Corradi, Valentina; Swanson, Norman R.

Working Paper

Some Recent Developments in Predictive Accuracy Testing With Nested Models and (Generic) Nonlinear Alternatives

Working Paper, No. 2003-16

Provided in Cooperation with:
Department of Economics, Rutgers University

Suggested Citation: Corradi, Valentina; Swanson, Norman R. (2003) : Some Recent Developments in Predictive Accuracy Testing With Nested Models and (Generic) Nonlinear Alternatives, Working Paper, No. 2003-16, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:
https://hdl.handle.net/10419/23173

Some Recent Developments in Predictive Accuracy Testing With Nested Models and (Generic) Nonlinear Alternatives*

Valentina Corradi[1] and Norman R. Swanson[2]

[1] University of Exeter
[2] Rutgers University

August 2002
revised: January 2003

Abstract

Forecasters and applied econometricians are often interested in comparing the predictive accuracy of nested competing models. A leading example of nestedness is when predictive ability is equated with "out-of-sample Granger causality". In particular, it is often of interest to assess whether historical data from one variable are useful when constructing a forecasting model for another variable, and hence our use of terminology such as "out-of-sample Granger causality" (see e.g. Ashley, Granger and Schmalensee (1980)). In this paper we examine and discuss three key issues one is faced with when constructing predictive accuracy tests, namely: the contribution of parameter estimation error, the choice of linear versus nonlinear models, and the issue of (dynamic) misspecification, with primary focus on the latter of these issues. One of our main conclusions is that there are a number of easy to apply statistics constructed using out of sample conditional moment conditions which are robust to the presence of dynamic misspecification under both hypothesis. We provide some new Monte Carlo findings and empirical evidence based on the use of such tests. In particular, we analyze the finite sample properties of the consistent out of sample test of Corradi and Swanson (2002) using data generating processes calibrated with U.S. money and output, and empirically investigate the (non)linear marginal predictive content of money for output. Our Monte Carlo evidence suggests that the tests perform adequately in finite samples, and our empirical evidence suggests that there is non useful (non)linear information in money growth that is not already contained in lags of output growth, when the objective is output growth prediction.

JEL classification: C22, C51.
Keywords: Conditional p-value, bootstrap, forecasting, out-of-sample predictive accuracy, parameter estimation error.

---

* Valentina Corradi, Department of Economics, University of Exeter, Streatham Court, Exeter EX4 4PU, U.K., V.Corradi@exeter.ac.uk. Norman R. Swanson, Department of Economics, Rutgers University, New Brunswick, New Jersey, 08901-1248, USA, nswanson@econ.rutgers.edu. The authors wish to thank the co-editor in charge, an anonymous referee, Graham Elliott and Allan Timmerman for useful comments and suggestions. Parts of this paper were written while the second author was visiting the University of California, San Diego, and he is grateful to faculty members there for hosting him and for providing a stimulating research environment.

1 Introduction

The discussion of forecast model comparison by Ashley, Granger and Schmalensee (1980) and Granger and Newbold (1986) are two of the main driving forces behind much of the current literature on predictive ability, although it has really only been over the past ten years or so that the issue of (out of sample) predictive accuracy evaluation has received increasing attention from both theoretical and applied perspectives. One of the most important recent contributions is the seminal paper of Diebold and Mariano (1995, DM), in which a general test of equal predictive accuracy between two competing models is proposed. Since then, efforts have been made to further generalize DM type tests in order to: account for parameter estimation error (see e.g. West (1996) and West and McCracken (1998)); allow for non differentiable loss functions together with parameter estimation error (McCracken (2000)); extend the DM framework to the case of integrated and cointegrated variables (see e.g. Clements and Hendry (1999a,b) and Corradi, Swanson and Olivetti (2001)); and address the issue of joint comparison of more than two competing models (see e.g. Sullivan, Timmermann and White (1999, 2001) and White (2000)).

In applied finance and in financial risk management, uncovering the best loss function specific model for the conditional mean often does not suffice. Therefore, attention has also recently focused on the issue of (conditional) forecast interval evaluation (see e.g. Christoffersen (1998), Christoffersen and Diebold (2000)) and, as a natural extension, the issue of predictive density evaluation (see e.g. Diebold, Gunther and Tay (1998), Bai (1998), Diebold, Hahn and Tay (1999), Clements and Smith (2000,2001), Corradi and Swanson (2001), Hong (2001), and the references cited therein). One of the common features of many of the papers cited above is that nonnested forecasting models are compared. However, forecasters and applied econometricians are often interested in comparing the predictive accuracy of nested competing models. The most obvious context in which nested models should be compared is when predictive ability is equated with "out-of-sample Granger causality", for example. In particular, it is often of interest to assess whether historical data from one variable are useful for constructing a forecasting model for another variable, hence our use of terminology such as "out-of-sample Granger causality".[1]

[1]Granger (1980) summarizes his personal viewpoint on testing for causality, and outlines what he considers to be a useful operational version of his original definition of causality (Granger (1969)). This operational version is based on a comparison of the one-step ahead predictive ability of competing models. However, the common practice is to test for Granger causality using in-sample F-tests.

A common problem that arises when comparing nested models is that the statistic vanishes in probability under the null of equal predictive ability. This is the case with the Diebold-Mariano statistic, for example. Therefore, efforts have been made to construct tests which have a nondegenerate limiting distribution under the null. A partial list of the recent contributions that address this by constructing a variety of new tests such as those based on the encompassing principle includes Harvey, Leybourne and Newbold (1997), McCracken (1999) and Clark and McCracken (2001). Related papers based on out of sample moment conditions include Chao, Corradi and Swanson (2001, CCS), Corradi and Swanson (2002, CS).

In this paper we confine our attention to the issue of evaluating nested models. In particular, we discuss and evaluate a variety of recent testing contributions that address what we feel are three of the key outstanding issues in predictive accuracy testing, namely: parameter estimation error, the use of nonlinear versus linear models, and dynamic misspecification. With regard to linear versus nonlinear models, it is worth noting that in applied time series analysis there has been a long standing debate concerning whether simple linear models (e.g. ARMA models) provide out of sample forecasts which are (at least) as accurate as more sophisticated nonlinear models. If this were shown to be the case, then there would be no point in using nonlinear models for out-of-sample prediction, even if the linear models could be shown to be incorrectly specified. This debate is addressed, for example, by Teräsvirta and Anderson (1992), Granger and Teräsvirta (1993), Swanson and White (1995, 1997), and the references cited therein. The notion of parameter estimation error (PEE) is crucial because confidence bands around point predictive accuracy tests can often be dominated by the effect of PEE, and failure to account for PEE can thus lead to very imprecise if not incorrect inferences (see e.g. West (1996), White (2000), Chao, Corradi and Swanson (2001), and the references contained therein). Finally, the area which these authors feel is most often overlooked in discussions of predictive accuracy is that of model misspecification, either dynamic or otherwise. This topic is crucial if one takes the objective view that all prediction models are approximations of some underlying (and perhaps highly complex) reality. As all models are typically parsimonious approximations, it is likely that *even* if the null model is correctly specified for a given information set, it is still likely to be dynamically misspecified, as it does not take into account all of the relevant history. Failing to take into account the possibility of dynamic misspecification under both hypotheses leads to incorrect inference, as critical values are generally incorrect in such cases. White (2000) tackles this problem by allowing for the comparison of many

models at once, rather than the comparison of only two models, as in Diebold and Mariano (1995). By allowing for many models and conducting data mining exercises, White implicitly assumes that there are many approximations of the truth that are probably worth evaluating, and he accounts for this feature of one's empirical investigation by designing a data snooping technique based on the bootstrap that accounts for sequential test bias associated with comparing many models. CS take the data snooping approach of White (2000) one step further by allowing for (dynamic) misspecification among competing prediction models (under both hypotheses), while at the same time ensuring test consistency against generic nonlinear alternatives. This departs from the usual practice of comparing the predictive accuracy of a finite and fixed set of linear and (less often) nonlinear models, and allows for the construction of ex ante Granger causality type predictive accuracy tests when the precise form of the nonlinearity is unknown. In this sense, these tests combine the consistent specification testing and predictive accuracy testing literatures.

