 20$
Size1	0.03	0.04	0.05	0.01	0.00	0.01	0.00	0.00	0.01
Size2	0.03	0.03	0.06	0.01	0.01	0.01	0.00	0.00	0.00
Power1	0.68	0.81	0.85	0.01	0.69	0.85	0.74	0.75	0.75
Power2	0.82	0.91	0.94	0.04	0.94	0.90	0.90	0.84	0.84
Power3	0.88	0.96	0.95	0.08	0.97	0.91	0.90	0.84	0.85
Power4	0.65	0.84	0.88	0.01	0.67	0.85	0.69	0.77	0.78
Power5	0.84	0.93	0.94	0.02	0.95	0.91	0.88	0.84	0.83
Power6	0.89	0.96	0.95	0.09	0.96	0.92	0.93	0.88	0.86
Power7	0.65	0.70	0.72	0.07	0.67	0.68	0.68	0.67	0.68
Power8	0.83	0.89	0.89	0.17	0.89	0.85	0.85	0.82	0.81
Power9	0.90	0.94	0.94	0.12	0.94	0.92	0.92	0.89	0.88
Power10	0.63	0.70	0.73	0.05	0.69	0.69	0.65	0.68	0.67
Power11	0.82	0.88	0.90	0.14	0.89	0.86	0.85	0.85	0.82
Power12	0.90	0.93	0.94	0.08	0.94	0.92	0.92	0.89	0.89
	Panel C: $a_2 = 0.9$								
	$l = 10$	$l = 20$	$l = 50$	$l = 10$	$l = 20$	$l = 50$	$l = 10$	$l = 20$	$l = 50$
Size1	0.01	0.03	0.06	0.00	0.00	0.01	0.00	0.00	0.00
Size2	0.01	0.02	0.03	0.00	0.01	0.01	0.00	0.00	0.01
Power1	0.37	0.54	0.72	0.00	0.53	0.75	0.36	0.56	0.63
Power2	0.64	0.75	0.83	0.13	0.82	0.82	0.60	0.69	0.75
Power3	0.68	0.78	0.88	0.16	0.86	0.84	0.67	0.71	0.77
Power4	0.35	0.54	0.74	0.00	0.43	0.76	0.32	0.52	0.67
Power5	0.61	0.75	0.85	0.07	0.80	0.82	0.60	0.67	0.77
Power6	0.71	0.80	0.87	0.16	0.86	0.84	0.67	0.72	0.78
Power7	0.48	0.54	0.60	0.06	0.49	0.55	0.44	0.54	0.58
Power8	0.67	0.77	0.81	0.24	0.81	0.75	0.65	0.68	0.72
Power9	0.76	0.83	0.89	0.14	0.89	0.86	0.70	0.76	0.79
Power10	0.47	0.56	0.60	0.06	0.50	0.56	0.42	0.54	0.58
Power11	0.67	0.75	0.81	0.21	0.82	0.75	0.66	0.70	0.73
Power12	0.75	0.84	0.89	0.18	0.89	0.85	0.71	0.77	0.79

* Notes: See notes to Table 2.

Table 4: Recursive Estimation Scheme - Rejection Frequencies of Various Tests with $T = 540$, $P = 0.5T$ *

