A Consistent Test for Nonlinear Out of Sample Predictive Accuracy*

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

[1] University of Exeter
[2] Purdue University

September 2001

Abstract

In this paper, we draw on both the consistent specification testing and the predictive ability testing literatures and propose an integrated conditional moment type predictive accuracy test that is similar in spirit to that developed in Bierens (1982,1990) and Bierens and Ploberger (1997). The test is consistent against generic nonlinear alternatives, and is designed for comparing nested models. One important feature of our approach is that the same loss function is used for in-sample estimation and out-of-sample prediction. In this way, we rule out the possibility that the null model can outperform the nesting generic alternative model. It turns out that the limiting distribution of the ICM type test statistic that we propose is a functional of a Gaussian process with a covariance kernel that reflects both the time series structure of the data as well as the contribution of parameter estimation error. As a consequence, critical values are data dependent and cannot be directly tabulated. One approach in this case is to obtain critical value upper bounds using the approach of Bierens and Ploberger (1997). Here, we establish the validity of a conditional p-value method for constructing critical values. The method is similar in spirit to that proposed by Hansen (1996) and Inoue (2001), although we additionally account for parameter estimation error. In a series of Monte Carlo experiments, the finite sample properties of three variants of the predictive accuracy test are examined. Our findings suggest that all three variants of the test have good finite sample properties when quadratic loss is specified, even for samples as small as 600 observations. However, non-quadratic loss functions such as linex loss require larger sample sizes (of 1000 observations or more) in order to ensure reasonable finite sample performance.

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

---

1 Introduction

In recent years, much attention has been given in the econometrics literature to the issue of predictive ability. One of the most important recent contributions is the seminal paper of Diebold and Mariano (DM: 1995), in which a quite general test of equal predictive accuracy between two competing models is proposed. Since then, efforts have been made to generalize DM type tests in order to: account for parameter estimation error (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 (Clements and Hendry (1999a,b, 2001) and Corradi, Swanson and Olivetti (2001)); and address the issue of joint comparison of more than two competing models (Sullivan, Timmermann and White (1999) and White (2000)). Other papers approach the issue of predictive accuracy testing via the use of encompassing and related tests (see e.g. Chao, Corradi and Swanson (2001), Clark and Mc-Cracken (2001), Harvey, Leybourne and Newbold (1997) and McCracken (1999)).[1] One of the common features of many of the papers cited above is that nonnested forecasting models are compared. However, 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 when constructing a forecasting model for another variable, hence our use of terminology such as "out-of-sample Granger causality".[2] Another feature of the above papers is that they compare a given and known set of models. More precisely, they either compare two different models or they compare a given benchmark (or reference) model with multiple alternative models. Needless to say, there may exist some other model which, although not included in the finite set of competing models, yields superior forecasts. This is a feature of predictive ability (or accuracy) tests which has been ad-

---

[1]Even though the DM paper signalled renewed interest in the area, it should be stressed that related tests had been proposed in the past (see e.g. Granger and Newbold (1986)).

[2]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, Granger concludes his discussion by noting that it is common practice to test for Granger causality using in-sample F-tests. This practice continues to be prevalent.

dressed in other areas of econometrics. For example, in the consistent specification testing literature it is customary to test the validity of a null (or reference) model against a generic alternative. This can be accomplished by constructing conditional moment tests which employ an infinite number of moment conditions (see e.g. Bierens (1982,1990), Bierens and Ploberger (1997), de Jong (1996), Hansen (1996), Lee, Granger and White (1993) and Stinchcombe and White (1998)).

In this paper, we draw on both the consistent specification and predictive ability testing literatures and propose a test for predictive accuracy which is consistent against generic nonlinear alternatives, and which is designed for comparing nested models. Broadly speaking, given a particular reference model, assume that the objective is to test whether there exists any unknown alternative model that has better predictive accuracy than the reference model, for a given loss function. A typical example is the case in which the reference model is a simple autoregressive model and we want to check whether a more accurate forecasting model can be constructed by including possibly unknown (non)linear functions of the past of the process or of the past of some other process(es).[3] Although this is the example that we focus on, the reference model can be any (non)linear model. One important feature of our approach is that the same loss function is used for in-sample estimation and out-of-sample prediction (see Granger (1993) and Weiss (1996)). In particular, we use least squares (LS) to estimate parameters when the loss function is quadratic, and use a more general $m-$estimator when the loss function is non-quadratic. In this way, we rule out the possibility that the null model can outperform the nesting generic alternative model. Thus, the null hypothesis is that the null and alternative models have equal predictive accuracy, while the alternative is that the null model is outperformed.