In the sequel, we review the recent literature and debate on predictive comparison of nested models, with particular emphasis on the results of CCS and CS, and provide new Monte Carlo and empirical evidence. The Monte Carlo experiments carried out below are based on data generating processes calibrated with U.S. money (M2) and output (industrial production), and are designed to evaluate the performance of CS type tests. Our empirical investigation focuses on the (non)linear marginal predictive content of M2 for industrial production, and our results suggest that there is no useful (non)linear information in money growth that is not already contained in lags of output growth, when the objective is output growth prediction.

The rest of the paper is organized as follows. Section 2 discusses linear and nonlinear predictive accuracy tests, and Section 3 generalizes the results of Section 2 to generic nonlinear alternatives. Section 4 discusses results from a series of Monte Carlo experiments and from the empirical examination of a macroeconomic dataset using a generic nonlinear test of predictive accuracy. Concluding remarks are gathered in Section 5.

2 Linear and Nonlinear Predictive Accuracy Tests

We begin by discussing out of sample tests of linear restrictions.[2] In empirical forecasting applications, one often starts from a restricted autoregressive (AR(p)) model,

$$x_t = \sum_{j=1}^{p} \beta_j^* x_{t-j} + u_{0,t} \tag{1}$$

and compare it with a "larger" unrestricted autoregressive model,[3]

$$x_t = \sum_{j=1}^{p} \beta_j^* x_{t-j} + \sum_{j=1}^{k} \alpha_j^* y_{t-j} + u_{1,t} \tag{2}$$

In-sample "Granger causality" tests of $H_0 : \alpha_j^* = 0, \forall j$ versus $H_A : \alpha_j^* \neq 0$, for some j, can then easily be constructed using Wald type statistics which have a limiting χ^2 distribution under H_0. For example, in the case of martingale difference errors under the null, and given a maintained assumption of conditional homoskedasticity, one commonly constructs $F = \frac{(RRSS-URSS)/k}{URSS/(T-k)}$ where $RRSS$ and $URSS$ are the sum of least squares residuals from the restricted and the unrestricted models, k is the number of restrictions, and $kF \xrightarrow{d} \chi_k^2$ under H_0. An out-of-sample analog to this test is proposed by Clark and McCracken (2001,CM). In their test, as in all recursive type predictive accuracy tests, one estimates (1) and (2) using observations $t = 1, 2, \ldots R$, and computes $\widehat{u}_{0,R+1} = x_{R+1} - \sum_{j=0}^{p-1} \widehat{\beta}_{R,j} x_{t-j}$ and $\widehat{u}_{1,R+1} = x_{R+1} - \sum_{j=0}^{p-1} \widehat{\beta}_{R,j} x_{t-j} - \sum_{j=0}^{k-1} \widehat{\alpha}_{R,j} y_{t-j}$, and then re-estimates the model using $R+1$ observations and constructs $\widehat{\beta}_{R+1,j}, \widehat{\alpha}_{R,j}$, $\widehat{u}_{0,R+2}$ and $\widehat{u}_{1,R+2}$. This procedure is repeated until sequences of P ex ante forecast errors have been constructed, with $P + R = T$, where T is the sample size. CM also suggest a new encompassing test statistic,

$$ENC\_NEW = P \frac{\overline{c}}{P^{-1} \sum_{t=R}^{T-1} \widehat{u}_{1,t+1}^2},$$

where $c_{t+1} = \widehat{u}_{0,t+1}(\widehat{u}_{0,t+1} - \widehat{u}_{1,t+1})$, and $\overline{c} = P^{-1} \sum_{t=R}^{T-1} c_{t+1}$. They show that the statistic above vanishes if $P/R \to 0$, while it has a nonstandard limiting distribution if $P/R \to \pi \neq 0$.[4] The ENC_NEW test is very easy to compute and critical values have been tabulated by the authors,

[2] Hereafter β^* denotes the best linear predictor of y_t given its past history. Analogously, in the sequel, $\delta^* = (\beta^*, \alpha^*)'$ denotes the best linear predictor of y_t given its past and the past of x_{t-1}.

[3] All results discussed below generalize straightforwardly to the case where both the restricted and unrestricted models contain the past of other explanatory variables and/or an intercept.

[4] Clark and McCracken (2001) also provide the asymptotic distributions of related encompassing tests (such as that of Harvey, Leybourne and Newbold (1997)) for the case of recursively estimated parameters.

for various values of π. However, the statistic is written in a non HAC (heteroskedasticity and autocorrelation) robust form, so that the critical values are valid only if the errors, under the null, form a martingale difference and conditionally homoskedastic sequence. The martingale difference error assumption is essentially equivalent to an assumption of correct dynamic misspecification under the null. However, as pointed out above, it is preferable to evaluate the relative performance of (dynamically) misspecified models. Indeed, if the null model is dynamically correctly specified, then there may be no compelling reason for performing an out of sample version of the particular test being used (such as the F test) in the first place. For example, and broadly speaking, in the case of correct specification under the null, t and F tests are known to be optimal. In a recent paper, Inoue and Killian (2002) provide Monte Carlo findings showing that out of sample tests are not more reliable than in sample tests in such cases. However, as their Monte Carlo DGPs impose dynamic correct specification under the null, it is perhaps not too surprising that the power of in sample tests is often higher, as one uses the entire sample. Under a similar set-up as that used by Inoue and Killian (2002), Rossi (2001) shows that out of sample tests are often suboptimal. These last two papers serve to underscore the importance of remembering that out of sample tests are not useful in all contexts, and in particular out of sample tests are likely to be most useful in contexts where (dynamic) misspecification is allowed under both hypotheses. However, in the case of misspecification one must be careful when obtaining distributional results for even the most standard out of sample tests.

Intuitively, in order to allow for dynamic misspecification under both hypothesis, we want to construct a HAC robust test, thus allowing for comparisons in the case of non martingale difference sequence scores. This property is crucial when comparing multiple (or even two) prediction models, as there is absolutely no reason to believe that one of the models is correctly specified, either dynamically or otherwise. The Diebold Mariano test can be written in a robust form, by simply scaling the numerator by a heteroskedasticity and autocorrelation consistent (HAC) robust (co)variance estimator (e.g. see Newey and West (1997)). However, it is easy to see that when comparing nested models the DM statistic vanishes in probability under the null. Consider

$$DM_P = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(f(\widehat{u}_{0,t+1}) - f(\widehat{u}_{1,t+1}) \right) / \widehat{\sigma}_P, \tag{3}$$

where $\widehat{\sigma}_P^2$ is a HAC variance estimator of $d_t = f(\widehat{u}_{0,t+1}) - f(\widehat{u}_{1,t+1})$, with f denoting a given loss

function. The null hypothesis here is that of equal predictive ability, and is written as

$$H_0' : E(f(u_{0,t+1})) - E(f(u_{1,t+1})) = 0.$$

Notice that when the numerator of (3) is expanded (via a mean value expansion) around the "true" parameter values we obtain

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (f(u_{0,t+1}) - f(u_{1,t+1})) - \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \nabla_\beta f|_{\overline{\beta}} (\widehat{\beta}_t - \beta^*) + \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \nabla_\delta f|_{\overline{\delta}} (\widehat{\delta}_t - \delta^*), \qquad (4)$$

where $\overline{\beta} \in (\widehat{\beta}_t, \beta^*)$, $\overline{\delta} \in (\widehat{\delta}_t, \delta^*)$, and $\beta^* = (\beta_1^*, \ldots, \beta_p^*)', \delta^* = (\beta^*, \alpha^*)'$ and "hat" denotes the quasi-maximum likelihood estimator, for example. The first term in (4) is identically equal to zero, under H_0, while the second and third terms vanish in probability if $P/R \to 0$ and/or if the same loss function is used for estimation and prediction. In fact, if we use the same loss function for estimation and prediction (i.e. the models are estimated by OLS and f is a quadratic loss function), then $\frac{1}{P} \sum_{t=R}^{T-1} \nabla_\beta f|_{\overline{\beta}}$ and $\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \nabla_\delta f|_{\overline{\delta}}$ converge in probability to their respective means, which are identically to zero, because of the first order conditions. Therefore the statistic vanishes in probability. McCracken (1999) shows that if $\pi > 0$, then $\sqrt{P}DM$ has a nondegenerate, non standard limiting distribution, whose critical values can be tabulated, as they are nuisance parameters free. Again, it should be stressed that such critical values are valid under the additional assumption that the errors form a conditionally homoskedastic martingale difference sequence. This, then, returns us to the problem of using a test that is not robust to misspecification.