Model	Assume $\pi = 0$			Assume $\pi > 0$		Recur Block Bootstrap		
	F	DM	CM	DM	CM	CCS-$l1$	CCS-$l2$	CCS-$l3$
Panel A: $a_2 = 0.3$								
Size1	0.11	0.01	0.06	0.10	0.10	0.20	0.21	0.20
Size2	0.11	0.01	0.07	0.11	0.11	0.17	0.17	0.17
Power1	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.94
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96
Power4	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.96
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
Power7	1.00	1.00	1.00	1.00	1.00	0.98	0.89	0.78
Power8	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.92
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
Power10	1.00	1.00	1.00	1.00	1.00	0.99	0.88	0.77
Power11	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.91
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
Panel B: $a_2 = 0.6$								
Size1	0.09	0.02	0.04	0.10	0.09	0.19	0.22	0.20
Size2	0.11	0.01	0.06	0.10	0.09	0.14	0.16	0.19
Power1	1.00	1.00	1.00	1.00	1.00	0.95	0.91	0.89
Power2	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.95
Power3	1.00	1.00	1.00	1.00	1.00	0.98	0.94	0.93
Power4	1.00	1.00	1.00	1.00	1.00	0.96	0.91	0.86
Power5	1.00	1.00	1.00	1.00	1.00	0.98	0.96	0.94
Power6	1.00	1.00	1.00	1.00	1.00	0.99	0.94	0.92
Power7	1.00	1.00	1.00	1.00	1.00	0.80	0.69	0.64
Power8	1.00	1.00	1.00	1.00	1.00	0.97	0.89	0.84
Power9	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.91
Power10	1.00	1.00	1.00	1.00	1.00	0.81	0.67	0.61
Power11	1.00	1.00	1.00	1.00	1.00	0.97	0.86	0.84
Power12	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.92
Panel C: $a_2 = 0.9$								
Size1	0.10	0.01	0.06	0.11	0.11	0.10	0.11	0.14
Size2	0.13	0.01	0.08	0.11	0.14	0.09	0.11	0.16
Power1	1.00	1.00	1.00	1.00	1.00	0.68	0.74	0.76
Power2	1.00	1.00	1.00	1.00	1.00	0.81	0.83	0.86
Power3	1.00	1.00	1.00	1.00	1.00	0.80	0.80	0.82
Power4	1.00	1.00	1.00	1.00	1.00	0.63	0.67	0.73
Power5	1.00	1.00	1.00	1.00	1.00	0.80	0.82	0.85
Power6	1.00	1.00	1.00	1.00	1.00	0.76	0.79	0.84
Power7	1.00	0.96	1.00	1.00	1.00	0.47	0.43	0.49
Power8	1.00	1.00	1.00	1.00	1.00	0.70	0.73	0.78
Power9	1.00	1.00	1.00	1.00	1.00	0.75	0.78	0.82
Power10	1.00	0.96	1.00	1.00	1.00	0.42	0.45	0.48
Power11	1.00	1.00	1.00	1.00	1.00	0.69	0.75	0.78
Power12	1.00	1.00	1.00	1.00	1.00	0.77	0.77	0.81

* Notes: See notes to Table 2. Test statistics, denoted by F, DM, CM, CS, and CCS are summarized in Table 1. Block lengths are denoted by $l1$, $l2$, and $l3$, so that $CCS - l3$ is the CCS test with block length $l3$. Block lengths corrspond to those used in Table 2 and 3, so that for $a_2 = 0.3$, $l1, l2, l3$=2,5,10. The block lengths for $a_2 = 0.6$ and $a_2 = 0.9$ are $l1, l2, l3$=4,10,20 and $l1, l2, l3$=10,20,50, respectively. $\pi = 0$ corresponds to the case where standard critical values based upon the assumption that parameter estimation error vanishes asymptotically are used (i.e. $\pi = lim_{T->\infty} P/R = 0$). $\pi > 0$ corresponds to the case where nonstandard critical values (see McCracken (2004)) based upon the assumption that parameter estimation error does not vanish asymptotically are used (i.e. $\pi = lim_{T->\infty} P/R > 0$). In this case, we assume that $\pi = 1$.

Table 5: Rolling Estimation Scheme - Rejection Frequencies of Various Tests with $T = 540$, $P = 0.5T$ *