The most natural approach to forming the desired test is to use an out of sample version of the integrated conditional moment (ICM) test of Bierens (1982,1990) and Bierens and Ploberger (1997). One reason why we propose an ICM type rather than, say, a DM type test, is that we use the same loss function throughout.[4] To explain the problem that arises with DM type tests more clearly, note first that the difference between functionals of the "true" forecast errors is identically zero when the null model is nested. Additionally, parameter estimation error vanishes whenever the

---

[3]For example, Swanson and White (1995,1997) compare the predictive accuracy of various linear models against neural network models using both in-sample and out-of-sample model selection criteria.

[4]The applicability of the ICM test in our context, as well as the importance of using the same loss function, both in- and out-of-sample, was pointed out to us by one of the referees.

same loss function is used both in- and out-of-sample, regardless of the value of $\pi = \lim_{T \to \infty} P/R$, where P is the length of the out-of-sample evaluation period, and R is the length of the in-sample regression period used to construct the first of a sequence of P recursive ex ante forecasts. This second characteristic of DM type tests follows from West (1996), where it is shown that parameter estimation error vanishes, regardless of the value of π, when the expectation of the derivative (with respect to the parameters of the model) of the out-of-sample loss function, evaluated at the probability limit of the estimated parameters, is equal to zero. However, this is always the case when the same loss function is used both in- and out-of-sample. As a consequence, when the same loss function is used throughout, and the null model is nested, the numerator of DM type tests vanishes in probability under the null.[5]

It should perhaps be stressed that our ICM type test differs from those developed by Bierens (1982,1990) and Bierens and Ploberger (1997) in a number of respects. First, parameters are estimated recursively. Second, the test statistic is computed over the out-of-sample forecast period. Third our null hypothesis is that the reference model is the best "loss function specific" predictor, for a given information set.

It turns out that the limiting distribution of the ICM type test statistic that we propose is a functional of a Gaussian process with a covariance kernel that reflects both the time series structure of the data as well as the contribution of parameter estimation error. As a consequence, critical values are data dependent and cannot be directly tabulated. We could in principle obtain critical value upper bounds using the approach of Bierens and Ploberger (1997). However, it is well known that inferences based on upper bounds is conservative. For this reason, we also consider another approach. In particular, we propose an extension of Inoue (2001) which allows for non vanishing parameter estimation error.[6] In practice, our approach is to simply augment the simulated statistics (used in the construction of the empirical distribution) by adding an extra term which is a strong consistent estimator of the contribution of parameter estimation error to the covariance kernel of the limiting distribution of the test statistic. We establish that under the null hypothesis the actual statistic and the simulated statistic have the same limiting distribution. Under the alternative, the

---

[5]Note that McCracken (1999) shows that a particular version of the DM test in which a null model is compared against a fixed alternative and the numerator is multiplied by $\sqrt{P}$ has a nonstandard limiting distribution. However, his approach does not apply in our case, as we consider generic alternatives.

[6]Inoue (2001) is an extension of the conditional p-value approach of Hansen (1996) to the case of non martingale difference errors.

actual statistic diverges to infinity at rate P, while the simulated statistic grows at most at rate l, where l plays the same role as the block length in the block bootstrap (e.g. see Künsch (1989)). Thus, conditional on the sample, we can generate many simulated statistics, and for each of these compute the relevant functional over γ, the nuisance parameter over which integration is done. The empirical distribution of these functionals can then be used to construct critical values for the test. In particular, we reject (do not reject) the null hypothesis whenever we obtain a value above (below) the $(1-\alpha)^{th}-$percentile, thereby ensuring that the test has asymptotic size equal to α, and unit asymptotic power.

The rest of the paper is organized as follows. Section 2.1 describes the set up and outlines the assumptions needed in the rest of the paper. Section 2.2 examines the asymptotic behavior of the test statistic and establishes the asymptotic validity of the conditional p-value approach. The findings from a series of Monte Carlo experiments are reported in Section 3, and Section 4 contains concluding remarks. All proofs are collected in an Appendix.