In order to address this problem, a simple out of sample test for forecast evaluation of nested linear models has been proposed by Chao, Corradi and Swanson (2001). The suggested test statistic is based on

$$m_P = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \widehat{u}_{0,t+1} Y_t, \qquad (5)$$

where $\widehat{u}_{0,t+1} = x_{t+1} - \sum_{j=0}^{p-1} \widehat{\beta}_{t,j} x_{t-j}$, $Y_t = (y_t, y_{t-1,} \ldots, y_{t-k-1})'$. It is important to note that tests formed using m_P do not require the restricted or unrestricted models to be dynamically correctly specified. In fact, if we form a test statistic by rescaling m_P by a HAC robust variance estimator, we immediately obtain a statistic that is robust to dynamic misspecification. In this case the hypotheses of interest are

$\widetilde{H}_0 : E(u_{0,t+1} y_{t-j}) = 0,\ j = 0, 1, \ldots k-1$ versus $\widetilde{H}_A : E(u_{0,t+1} y_{t-j}) \neq 0$ for some $j,\ j = 0, 1, \ldots k-1$

Given mild moment and memory restrictions, it turns out that, under the null,

$$m_p'(\widehat{S}_{11} + 2(1-\pi^{-1}\ln(1+\pi))\widehat{F}'\widehat{M}\widehat{S}_{22}\widehat{M}\widehat{F} - 2(1-\pi^{-1}\ln(1+\pi))(\widehat{F}'\widehat{M}\widehat{S}_{12} + \widehat{S}_{12}'\widehat{M}\widehat{F}))^{-1}m_p \xrightarrow{d} \chi^2_k,$$

where $\widehat{F} = \frac{1}{P}\sum_{t=R}^{T} X_t Y_t'$, $\widehat{M} = \left(\frac{1}{P}\sum_{t=R}^{T-1} X_t X_t'\right)^{-1}$, and

$$\begin{aligned}\widehat{S}_{11} &= \frac{1}{P}\sum_{t=R}^{T-1}(\widehat{u}_{0,t+1}Y_t - \widehat{\mu}_1)(\widehat{u}_{0,t+1}Y_t - \widehat{\mu}_1)' + \frac{1}{P}\sum_{\tau=\tau}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{u}_{0,t+1}Y_t - \widehat{\mu}_1)(\widehat{u}_{0,t+1-\tau}Y_{t-\tau} - \widehat{\mu}_1)' \\ &\quad + \frac{1}{P}\sum_{t=\tau}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{u}_{0,t+1-\tau}Y_{t-\tau} - \widehat{\mu}_1)(\widehat{u}_{0,t+1}Y_t - \widehat{\mu}_1)',\end{aligned}$$

with $\widehat{\mu}_1 = \frac{1}{P}\sum_{t=R}^{T-1}\widehat{u}_{0,t+1}Y_t$,

$$\widehat{S}_{12}' = \frac{1}{P}\sum_{\tau=0}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{u}_{0,t+1-\tau}Y_{t-\tau} - \widehat{\mu}_1)\,(X_{t-1}\widehat{u}_{0,t})' + \frac{1}{P}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(\widehat{u}_{0,t+1}Y_t - \widehat{\mu}_1)\,(X_{t-1-\tau}\widehat{u}_{0,t-\tau})',$$ and

$$\begin{aligned}\widehat{S}_{22} &= \frac{1}{P}\sum_{t=R}^{T-1}(X_{t-1}\widehat{u}_{0,t})\,(X_{t-1}\widehat{u}_{0,t})' + \frac{1}{P}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(X_{t-1}\widehat{u}_{0,t})\,(X_{t-1-\tau}\widehat{u}_{0,t-\tau})' \\ &\quad + \frac{1}{P}\sum_{\tau=1}^{l_T} w_\tau \sum_{t=R+\tau}^{T-1}(X_{t-1-\tau}\widehat{u}_{0,t-\tau})\,(X_{t-1}\widehat{u}_{0,t})',\end{aligned}$$

for $w_\tau = 1 - \frac{\tau}{l_T+1}$ and where $l_T/T^{1/4} \to 0$ as $T, l_T \to \infty$. This test is consistent against the alternative that the nesting model outperforms the nested one. The statistic can also easily be generalized in order to compare the null model against given nonlinear alternatives. This can be accomplished by using more general nonlinear test functions, such as the exponential (as in Bierens (1990)), a neural network with sigmoidal activation function, or some other generically comprehensive function (see e.g. Stinchcombe and White (1998)). In this case, we can construct nonlinear predictive accuracy tests based on $m_P = \frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\widehat{u}_{0,t+1}g(\gamma'Y_t)$, where $\gamma \in \Gamma$ is a nuisance parameter unidentified under the null hypothesis (for a detailed survey of nonlinearity tests used in economics, see Granger and Teräsvirta (1993)). If we confine attention to a finite grid of values for the nuisance parameter, γ, following the approach suggested by Lee, White and Granger (1993, LWG) in the context of (in-sample) testing for neglected nonlinearities, we can set $g(\gamma'Y_t) = \gamma'Y_t + (1+\exp(-\gamma'Y_t))^{-1}$, where γ is a $k \times 1$ vector.[5] The above distributional results then hold, with Y_t replaced by $g(\gamma'Y_t)$. Note that both the finite sample size and power in this case

[5]Different sets of weights, say γ_1 and γ_2, can be chosen for the linear and nonlinear components of the model.

depend on the specific γ which is used. Following LWG however, we can randomly draw l different sets of γ and compute l different statistics, say. Let $PV_1, \dots PV_l$ be the p-values associated with the l different statistics, so that $PV_1 \leq PV_2 \dots \leq PV_l$. LWG suggest rejecting the null at 5% if there is a $j = 1, \dots l$ such that $PV_j \leq 0.05/(l - j - 1)$. An alternative to the above approach which is discussed in the next section is to construct a test that is consistent against generic nonlinear alternatives.

3 A Predictive Accuracy Test That Is Consistent Against Generic Nonlinear Alternatives

As above, our discussion begins by specifying a simple AR(1) as our reference model, although the results discussed here generalize in a straightforward manner to the case where the reference model is a possibly nonlinear AR(p) model (see e.g. Granger and Teräsvirta (1993)). In addition, and for ease of exposition, we again confine our attention to one-step ahead forecasts, although extension to multi-step ahead forecasts follows directly. As we do not in general assume that the reference or alternative models are dynamically correctly specified, we do not explicitly write down data generating processes. Nevertheless, we can define the "true" one-step ahead forecast errors for the reference model (say model 0) and for the generic alternative model (say model 1). More precisely, let the reference model be

$$x_t = \beta_1^* + \beta_2^* x_{t-1} + u_{0,t}, \tag{6}$$

where $\beta^* = (\beta_1^*, \beta_2^*)' = \arg\min_{\beta \in B} E(f(x_t - \beta_1 - \beta_2 x_{t-1}))$, $\beta = (\beta_1, \beta_2)'$, x_t is a scalar, and in this case the same loss function, f, is used both for in-sample estimation and out-of-sample prediction evaluation. Additionally, B is a generic compact set defined on the real line. The generic alternative model is:

$$x_t = \delta_1^*(\gamma) + \delta_2^*(\gamma) x_{t-1} + \delta_3^*(\gamma) g(z^{t-1}, \gamma) + u_{1,t}(\gamma), \tag{7}$$

where $\delta^*(\gamma) = (\delta_1^*(\gamma), \delta_2^*(\gamma), \delta_3^*(\gamma))' = \arg\min_{\delta \in \Delta} E(f(x_t - \delta_1 - \delta_2 - \delta_3 g(z^{t-1}, \gamma)))$,
$\delta(\gamma) = (\delta_1(\gamma), \delta_2(\gamma), \delta_3(\gamma))'$, $\gamma \in \Gamma$, with Γ a compact subset of $\Re^d$, $z^{t-1} = (z_{1,t-1}, z_{1,t-2}, \dots)$ is a finite vector of lagged variables, possibly including lags of x_t, and g is defined as above (for example, $g(z^{t-1}, \gamma) = \exp(\sum_{i=1}^q \gamma_i \Phi(z_{t-i}))$, or $g(z^{t-1}, \gamma) = 1/(1 + \exp(c - \sum_{i=1}^q \gamma_i \Phi(z_{t-i})))$, with