Model	Assume $\pi = 0$			Assume $\pi > 0$		Recur Block Bootstrap		
	F	DM	CM	DM	CM	CCS-$l1$	CCS-$l2$	CCS-$l3$
Panel A: $a_2 = 0.3$								
Size1	0.11	0.00	0.06	0.07	0.09	0.17	0.17	0.16
Size2	0.10	0.00	0.06	0.10	0.09	0.14	0.14	0.13
Power1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
Power2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
Power4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
Power5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Power7	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.81
Power8	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.94
Power9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Power10	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.80
Power11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95
Power12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Panel B: $a_2 = 0.6$								
Size1	0.10	0.01	0.05	0.08	0.08	0.13	0.14	0.16
Size2	0.13	0.01	0.06	0.10	0.10	0.11	0.13	0.15
Power1	1.00	1.00	1.00	1.00	1.00	0.99	0.94	0.93
Power2	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.96
Power3	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.96
Power4	1.00	1.00	1.00	1.00	1.00	0.99	0.94	0.91
Power5	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.96
Power6	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.96
Power7	1.00	0.99	1.00	1.00	1.00	0.86	0.71	0.67
Power8	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.90
Power9	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.95
Power10	1.00	0.99	1.00	1.00	1.00	0.84	0.72	0.68
Power11	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.91
Power12	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.94
Panel C: $a_2 = 0.9$								
Size1	0.11	0.02	0.04	0.08	0.10	0.07	0.12	0.13
Size2	0.13	0.02	0.06	0.12	0.13	0.06	0.07	0.13
Power1	1.00	1.00	1.00	1.00	1.00	0.72	0.72	0.80
Power2	1.00	1.00	1.00	1.00	1.00	0.85	0.83	0.86
Power3	1.00	1.00	1.00	1.00	1.00	0.80	0.82	0.82
Power4	1.00	1.00	1.00	1.00	1.00	0.70	0.71	0.78
Power5	1.00	1.00	1.00	1.00	1.00	0.83	0.82	0.87
Power6	1.00	1.00	1.00	1.00	1.00	0.81	0.80	0.85
Power7	1.00	0.97	1.00	1.00	1.00	0.46	0.44	0.49
Power8	1.00	1.00	1.00	1.00	1.00	0.80	0.78	0.82
Power9	1.00	1.00	1.00	1.00	1.00	0.87	0.83	0.87
Power10	1.00	0.97	1.00	1.00	1.00	0.48	0.48	0.53
Power11	1.00	1.00	1.00	1.00	1.00	0.80	0.81	0.82
Power12	1.00	1.00	1.00	1.00	1.00	0.83	0.84	0.87

* Notes: See notes to Table 4.

Table 6: Mean Square Forecast Errors and the Marginal Predictive Content of $M2$ for Output*

	Recursive		Rolling	
Start Year	*small model*	*big model*	*small model*	*big model*
1974	0.0000398	0.0000395	**0.0000393**	0.0000410
1975	0.0000359	0.0000355	**0.0000345**	0.0000363
1976	0.0000352	0.0000349	**0.0000338**	0.0000352
1977	0.0000353	0.0000350	**0.0000340**	0.0000345
1978	0.0000350	0.0000348	**0.0000337**	0.0000344
1979	0.0000345	0.0000341	**0.0000334**	0.0000339
1980	0.0000343	0.0000338	**0.0000333**	0.0000336
1981	0.0000328	0.0000325	**0.0000323**	0.0000330
1982	0.0000321	**0.0000317**	**0.0000317**	0.0000323
1983	0.0000285	**0.0000280**	0.0000284	0.0000287
1984	0.0000271	**0.0000261**	0.0000274	0.0000271
1985	0.0000281	**0.0000271**	0.0000280	0.0000281
1986	0.0000284	**0.0000274**	0.0000281	0.0000283
1987	0.0000281	**0.0000272**	0.0000278	0.0000279
1988	0.0000279	**0.0000266**	0.0000272	0.0000270
1989	0.0000292	**0.0000279**	0.0000285	0.0000280
1990	0.0000281	**0.0000269**	0.0000275	**0.0000269**
1991	0.0000278	0.0000269	0.0000269	**0.0000267**
1992	0.0000278	0.0000271	**0.0000267**	0.0000270
1993	0.0000288	0.0000280	**0.0000275**	0.0000279

* Notes: For the empirical work, the variables used are monthly observations of industrial production (IP), the wholesale price index (P), the secondary market rate on 90-day U.S. Treasury bills (R), the interest rate on three-month prime commercial paper (C) and Divisia monetary aggregates of money supply ($M2$). The sample period is 1959-01 to 2003-12.
We define the small model as:

$$y_t = \theta^\dagger_{1,1} + \theta^\dagger_{1,2} y_{t-1} + \theta^\dagger_{1,3} z_{1,t-1} + u_{1,t}$$