2 Comparing the Predictive Accuracy of the Linear Model Against Generic Nonlinear Alternatives

2.1 Set-up and Assumptions

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, say based on the application of in-sample nonlinearity tests such as the Ramsey (1969) RESET test, the Luukkonen, Saikkonen, and Teräsvirta (1988) smooth transition autoregressive test, or the Lee, White and Granger (1993) neural network test (for a detailed discussion of numerous nonlinearity tests, see Granger and Teräsvirta (1993)).

The debate on the usefulness of linear versus nonlinear models discussed in the previous paragraph is addressed, for example, in Swanson and White (1997), who find that adaptive (vector) autoregressive models are not outperformed (in a predictive sense) by more sophisticated adaptive neural network models. Teräsvirta and Anderson (1992), on the other hand, find some evidence

that smooth transition autoregressive models can be useful for characterizing nonlinearities in business cycles. Regardless of the outcome of this debate, however, it is natural to use linear models as benchmarks when evaluating the usefulness of more general (non)linear models. We follow suit, and use a simple autoregressive model as our reference model.[7] Further, we confine our attention to one-step ahead forecasts. 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:

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

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, f is the loss function used both for in-sample estimation and out-of-sample prediction evaluation, and B is a generic compact set defined on the real line. The generic alternative model is:

$$x_t = \delta_1^*(\gamma) + \delta_2^*(\gamma) x_{t-1} + \delta_3^*(\gamma) g(z^{t-1}, \gamma) + u_{1,t}(\gamma), \tag{2}$$

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. The alternative model is called "generic" because of the presence of $g(z^{t-1}, \gamma)$, 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). Examples of $g(z^{t-1}, \gamma)$ include: $g(z^{t-1}, \gamma) = \exp(\sum_{i=1}^q \gamma_i \Phi(z_{t-i}))$, 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 Granger causality)[8]; or both. Note also that Assumptions **A2** (below) ensures that β^* and $\delta^*(\gamma)$ are uniquely identified. The hypotheses of interest are:

$$H_0 : E(f(u_{0,t+1}) - f(u_{1,t+1}(\gamma))) = 0 \text{ versus } H_A : E(f(u_{0,t+1}) - f(u_{1,t+1}(\gamma))) > 0. \tag{3}$$

---

[7]The results obtained in the sequel 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)).

[8]In a recent interesting paper Rothman, van Dijk and Franses (2000) discusses alternative approaches to nonlinear Granger causality testing.

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, as long as the errors above are loss function specific forecast errors. (Below, we discuss how to form appropriate recursively estimated forecast errors in our context.) 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 $E(f'(u_{0,t+1})g(z^t, \gamma)) = 0$. Thus, we can once again restate H_0 and H_A as:

$$H_0 : E(f'(u_{0,t+1})g(z^t, \gamma)) = 0 \text{ versus } H_A : E(f'(u_{0,t+1})g(z^t, \gamma)) \neq 0, \tag{4}$$

for $\forall \gamma \in \Gamma$, except for a subset with zero Lebesgue measure.

It is now clear that we can implement an integrated conditional moment type test. The null hypothesis in (4) corresponds to that of equal predictive ability of model (1) and (2). 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 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:

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

6

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:

$$M_P = \int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma, \tag{5}$$

where $\int_\Gamma \phi(\gamma) d\gamma = 1$, and $\phi(\gamma) \geq 0$, and absolutely continuous with respect to Lebesgue measure.

$$m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} f'(\widehat{u}_{0,t+1}) g(z^t, \gamma). \tag{6}$$

Two comments are worth making at this point. First, in this setup, R is defined to be the length of the "initial" in sample-period, and P is the length of the out-of-sample period, where the sample size is $T = R + P$. We begin by estimating the reference model using the first R observations, and computing a single one-step ahead forecast error, $\widehat{u}_{0,R+1}$, where $\widehat{u}_{0,R+1} = x_{R+1} - \widetilde{x}_R'\widehat{\beta}_R$. The forecasting model is then re-estimated using $R + 1$ observations, and a new "real-time" forecast error is constructed. This process is continued until the entire sample is exhausted, and a sequence of P one-step ahead forecast errors is constructed. Second, Monte Carlo experiments reported on in Section 3 examine other functionals of $m_P(\gamma)$, including $M_P^{\sup} = \sup_{\gamma \in \Gamma} |m_P(\gamma)|$ and $|M_P| = \int_\Gamma |m_p(\gamma)| \phi(\gamma) d\gamma$.