$c \neq 0$ and Φ a measurable one to one mapping from $\Re$ to a bounded subset of $\Re$). In general, z^{t-1} could contain: lags of the dependent variable (when testing for neglected nonlinearity); lags of other variables (when testing for nonlinear out-of-sample Granger causality - see also Rothman, van Dijk and Franses (2001)); or both. Analogous to the DM test, and along the lines discussed in CS (2002), the hypothesis of interest is:

$$H_0 : E(f(u_{0,t+1}) - f(u_{1,t+1}(\gamma))) = 0 \text{ versus } H_A : E(f(u_{0,t+1}) - f(u_{1,t+1}(\gamma))) > 0. \tag{8}$$

Clearly, the reference model is nested within the alternative model, and given the definitions of β^* and $\delta^*(\gamma)$, the null model can never outperform the alternative. 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. It follows that H_0 and H_A can be restated as:

$$H_0 : \delta_3^*(\gamma) = 0 \text{ versus } H_A : \delta_3^*(\gamma) \neq 0,$$

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

$$E((f'(x_{t+1} - \delta_1^*(\gamma) - \delta_2^*(\gamma)x_t - \delta_3^*(\gamma)g(z^t,\gamma))) \times \begin{pmatrix} -1 \\ -x_t \\ -g(z^t,\gamma) \end{pmatrix}) = 0,$$

where f' denotes the first derivative of f with respect to its argument. Thus, under H_0 we have that $\delta_3^*(\gamma) = 0$, $\delta_1^*(\gamma) = \beta_1^*$, $\delta_2^*(\gamma) = \beta_2^*$, and $E(f'(u_{0,t+1})g(z^t,\gamma)) = 0$. Now, we can once again restate H_0 and H_A as:

$$H_0 : E(f'(u_{0,t+1})g(z^t,\gamma)) = 0 \text{ versus } H_A : E(f'(u_{0,t+1})g(z^t,\gamma)) \neq 0, \tag{9}$$

for $\forall \gamma \in \Gamma$, except for a subset with zero Lebesgue measure, which is the "generic" version of the hypotheses tested using the m_P statistic. It is thus clear that we can implement an integrated conditional moment type test (see e.g. Bierens (1982,1990) and Bierens and Ploberger (1997)). The null hypothesis in (9) corresponds to that of equal predictive ability of models (6) and (7). When $z^t = (y_t, ..., y_{t-q})$ or $z^t = (x_t, ..., x_{t-q})$, say, and when the loss function is quadratic, H_0 corresponds to correct specification of the conditional mean, given z^t. In fact, in the quadratic loss case the conditional mean is the best mean square predictor. When the loss function is a linex (i.e. $f(u) = e^{au} - au - 1$), it has been shown (see e.g. Christoffersen and Diebold (1997)) that, under conditional normality, the best predictor, given the information in z^t, is $E(x_{t+1}|z^t) + 0.5aVar(x_{t+1}|z^t)$. Here,

the joint correct specification of the conditional mean and conditional variance are implicit to the null hypothesis. When $z^t = (y_t, ..., y_{t-q})$, the null hypothesis can be interpreted as no Granger causality from y_t to x_t, in the sense that the past y_t does not help to predict x_t, either linearly or nonlinearly. Before writing down the test statistic, it is worth noting that we use an $m-$estimator in order to obtain a consistent estimator of β^*. In particular, define:

$$\widehat{\beta}_t = (\widehat{\beta}_{1,t}, \widehat{\beta}_{2,t})' = \arg\min_{\beta \in B} \frac{1}{t} \sum_{j=2}^{t} f(x_j - \beta_1 - \beta_2 x_{j-1}).$$

Also, define $\widehat{u}_{0,t+1} = x_{t+1} - \widetilde{x}_t' \widehat{\beta}_t$, where $\widetilde{x}_t = (1, x_t)'$. The test statistic is:

$$M_P = \int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma, \tag{10}$$

and

$$m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} f'(\widehat{u}_{0,t+1}) g(z^t, \gamma), \tag{11}$$

where $\int_\Gamma \phi(\gamma) d\gamma = 1$, $\phi(\gamma) \geq 0$, and $\phi(\gamma)$ is absolutely continuous with respect to Lebesgue measure. Note that $\phi(\gamma)$ is a weighting function defined over the nuisance parameter space Γ. For example, the (simple) average statistic corresponds to the case in which $\phi(\gamma)$ is uniformly distributed over Γ. Needless to say, other functionals of $m_P(\gamma)$, including $M_P^{\sup} = \sup_{\gamma \in \Gamma} |m_P(\gamma)|$ and $|M_P| = \int_\Gamma |m_p(\gamma)| \phi(\gamma) d\gamma$ can be constructed. These alternative test statistics are examined (along with M_P) in the next section.

Given mild memory, moment, smoothness and identifiability conditions, it is shown in CS that $M_P \xrightarrow{d} \int_\Gamma Z(\gamma)^2 \phi(\gamma) d\gamma$, where Z is a Gaussian process with a covariance kernel that reflects both the dependence structure of the data and, for $\pi > 0$, the effect of parameter estimation error. Hence, critical values are data dependent and cannot be tabulated. One possibility in this case is to use the upper bounds suggested by Bierens and Ploberger (1997). However, it is well known that inference based on these upper bounds is conservative. In addition, note that these bounds are not valid if we take different functionals over $m_P(\gamma)$, such as the supremum statistic, $\sup_{\gamma \in \Gamma} |m_P(\gamma)|$. Another approach for obtaining data dependent but asymptotically correct critical values is to use the bootstrap, see for example White (2000) and Corradi and Swanson (2001). An alternative approach which avoids resampling, and which CS use, is a modification of the conditional p-value approach of Hansen (1996) and Inoue (2001), where ϵ_t, η_t are iid $N(0, 1/l)$ random variables, with

$E(\epsilon_t \eta_s) = 0$, $\forall t, s$, where l plays the role of the blocksize in a block bootstrap, and where the "simulated" statistic is:

$$m_P^*(\gamma) = m_P^{*(1)}(\gamma) + m_P^{*(2)}(\gamma),$$

with

$$m_P^{*(1)}(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} \left(f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) + \widehat{\Pi} \widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T) \right), \tag{12}$$

$$\widehat{u}_{0,i+1} = x_{i+1} - \widetilde{x}_i \widehat{\beta}_i$$

and

$$m_P^{*(2)}(\gamma) = (2\widehat{\Pi} - \widehat{\Pi}^2)^{1/2} \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \eta_t \sum_{i=t}^{t+l-1} \widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T),$$