where

$$\theta^\dagger_1 = (\theta^\dagger_{1,1}, \theta^\dagger_{1,2}, \theta^\dagger_{1,3})' = \arg \min_{\theta_1 \in \Theta_1} E(q_1(y_t - \theta_{1,1} - \theta_{1,2} y_{t-1} - \theta_{1,3} z_{1,t-1})) \text{ is defined conformably,}$$

$$y_t = (\Delta \log IP_t, \Delta \log P_t, \Delta R_t)'$$

and

$$z_{1,t-1} = C_{t-1} - R_{t-1}.$$

We further define the generic alternative (big) model as:

$$y_t = \theta^\dagger_{2,1}(\gamma) + \theta^\dagger_{2,2}(\gamma) y_{t-1} + \theta^\dagger_{2,3}(\gamma) z_{1,t-1} + \theta^\dagger_{2,4}(\gamma) w(Z^{t-1}, \gamma) + u_{2,t}(\gamma)$$

where

$$\theta^\dagger_2(\gamma) = (\theta^\dagger_{2,1}(\gamma), \theta^\dagger_{2,2}(\gamma), \theta^\dagger_{2,3}(\gamma), \theta^\dagger_{2,4}(\gamma))' = \arg \min_{\theta_2 \in \Theta_2} E(q_1(y_t - \theta_{2,1} - \theta_{2,2} y_{t-1} - \theta_{2,3} z_{1,t-1} - \theta_{2,4} w(Z^{t-1}, \gamma)))$$

and

$$\begin{aligned} y_t &= (\Delta \log IP_t, \Delta \log P_t, \Delta \log M2_t, \Delta R_t) \\ z_{1,t-1} &= C_{t-1} - R_{t-1} \\ z_{2,t-1} &= \log M2_{t-1} - \log IP_{t-1} - \log P_{t-1}. \end{aligned}$$

$z_{1,t-1}$ and $z_{2,t-1}$ can be interpreted as vector error correction terms. Mean square forecast errors are reported for the small and big models as defined above. Since we are interested in examining the (non)linear marginal predictive content of money for income, our forecasting analysis and test statistics are constructed based on estimates of the first equation in the vector error correction model specified above. All predictions are 1-step ahead output and predictive periods begin in the year given in the first column of entries in the table. Entries in bold represent the lowest MSFE for the corresponding year in which prediction started.

Table 7: Tests for the Marginal Predictive Content of $M2$ for Output*

	Prediction Period Begins in		
Test Statistic	1987($\pi = 0.6$)	1981($\pi = 1.0$)	1978($\pi = 1.4$)
Panel A: Sig Level = 5%; Recursive			
CS (Recur Block Bootstrap)	no reject	no reject	no reject
CCS (Recur Block Bootstrap)	no reject	reject	no reject
F	reject	reject	reject
DM (Tabulated CVs)	reject	no reject	no reject
CM (Tabulated CVs)	reject	reject	reject
Panel B: Sig Level = 10%; Recursive			
CS (Recur Block Bootstrap)	no reject	no reject	no reject
CCS (Recur Block Bootstrap)	reject	reject	no reject
F	reject	reject	reject
DM (Tabulated CVs)	reject	reject	reject
CM (Tabulated CVs)	reject	reject	reject
Panel C: Sig Level = 5%; Rolling			
CS (Recur Block Bootstrap)	no reject	no reject	no reject
CCS (Recur Block Bootstrap)	no reject	no reject	no reject
F	reject	reject	reject
DM (Tabulated CVs)	no reject	no reject	no reject
CM (Tabulated CVs)	reject	no reject	no reject
Panel D: Sig Level = 10%; Rolling			
CS (Recur Block Bootstrap)	no reject	no reject	no reject
CCS (Recur Block Bootstrap)	reject	reject	no reject
F	reject	reject	reject
DM (Tabulated CVs)	no reject	no reject	no reject
CM (Tabulated CVs)	reject	no reject	reject

* Notes: Entries denote either rejection (reject) or failure to reject (no reject) the null hypothesis that $M2$ has no marginal predictive content for output. Entries denote nominal 5% and 10% level test rejection based on critical values constructed using the approach signified in brackets in the first column of the table. The models are as described in the notes to Table 6. All models use monthly data and all predictions are based on 1-step ahead recursive and rolling schemes.