Hereafter, let $f_t(\beta) = f(x_t - \beta_1 - \beta_2 x_{t-1})$, with $f_t'(\beta)$ defined analogously. Further, let $f_{t+1}(\widehat{\beta}_t)$ $= f(\widehat{u}_{0,t+1}) = f(x_{t+1} - \widetilde{x}_t'\widehat{\beta}_t)$, with $f_{t+1}'(\widehat{\beta}_t)$ again defined analogously. Finally, the operators $\nabla_\beta(\cdot)$, and $\nabla_\beta^2(\cdot)$ denote first and second derivatives with respect to β, respectively. In the sequel, the following assumptions are used.

Assumption A1: (i) (x_t, z_t) is a strictly stationary and absolutely regular strong mixing sequence with size $-4(4 + \psi)/\psi$, $\psi > 0$, (ii) f is three times continuously differentiable in β, over the interior of B, and $\nabla_\beta f, \nabla_\beta^2 f, \nabla_\beta f', \nabla_\beta^2 f'$ are $2r-$dominated[9] uniformly in B, with $r \geq 2(2 + \psi)$, (iii) $E\left(-\nabla_\beta^2 f_t(\beta)\right)$ is negative definite, uniformly in B, (iv) g is a bounded, twice continuously differentiable function on the interior of Γ and $\nabla_\gamma g(z^t, \gamma)$ is bounded uniformly in Γ and (v) $\nabla_\gamma \nabla_\beta f_t'(\beta) g(z^{t-1}, \gamma)$ is continuous on $B \times \Gamma$, Γ a compact subset of R^d and is $2r-$dominated uniformly in $B \times \Gamma$, with $r \geq 2(2 + \psi)$.

Assumption A2: (i) $E(f'(x_t - \beta_1 - \beta_2 x_{t-1})) > E(f'(x_t - \beta_1^* - \beta_2^* x_{t-1}))$, $\forall \beta \neq \beta^*$ and (ii) $E(f'(x_t - \delta_1 - \delta_2 x_{t-1} - \delta_3 g(z^{t-1}, \gamma))) > \inf_\gamma E(f'(x_t - \delta_1^*(\gamma) - \delta_2^*(\gamma) x_{t-1} - \delta_3^*(\gamma) g(z^{t-1}, \gamma)))$ for $\delta \neq \delta^*(\gamma)$.

---

[9]Let $\nabla_\beta f(x_t - \beta_1 - \beta_2 x_{t-1})_i$ be the $i-th$ element of $\nabla_\beta f(x_t - \beta_1 - \beta_2 x_{t-1})_i$, $i = 1, 2$. By $r-$domination, we mean that $\sup_{\beta \in B} |\nabla_\beta f(x_t - \beta_1 - \beta_2 x_{t-1})|_i \leq h(x_t)$, with $E((h(x_t))^r) < \infty$.

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

Assumption A4: For any t, s; $\forall\ i, j, k = 1, 2$; and for $\Delta < \infty$:

(i) $E\left(\sup_{\beta \times \gamma \times \gamma^+ \in B \times \Gamma \times \Gamma} \left| f_t'(\beta) g(z^{t-1}, \gamma) \nabla_\beta^k f_s'(\beta) g(z^{s-1}, \gamma^+) \right|^4 \right) < \Delta,$

where $\nabla_\beta^k(\cdot)$ denotes the $k-$th element of the derivative of its argument with respect to β.

(ii) $E\left(\sup_{\beta \in B} \left| \left(\nabla_\beta^k (\nabla_\beta^i f_t(\beta)) \nabla_\beta^j f_s(\beta) \right) \right|^4 \right) < \Delta$, and

(iii) $E\left(\sup_{\beta \times \gamma \in B \times \Gamma} \left| \left(f_t'(\beta) g(z^{t-1}, \gamma) \nabla_\beta^k (\nabla_\beta^j f_s(\beta)) \right) \right|^4 \right) < \Delta.$