Hereafter, let $f_t(\beta) = f(x_t - \beta_1 - \beta_2 x_{t-1})$, with $f_t'(\beta)$ defined analogously. Further, $f_{t+1}(\widehat{\beta}_t) = f(\widehat{u}_{0,t+1}) = f(x_{t+1} - \widetilde{x}_t' \widehat{\beta}_t)$, with $f_{t+1}'(\widehat{\beta}_t)$ again defined analogously, and the operators $\nabla_\beta(\cdot)$, and $\nabla_\beta^2(\cdot)$ denote first and second derivatives with respect to β, respectively. Finally, $\widehat{\Pi} = (1 - \widehat{\pi}^{-1} \ln(1 + \widehat{\pi}))$, with $\widehat{\pi} = P/R$, $\widehat{F}(\widehat{\beta}_T, \gamma) = \frac{1}{T} \sum_{t=q+1}^{T} \nabla_\beta f_t'(\widehat{\beta}_T) g(z^{t-1}, \gamma)$, and $\widehat{B}(\widehat{\beta}_T)^{-1} = \left(-\frac{1}{T} \sum_{t=2}^{T} \nabla_\beta^2 f_t(\widehat{\beta}_T) \right)^{-1}$. Then, under H_0,

$$M_P^* = \int_\Gamma m_P^*(\gamma)^2 \phi(\gamma) d\gamma \xrightarrow{d^*} \int_\Gamma Z(\gamma)^2 \phi(\gamma) d\gamma, \quad a.s. - \omega,$$

where d^* denotes convergence in distribution with respect to P^*, and P^* is the probability law governing ϵ_t and η_t, conditional on the sample. Corradi and Swanson (2002) show that M_P and M_P^* have the same limiting distribution, conditional on the sample and for all samples except a set of measure zero, under H_0. Additionally, under H_A, M_P diverges to infinity at rate P while M_P^* diverges at most at rate l, conditionally on the sample and for all samples except a set of measure zero. Thus, for any independent draw of ϵ_t, η_t, $t = R, \ldots, T-1$, it suffices to compute M_P^*. By carrying out a large number of draws of ϵ_t, η_t, and forming numerous simulated statistics, the percentiles of this simulated statistic can be obtained. The decision rule in this case is to reject (do not reject) H_0 if the value of M_P which is obtained is above (equal to or below) the $(1-\alpha)th-$percentile of the simulated empirical distribution. This rule provides a test with asymptotic size equal to α, and unit asymptotic power.

4 An Empirical Example and Monte Carlo Results

The M_P test outlined in the previous section, which has power against generic nonlinear alternatives, has been examined via a series of Monte Carlo experiments in CS. However, in that paper the authors do not provide an empirical example illustrating how to implement the test in practice, and do not calibrate their Monte Carlo experiments to actual economic data. In this section we carry out a series of Monte Carlo experiments and an empirical investigation based on monthly U.S. money (M2 from January 1959 until May 2002) and output (industrial production - IP - from January 1959 until May 2002).[6,7] The dataset used is similar to that used by Amato and Swanson (2001, AS), although AS do not carry out nonlinear out of sample Granger causality tests, instead focusing on mean square forecast error based measures of alternative linear models of industrial production both with and without money as an explanatory variable. Additionally, AS use real-time data, while we do not.

To summarize, we began our analysis by fitting $VEC(p)$ models of the form

$$\Delta \log Y_t = a + b(L)\Delta \log Y_{t-1} + cZ_{t-1} + \epsilon_t,$$

where $Y_t = (IP_t, M2_t)'$, $b(L)$ is a conformably defined lag polynomial in the parameters associated with the lags of Y_t and ϵ_t is an error term. In addition, $Z_{t-1} = d \log Y_{t-1}$ is an $r \times 1$ vector of $I(0)$ variables, and r is the rank of the cointegrating space. As models with $r = 0$ were found to perform (i.e. predict) as well as models with $r \neq 0$, r is hereafter set to zero. In addition, it is

[6]These data were downloaded from the website of the Federal Reserve Bank of St. Louis. It should be noted that the data are not real-time, in the sense that a new vector of observations is not gathered at each point in time for each variable; with each vector corresponding to the entire history of observations that were available in real-time at a particular calendar date. In this way, all revisions to each variable not examined. Instead, a snapshot of revised, partially revised, and first available data taken in July of 2002 is examined. Because of this, truly ex-ante forecasts that only use information (before revision) available at a given point of time cannot be constructed. For this reason, our empirical example is meant only as an illustration of how to use the CS test. For a detailed discussion of real-time data, see Diebold and Rudebusch (1991), Swanson, Ghysels and Callan (1999), Croushore and Stark (2001), Ghysels, Swanson and Callan (2002), and the references contained therein.

[7]For simplicity, and because our empirical example is meant only as an illustration, we consider only money and output (and exclude other variables such as prices and interest rates). Additionally, keep in mind that allowing for (dynamic) misspecification under both hypotheses at the very least allows us to ensure that our critical values are asymptotically valid, even if the "right" number of lags does not enter into the null model.

assumed that each equation in the above model has the same number of lags (chosen using the Schwarz information criterion (SIC)), so that, without loss of efficiency (in the context of standard maximum likelihood estimation), we hereafter focus on generalized versions of the first equation in the system, namely,

$$\Delta \log IP_t = a + b(L)\Delta \log Y_{t-1} + \epsilon_t.$$

and the second equation in the system, namely,

$$\Delta \log M2_t = a + b(L)\Delta \log Y_{t-1} + \epsilon_t.$$

In particular, the null model was set equal to an AR(1) model either in *IP* or *M2*, with lags selected by the SIC (in all cases, a lag order of 1 was chosen). Then we set $P = 0.5T$ and constructed sequences of 1-step ahead prediction of either *IP* or *M2*. M_P type tests were then constructed using generic nonlinear functions of lags of both $M2$ and *IP* (in all cases, one lag was used) in order to check for the presence of generic nonlinear predictive ability (i.e. nonlinear out of sample Granger causality) from either *M2* to *IP* or *IP* to *M2*. In the simple case in which the causality is linear, this approach reduces to considering alternative models of the type given above for $\Delta \log IP_t$ and $\Delta \log M2_t$. However, by constructing M_P type tests, we also allow for the possibility that other (non)linear functions of the variables are useful for improving predictive performance. In particular, we set $g(z^{t-1}, \gamma) = \exp(\sum_{i=1}^{2}(\gamma_i \tan^{-1}((z_{i,t-1} - \overline{z}_i)/2\widehat{\sigma}_{z_i})))$, with $z_{1,t-1} = IP_{t-1}$, $M2_{2,t-1} = x_{t-1}$, and γ_1, γ_2 scalars. Additionally, we define $\Gamma = [0.0, 5]$x$[0.0, 5]$, and consider a grid equal to 0.1, so that overall we have 10000 (100×100) evaluation points (with the point {0,0} being omitted). The statistics M_P and $|M_P|$ are computed as simple averages over the 10000 evaluation points, while $M_P^{\sup}$ is computed as the maximum over the evaluations points. The loss function used throughout is quadratic. Finally, conditional p-values were constructed using 100 simulated statistics, and l was set equal to $\{30, 40, 50\}$.

Results based on this setup for a number of different sample periods are contained in Table 1. In the table, statistic values are reported in the first column of numerical entries, and 90^{th} percentiles of the empirical distribution of the simulated M_P^* statistics for various values of l are given in the remaining columns of numerical entries. The results that emerge upon inspection of the table are quite clear-cut. In particular, money does not appear useful for predicting output regardless of sample period.[8] This finding is in disagreement with the in-sample findings reported in Swanson

[8]Note, though, that this is a specialized finding, as the information set given as z^t includes only one lag of money

(1998) and elsewhere that models with money are preferred to the smaller models without money, and is in agreement with Chao, Corradi and Swanson (2001, CCS) who find that use of the m_P test suggests that there is actually nothing to choose between the larger model with money, and the smaller model without. We thus have evidence that the results of the linear m_P test performed by CCS carry through to the case of generic nonlinear alternatives.

In order to shed light on the robustness of this finding, we carried out a series of Monte Carlo experiments with data generating processes parameterized according to the actual data that were used in our empirical investigation. In particular, and following loosely along the lines of the experiments carried out in CS data were generated as follows:

$y_t = b_1 + b_2 y_{t-1} + u_{1,t},\ u_{1,t} \sim iidN(0, \theta_1)$

Size1: $x_t = a_1 + a_2 x_{t-1} + u_{2,t},\ u_{2,t} \sim iidN(0, \theta_2)$

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

Power1 : $x_t = c_1 + c_2 x_{t-1} + \exp(\tan^{-1}(y_{t-1}/2)) + u_{2,t}$

Power2 : $x_t = c_1 + c_2 x_{t-1} + 2\exp(\tan^{-1}(y_{t-1}/2)) + u_{2,t}$

Power3 : $x_t = c_1 + c_2 x_{t-1} + y_{t-1} + u_{2,t}$

Power4 : $x_t = c_1 + c_2 x_{t-1} + 2y_{t-1} + u_{2,t}$

Power5 : $x_t = c_1 + c_2 x_{t-1} + y_{t-1} 1\{y_{t-1} > a_1/(1-a_2)\} + u_{2,t}$

Power6 : $x_t = c_1 + c_2 x_{t-1} + 2y_{t-1} 1\{y_{t-1} > a_1/(1-a_2)\} + u_{2,t}$

Power7 : $x_t = c_1 + c_2 x_{t-1} + \exp(\tan^{-1}(y_{t-1}/2)) + a_3 u_{2,t-1} + u_{2,t}$

Power8 : $x_t = c_1 + c_2 x_{t-1} + 2\exp(\tan^{-1}(y_{t-1}/2)) + a_3 u_{2,t-1} + u_{2,t}$

Power9 : $x_t = c_1 + c_2 x_{t-1} + y_{t-1} + a_3 u_{2,t-1} + u_{2,t}$

Power10: $x_t = c_1 + c_2 x_{t-1} + 2y_{t-1} + a_3 u_{2,t-1} + u_{2,t}$

Power11: $x_t = c_1 + c_2 x_{t-1} + y_{t-1} 1\{y_{t-1} > a_1/(1-a_2)\} + a_3 u_{2,t-1} + u_{2,t}$

Power12: $x_t = c_1 + c_2 x_{t-1} + 2y_{t-1} 1\{y_{t-1} > a_1/(1-a_2)\} + a_3 u_{2,t-1} + u_{2,t},$

In order to calibrate the DGPs, we estimated models using $y_t = \Delta \log M2_t$ and $x_t = \Delta \log IP_t$ for the entire sample period of our historical data. The reference models (*Size1* and *Size2*) are AR(1) and ARMA(1,1) processes. Following our above discussion, the null hypothesis is that no competing model outperforms the reference model. The alternative models all include (non)linear

and output. However, we experimented with more lags of money and output and our findings remained the same. Note also that lags of other variables were not included, so that other variables may be useful, even if money is not. Further investigation of this possibility is left to future research.)

functions of y_{t-1}. Thus, our focus is on DGPs parameterized to be similar to models constructed when testing for (non)linear out-of-sample Granger causality from money to output. The functional forms that are specified under the alternative include: (i) exponential (*Power1, Power2*); (ii) linear (Power3, Power4); and (iii) self exciting threshold (*Power5, Power6*). In addition, *Power7-Power12* are the same as *Power1-Power6*, except that an MA(1) term is added. Notice that *Power1* and *Power2* include a nonlinear term that is similar in form to the test function, $g(\cdot)$. Also, *Power3* and *Power4* serve as linear causality benchmarks. As in our empirical example, in all experiments we set $g(z^{t-1},\gamma) = \exp(\sum_{i=1}^{2}(\gamma_i \tan^{-1}((z_{i,t-1} - \overline{z}_i)/2\widehat{\sigma}_{z_i})))$, with $z_{1,t-1} = y_{t-1}$, $z_{2,t-1} = x_{t-1}$, and γ_1, γ_2 scalars. Additionally, we again define $\Gamma = [0.0, 5]$x$[0.0, 5]$, and consider a grid equal to 0.1, so that overall we have 10000 (100×100) evaluation points (with the point {0,0} being omitted). The M_P, $|M_P|$, and $M_P^{\sup}$ statistics are also computed as discussed above, and the loss function is again quadratic. Based on estimating *Size1* using our historical data we set $b_0 = 2$, $b_1 = 0.7$, $a_1 = 2$, $a_2 = \{0.2, 0.4\}$, $c_1 = 0.0$ and $c_2 = \{0.2, 0.4\}$. Further, based on examination of the residuals from our regressions, we set $a_3 = -0.1$, and as above, conditional p-values were constructed using 100 simulated statistics, and l was set equal to $\{30, 40, 50\}$. Experiments were carried out for sample sizes of $T = 300, 600$, and 1000 observations, with $P = 0.5T$.All results are based on 5000 Monte Carlo replications.

Our findings are summarized in Tables 1-3. The first column in the tables state the model type (e.g. *Size1).* In addition, sample sizes, l values, and versions of the statistic that are reported on are given in the tables. All numerical entries represent rejection frequencies, where rejection (or not) is based on the use of the 90^{th} percentile of the empirical distribution of the simulated M_P^* statistics (see above discussion). As above, results are clear-cut. Under H_0, the empirical level of the test is rather close to the nominal 10% level, regardless of whether M_P, $M_P^{\sup}$, or $|M_P|$ is used (with values usually between 0.10 and 0.15), and there is substantive improvement as the sample grows from 300 to 600 observations, with little left to gain when moving from 600 to 1000 observations. Power also increases markedly as the sample size increases, and there appears much to gain not only when moving from 300 to 600 observations, but also when moving from 600 to 1000 observations. In addition, our findings are rather robust to the choice of the lag truncation parameter l. Overall, the results from our Monte Carlo experiments suggest that the tests perform quite well with the types of DGPs and sample sizes encountered in our empirical illustration.

5 Concluding Remarks

We discuss a number of recent advances in the literature on predictive accuracy testing, with focus on easy to apply (and general) tests such as that due to Diebold and Mariano (1995), as well as more complex tests such as that of White (2000). Our primary focus concerns the issue of (dynamic) misspecification, and we argue that it is reasonable in most forecasting contexts to assume that misspecification may be present under both hypotheses (i.e. all models being compared may be misspecified). In this case, many of the tests that have recently appeared in the literature may not be valid using standard critical values, although certain tests such as the conditional moment tests discussed in Chao, Corradi and Swanson (2001) and Corradi and Swanson (2002) are. In addition to overviewing out of sample conditional moment tests we also provide new Monte Carlo and empirical evidence on their usefulness, and find, for example, that a generic nonlinear test of predictive accuracy suggests that lags of money growth are not useful for predicting output growth.

6 References

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

Ashley, R., C.W.J. Granger, and R. Schmalensee, (1980), Advertising and Aggregate Consumption: An Analysis of Causality, *Econometrica,* 48, 1149-1167.

Bai, J., (1998), Testing Parametric Conditional Distributions of Dynamic Models, Working Paper, Boston College.

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

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

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

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

Christoffersen, P., (1998), Evaluating Interval FOrecasts, *International Economic Review*, 39, 841-862.

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

Christoffersen, P. and F.X. Diebold, (2000), How Relevant Is Volatility Forecasting for Financial Risk Management?, *Review of Economics and Statistics*, 82, 12-22.

Croushore, D. and T. Stark, (2001), A Real-Time Data Set for Macroeconomists, *Journal of Econometrics*, 105, 111-130.

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

Clements, M.P. and D.F. Hendry, (1999a), *Forecasting Economic Time Series*: *The Zeuthen Lectures on Economic Forecasting,* MIT Press: Cambridge.

Clements, M.P. and D.F. Hendry, (1999b), On Winning Forecasting Competitions in Economics, *Spanish Economic Review*, 1, 123-160.

Clements, M.P. and J. Smith, (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 J. Smith, (2001), Evaluating Multivariate Forecast Densities: A Comparison of Two Approaches, *International Journal of Forecasting,* forthcoming.

Corradi, V. and N.R. Swanson, (2001), Bootstrap Conditional Distribution Tests Under Dynamic Misspecification, Mimeo, Exeter University and Rutgers University.

Corradi, V. and N.R. Swanson, (2002), A Consistent Test for Nonlinear Out of Sample Predictive Accuracy, *Journal of Econometrics,* forthcoming.

Corradi, V., N.R. Swanson and C. Olivetti, (2001), Predictive Ability with Cointegrated Variables, *Journal of Econometrics,* 104, 315-358.