The assumptions stated above are essentially memory, moment, smoothness and identifiability conditions. **A1** requires (x_t, z_t) to be strictly stationary and absolutely regular.[10] This memory condition is stronger than $\alpha-$mixing, but weaker than (uniform) $\phi-$mixing. In addition to the smoothness conditions placed on f and g, the absolute regularity assumption is needed to establish stochastic continuity along the lines of Hansen (1996). Additionally, the differentiability of the loss function assumed in **A1** is a rather standard assumption (see e.g. West (1996)), although it rules out interesting cases such as comparisons based on non-differentiable loss functions. Finally, we assume that g is differentiable, thus ruling out threshold type nonlinearities, for example. However, this does not mean that our test has no power against data generated according to threshold nonlinear processes (see Monte Carlo results for further details). Note that all the examples for g outlined above satisfy **A1(iv)**. **A2** is a standard unique identifiability condition for $m-$estimators in the case of strictly stationary observations. A sufficient condition for **A2** to hold is that f be strictly convex, which is the case when the loss is quadratic or linear exponential (linex), for example. (Linex loss is defined as $f(u) = e^{au} - au - 1$, when $a > 0$ $(a < 0)$ positive (negative) errors are more (less) costly than negative (positive) errors).[11] In the Monte Carlo analysis reported below, we consider both quadratic and linex loss. **A3** states the allowable rate of growth for the length of the regression period, R, and the prediction period, P. We require P to grow at a rate not faster than R. Finally, **A4** is a uniform moment condition which is used in showing that the limiting distribution of the simulated statistic and the limiting distribution of the actual statistic have the same covariance structure, uniformly in γ.

---

[10]The definition of $\beta-$mixing coefficient first appeared in a paper by Wolkonski and Rozanov (1959). For a detailed treatment of $\beta-$mixing and absolute regularity see also the Doukhan (ch.1.1 and 1.2 (1995)).

[11]Prediction using linex loss function is studied in Zellner (1993), Christoffersen and Diebold (1996, 1997) and Granger (1999), for example.

2.2 Asymptotic Results

We begin by obtaining the limiting distribution of $\int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma$.

Theorem 1: Let A1, A2 and A3 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 (6) and Z is a Gaussian process with covariance kernel given by:

$$
\begin{aligned}
K(\gamma_1, \gamma_2) &= S_{ff}(\gamma_1, \gamma_2) + 2\Pi F(\beta^*, \gamma_1)' B(\beta^*)^{-1} S_{hh} B(\beta^*)^{-1} F(\beta^*, \gamma_2) + \Pi F(\beta^*, \gamma_1)' B(\beta^*)^{-1} S_{fh}(\gamma_2) \\
&\quad + \Pi F(\beta^*, \gamma_2)' B(\beta^*)^{-1} S_{fh}(\gamma_1),
\end{aligned}
$$

with $F(\beta^*, \gamma) = E(\nabla_\beta(f'_{t+1}(\beta^*)g(z^t, \gamma)))$, $B(\beta^*)^{-1} = (-E(\nabla^2_\beta f(\beta^*)))^{-1}$,

$S_{ff}(\gamma_1, \gamma_2) = \sum_{j=-\infty}^{\infty} E(f'_q(\beta^*)g(z^{q-1}, \gamma_1)f'_{q+j}(\beta^*)g(z^{q+j-1}, \gamma_2))$,

$S_{hh} = \sum_{j=-\infty}^{\infty} E(\nabla_\beta f_1(\beta^*)\nabla_\beta f_{1+j}(\beta^*)')$,

$S_{fh}(\gamma_1) = \sum_{j=-\infty}^{\infty} E(f'_q(\beta^*)g(z^{q-1}, \gamma_1)\nabla_\beta f_{q+j}(\beta^*)')$,

$\Pi = 1 - \pi^{-1}\ln(1 + \pi)$, for $\pi > 0$ and $\Pi = 0$ for $\pi = 0$, $z^q = (z_1, ..., z_q)'$, and γ, γ_1, γ_2 are generic elements of Γ.