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

Diebold, F.X. and G.D. Rudebusch, (1991), Forecasting Output With the Composite Leading Index: A Real Time Analysis, *Journal of the American Statistical Association*, 86, 603-610.

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

Diebold, F.X., J. Hahn and A.S. Tay, (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.

Ghysels, E., N.R. Swanson, and M. Callan, (2002), Monetary Policy Rules with Model and Data Uncertainty, *Southern Economic Journal,* forthcoming.

Granger, C.W.J., (1969), Investigating Causal Relations by Econometric Models and Cross Spectral Methods, *Econometrica*, 37, 424-438.

Granger, C.W.J., (1980), Testing for Causality: A Personal Viewpoint, *Journal of Economic Dynamics and Control,* 2, 329-352.

Granger, C.W.J. and P. Newbold, *Forecasting Economic Time Series*, Academic Press: San Diego.

Granger, C.W.J. and T. Teräsvirta, (1993), *Modelling Nonlinear Economic Relationships*, Oxford University Press: Oxford.

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

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

Hong, Y., (2001), Evaluation of Out of Sample Probability Density Forecasts with Applications to

S&P 500 Stock Prices, Mimeo, Cornell University.

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

Inoue, A. and L. Killian, (2002), In Sample or Out of Sample Tests of Predictability: Which One Should We Use?, Mimeo, University of Michigan and North Carolina State University, Raleigh.

Lee, T.-H., H. White and C.W.J. Granger, (1993), Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests, *Journal of Econometrics*, 56, 269-290.

Li, F. and G. Tkacz, (2001), A Consistent Test for Conditional Density Functions with Time Dependent Data, Working Paper, Bank of Canada.

McCracken, M.W., (1999), Asymptotics for Out of Sample Tests of Causality, Working Paper, Louisiana State University.

McCracken, M.W., (2000), Robust Out of Sample Inference, *Journal of Econometrics,* 99, 195-223.

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

Rossi, B., 2001, Optimal Tests for Nested Model Selection with Underlying Parameter Instability, Mimeo, Duke University.

Rothman, P., D.J.C. van Dijk and P.H. Franses, (2001), A Multivariate STAR Analysis of the Relationship Between Money and Output, *Macroeconomic Dynamics*, 5, 506-532.

Stinchcombe, M. and H. White, (1998), Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative, *Econometric Theory*, 14, 295-324.

Sullivan, R., A. Timmerman and H. White, (1999), Data-Snooping, Technical Trading Rules and the Bootstrap, *Journal of Finance,* 54, 1647-1692.

Sullivan, R., A. Timmerman and H. White, (2001), Dangers of Data-driven Inference: The Case of Calendar Effects in Stock Returns, *Journal of Econometrics,* 104, 249-286.

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

Swanson, N.R., E. Ghysels and M. Callan, (1999), A Multivariate Time Series Analysis of the Data Revision Process for Industrial Production and the Composite Leading Indicator, *in Festschrift in Honor of C.W.J. Granger*, eds. R.F. Engle and H. White, Oxford University Press, Oxford.

Swanson, N.R. and H. White, (1995), A Model Selection Approach to Assessing the Information in

the Term Structure Using Linear Models and Artificial Neural Networks, *Journal of Business and Economic and Statistics*, 13, 265-275.

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

Teräsvirta, T. and H. Anderson, (1992), Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models, *Review of Journal of Applied Econometrics*, 7, 119-136.

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

West, K. and M.W. McCracken, (1998), Regression Based Tests of Predictive Ability, *International Economic Review*, 39, 817-840.

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

Table 1: Generic Nonlinear Predictive Accuracy Test Illustration – Money and Income *

Statistic	Statistic Value	90^{th}Percentiles		
		l=30	l=40	l=50
Sample=1959:2-2002:5 – Money -> IP				
M_P	8.445	22.030	16.339	13.093
M_P^{sup}	17.538	28.676	24.750	22.154
$\|M_P\|$	1.251	1.947	1.698	1.493
Sample=1959:2-2002:5 – IP -> Money				
M_P	1.173	2.961	2.274	1.726
M_P^{sup}	6.262	10.318	8.971	7.977
$\|M_P\|$	0.516	0.739	0.654	0.576
Sample=1970:1-2002:5 – Money -> IP				
M_P	0.006	0.008	0.008	0.007
M_P^{sup}	0.195	0.235	0.215	0.208
$\|M_P\|$	0.063	0.061	0.066	0.065
Sample=1970:1-2002:5 – IP -> Money				
M_P	0.016	0.007	0.008	0.006
M_P^{sup}	0.696	0.482	0.504	0.424
$\|M_P\|$	0.068	0.051	0.049	0.045
Sample=1984:1-2002:5 – Money -> IP				
M_P	0.026	0.111	0.105	0.062
M_P^{sup}	0.695	1.470	1.408	1.122
$\|M_P\|$	0.097	0.197	0.209	0.141
Sample=1984:1-2002:5 – IP -> Money				
M_P	0.055	0.056	0.049	0.032
M_P^{sup}	1.452	1.320	1.151	0.925
$\|M_P\|$	0.111	0.147	0.132	0.099
Sample=1959:2-1979:12 – Money -> IP				
M_P	0.150	0.412	0.421	0.347
M_P^{sup}	1.977	3.364	3.497	3.048
$\|M_P\|$	0.220	0.343	0.310	0.326
Sample=1959:2-2002:5 – IP -> Money				
M_P	0.007	0.061	0.055	0.051
M_P^{sup}	0.224	1.135	1.138	1.164
$\|M_P\|$	0.069	0.147	0.144	0.132

* Notes: Entries are statistic values and 90-percentile values taken from the empirical distribution of simulated statistics constructed as discussed above. The null hypothesis of the tests is equal predictive accuracy between a simple linear AR(1) model and a generic (non)linear laternative. Panels with "Money -> IP" have money as the additional variable for which marginal additional generic nonlinear predictive accuracy (relative to a simpler model with only IP). In all cases, the number of lags is selected using the Schwarz Information Criterion. Panels with "IP -> Money" are analogous to those discussed above, except that the marginal predictive content of IP for Money is being tested. All models are estimated using rolling windows of data. See above for further details.

Table 2: Monte Carlo Rejection Frequencies Based on Quadratic Loss, T=300 *

Model	l=30			l=40			l=50		
	M_P	M_P^{sup}	$\|M_P\|$	M_P	M_P^{sup}	$\|M_P\|$	M_P	M_P^{sup}	$\|M_P\|$
Panel A: $a_2 = 0.2$									
Size1	0.120	0.122	0.116	0.144	0.142	0.154	0.192	0.202	0.186
Size2	0.120	0.114	0.122	0.150	0.152	0.156	0.188	0.188	0.200
Power1	0.140	0.118	0.148	0.186	0.176	0.190	0.246	0.236	0.260
Power2	0.232	0.206	0.258	0.284	0.244	0.318	0.340	0.304	0.376
Power3	0.376	0.328	0.406	0.434	0.386	0.468	0.496	0.470	0.520
Power4	0.444	0.376	0.480	0.522	0.476	0.554	0.566	0.540	0.576
Power5	0.324	0.302	0.346	0.388	0.362	0.418	0.474	0.446	0.486
Power6	0.348	0.312	0.396	0.432	0.394	0.468	0.518	0.506	0.536
Power7	0.130	0.124	0.138	0.176	0.170	0.184	0.238	0.224	0.248
Power8	0.224	0.200	0.246	0.288	0.244	0.308	0.350	0.314	0.366
Power9	0.386	0.344	0.428	0.448	0.402	0.470	0.506	0.474	0.520
Power10	0.450	0.388	0.490	0.518	0.484	0.564	0.564	0.538	0.580
Power11	0.332	0.298	0.360	0.392	0.360	0.414	0.474	0.446	0.496
Power12	0.356	0.322	0.398	0.440	0.408	0.468	0.522	0.502	0.540
Panel B: $a_2 = 0.4$									
Size1	0.118	0.112	0.126	0.144	0.148	0.148	0.192	0.190	0.194
Size2	0.124	0.116	0.122	0.148	0.132	0.148	0.196	0.192	0.200
Power1	0.134	0.126	0.146	0.188	0.184	0.190	0.236	0.222	0.240
Power2	0.246	0.192	0.252	0.300	0.252	0.314	0.364	0.318	0.386
Power3	0.406	0.356	0.438	0.462	0.424	0.494	0.516	0.498	0.530
Power4	0.454	0.398	0.512	0.536	0.488	0.578	0.572	0.552	0.582
Power5	0.344	0.322	0.380	0.398	0.372	0.418	0.482	0.464	0.486
Power6	0.386	0.352	0.414	0.468	0.430	0.492	0.538	0.518	0.558
Power7	0.122	0.124	0.142	0.190	0.184	0.194	0.234	0.236	0.246
Power8	0.244	0.202	0.258	0.292	0.256	0.316	0.352	0.314	0.388
Power9	0.404	0.360	0.440	0.460	0.418	0.490	0.508	0.494	0.534
Power10	0.460	0.408	0.514	0.528	0.490	0.576	0.568	0.552	0.582
Power11	0.344	0.318	0.380	0.402	0.372	0.426	0.484	0.462	0.482
Power12	0.384	0.362	0.426	0.462	0.426	0.494	0.538	0.510	0.552

* Notes: All entries are rejection frequencies of the null hypothesis of equal predictive accuracy based on 10% nominal size critical values constructed using the conditional p-value approach discussed in Section 2. For all models denoted Power$i$, $i = 1, ..., 12$, data are generated with (non) linear Granger causality. In all experiments, the ex ante forecast period is of length P, which is set equal to 0.5T, where T is the sample size. All models are estimated using rolling windows of data. See above for further details.

Table 3: Monte Carlo Rejection Frequencies Based on Quadratic Loss, T=600 *

Model	l=30			l=40			l=50		
	M_P	M_P^{sup}	$\lvert M_P\rvert$	M_P	M_P^{sup}	$\lvert M_P\rvert$	M_P	M_P^{sup}	$\lvert M_P\rvert$
Panel A: $a_2 = 0.2$									
Size1	0.082	0.080	0.088	0.090	0.090	0.096	0.098	0.102	0.104
Size2	0.076	0.076	0.092	0.086	0.096	0.100	0.102	0.100	0.110
Power1	0.172	0.152	0.192	0.192	0.182	0.206	0.192	0.164	0.204
Power2	0.352	0.266	0.402	0.344	0.284	0.400	0.332	0.270	0.402
Power3	0.602	0.514	0.682	0.568	0.506	0.646	0.540	0.482	0.606
Power4	0.688	0.618	0.760	0.656	0.606	0.716	0.628	0.562	0.688
Power5	0.534	0.478	0.582	0.512	0.458	0.548	0.492	0.454	0.526
Power6	0.586	0.534	0.646	0.584	0.536	0.626	0.532	0.480	0.592
Power7	0.164	0.144	0.178	0.178	0.154	0.188	0.174	0.154	0.184
Power8	0.338	0.270	0.406	0.352	0.272	0.402	0.348	0.252	0.402
Power9	0.614	0.536	0.690	0.576	0.510	0.642	0.552	0.494	0.608
Power10	0.692	0.620	0.774	0.662	0.608	0.720	0.628	0.562	0.690
Power11	0.538	0.488	0.588	0.520	0.474	0.558	0.492	0.440	0.534
Power12	0.584	0.536	0.654	0.590	0.532	0.638	0.518	0.468	0.592
Panel B: $a_2 = 0.4$									
Size1	0.096	0.078	0.100	0.106	0.104	0.118	0.114	0.104	0.120
Size2	0.094	0.092	0.110	0.112	0.114	0.124	0.114	0.106	0.128
Power1	0.146	0.132	0.168	0.174	0.146	0.184	0.164	0.140	0.184
Power2	0.326	0.244	0.394	0.336	0.252	0.418	0.332	0.246	0.402
Power3	0.626	0.558	0.700	0.594	0.524	0.656	0.558	0.518	0.616
Power4	0.698	0.630	0.768	0.664	0.608	0.734	0.624	0.576	0.692
Power5	0.548	0.484	0.600	0.510	0.468	0.578	0.492	0.458	0.546
Power6	0.596	0.550	0.656	0.592	0.556	0.640	0.540	0.498	0.594
Power7	0.134	0.124	0.156	0.158	0.130	0.176	0.160	0.140	0.180
Power8	0.334	0.236	0.396	0.336	0.244	0.410	0.320	0.236	0.398
Power9	0.630	0.566	0.706	0.596	0.518	0.660	0.572	0.516	0.628
Power10	0.700	0.632	0.778	0.664	0.614	0.738	0.628	0.572	0.700
Power11	0.556	0.496	0.606	0.520	0.464	0.574	0.506	0.460	0.544
Power12	0.602	0.540	0.662	0.592	0.554	0.638	0.536	0.498	0.600

* See notes to Table 2.

Table 4: Monte Carlo Rejection Frequencies Based on Quadratic Loss, T=1000 *

Model	l=30			l=40			l=50		
	M_P	M_P^{sup}	$\lvert M_P\rvert$	M_P	M_P^{sup}	$\lvert M_P\rvert$	M_P	M_P^{sup}	$\lvert M_P\rvert$
Panel A: $a_2 = 0.2$									
Size1	0.096	0.090	0.102	0.104	0.104	0.108	0.098	0.104	0.110
Size2	0.096	0.092	0.102	0.110	0.112	0.110	0.108	0.112	0.110
Power1	0.302	0.224	0.348	0.296	0.244	0.354	0.290	0.252	0.358
Power2	0.562	0.432	0.676	0.570	0.428	0.652	0.544	0.422	0.630
Power3	0.806	0.738	0.868	0.770	0.706	0.848	0.746	0.706	0.824
Power4	0.836	0.780	0.892	0.810	0.754	0.880	0.802	0.746	0.860
Power5	0.750	0.680	0.812	0.720	0.662	0.782	0.706	0.648	0.766
Power6	0.762	0.730	0.816	0.728	0.704	0.792	0.730	0.676	0.774
Power7	0.260	0.206	0.330	0.274	0.218	0.322	0.272	0.226	0.344
Power8	0.552	0.432	0.680	0.562	0.438	0.660	0.538	0.428	0.640
Power9	0.810	0.748	0.874	0.778	0.724	0.864	0.762	0.696	0.840
Power10	0.840	0.788	0.892	0.822	0.754	0.882	0.808	0.746	0.856
Power11	0.754	0.670	0.812	0.728	0.668	0.790	0.708	0.648	0.772
Power12	0.764	0.730	0.822	0.728	0.698	0.802	0.726	0.680	0.776
Panel B: $a_2 = 0.4$									
Size1	0.102	0.102	0.102	0.118	0.122	0.110	0.114	0.110	0.116
Size2	0.102	0.098	0.102	0.120	0.122	0.116	0.110	0.120	0.116
Power1	0.238	0.194	0.314	0.250	0.192	0.306	0.262	0.200	0.298
Power2	0.532	0.388	0.660	0.528	0.390	0.654	0.516	0.368	0.648
Power3	0.808	0.746	0.874	0.786	0.734	0.866	0.764	0.692	0.818
Power4	0.840	0.790	0.906	0.818	0.766	0.886	0.804	0.736	0.860
Power5	0.738	0.682	0.802	0.720	0.648	0.774	0.694	0.634	0.752
Power6	0.764	0.722	0.812	0.728	0.696	0.802	0.712	0.668	0.786
Power7	0.238	0.182	0.292	0.222	0.174	0.292	0.246	0.182	0.290
Power8	0.538	0.402	0.652	0.550	0.406	0.652	0.522	0.400	0.642
Power9	0.814	0.754	0.876	0.788	0.738	0.876	0.764	0.706	0.836
Power10	0.844	0.790	0.906	0.824	0.772	0.886	0.804	0.746	0.862
Power11	0.746	0.690	0.810	0.734	0.654	0.792	0.694	0.636	0.760
Power12	0.768	0.730	0.822	0.726	0.694	0.808	0.712	0.666	0.792

* See notes to Table 2.