(ii) Under H_A, for $\varepsilon > 0$ and $\delta < 1$,

$$\lim_{P \to \infty} \Pr\left(\frac{1}{P^\delta} \int_\Gamma m_P(\gamma)^2 \phi(\gamma) d\gamma > \varepsilon\right) = 1.$$

When f is a quadratic function, it follows that $\widehat{\beta}_t$ is the LS estimator, computed using observations up to time t, $m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \widehat{u}_{0,t+1} g(z^t, \gamma)$, $F(\beta^*, \gamma) = E\left(-\widetilde{x}'_t g(z^t, \gamma)\right)$, $B(\beta^*)^{-1} = (-E(\widetilde{x}_t \widetilde{x}'_t))^{-1}$, and $\nabla_\beta f_{t+1}(\beta^*) = \widetilde{x}_t u_{0,t+1}$, for example. Note also 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, 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), as $K(\gamma_1, \gamma_2)$ satisfies their Lemma 1. Thus, from their Theorem 7, it follows that:

$$\lim_{P \to \infty} \Pr\left(M_P > \varepsilon \int_\Gamma \widehat{K}(\gamma, \gamma) \phi(\gamma) d\gamma\right) \leq \Pr(\overline{W} > \varepsilon),$$

where $\overline{W} = \sup_{s \geq 1} \frac{1}{s} \sum_{j=1}^{s} \xi_j^2$, with ξ_j an i.i.d. $N(0, 1)$ random variable, and where $\widehat{K}(\gamma, \gamma)$ is a consistent estimator of $K(\gamma, \gamma)$. For example, we can set:

$$\widehat{K}(\gamma, \gamma) = \widehat{S}_{ff}(\gamma, \gamma) + 2\widehat{\Pi}\widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \widehat{S}_{hh} \widehat{B}(\widehat{\beta}_T)^{-1} \widehat{F}(\widehat{\beta}_T, \gamma) + 2\widehat{\Pi}\widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \widehat{S}_{fh}(\gamma),$$

where $\widehat{\Pi} = (1 - \widehat{\pi}^{-1} \ln(1 + \widehat{\pi}))$, with $\widehat{\pi} = P/R$, $\widehat{F}(\widehat{\beta}_T, \gamma) = \frac{1}{T} \sum_{t=q+1}^{T} \nabla_\beta f_t'(\widehat{\beta}_T) g(z^{t-1}, \gamma)$, $\widehat{B}(\widehat{\beta}_T)^{-1} = \left(-\frac{1}{T} \sum_{t=2}^{T} \nabla_\beta^2 f_t(\widehat{\beta}_T) \right)^{-1}$,

$\widehat{S}_{ff}(\gamma, \gamma) = \frac{1}{P} \sum_{t=R}^{T-1} (f'(\widehat{u}_{0,t+1}) g(z^t, \gamma))^2 + \frac{2}{P} \sum_{\tau=1}^{b_T} w_\tau \sum_{t=\tau+R}^{T-1} f'(\widehat{u}_{0,t+1}) g(z^t, \gamma) f'(\widehat{u}_{0,t-\tau+1}) g(z^{t-\tau}, \gamma)$,

$\widehat{S}_{hh} = \frac{1}{P} \sum_{t=R+1}^{T} \nabla_\beta f_t(\widehat{\beta}_T) \nabla_\beta f_t(\widehat{\beta}_T)' + \frac{2}{P} \sum_{\tau=1}^{b_T} w_\tau \sum_{t=\tau+R}^{T} \nabla_\beta f_t(\widehat{\beta}_T) \nabla_\beta f_{t-\tau}(\widehat{\beta}_T)'$,

$\widehat{S}_{fh}(\gamma) = \frac{1}{P} \sum_{t=R}^{T-1} f'(\widehat{u}_{0,t+1}) g(z^t, \gamma) \nabla_\beta f_t(\widehat{\beta}_T) + \frac{2}{P} \sum_{\tau=1}^{b_T} w_\tau \sum_{t=\tau+R}^{T-1} f'(\widehat{u}_{0,t+1}) g(z^t, \gamma) \nabla_\beta f_{t-\tau}(\widehat{\beta}_T)$,

with w_τ an appropriately defined weighting function (e.g. $w_\tau = 1 - \frac{\tau}{b_T+1}$) and b_T an appropriately defined lag truncation parameter (e.g. such that b_T grows with T and $b_T = o(T^{1/4})$).

Under A1-A4, by an argument similar to that used in the proof of Lemma A in the appendix, it follows that $\widehat{K}(\gamma, \gamma) - K(\gamma, \gamma) \xrightarrow{pr} 0$, uniformly in γ.[12] Bierens and Ploberger (1997, pp. 1144) provide the 10%, 5% and 1% quantiles of $\overline{W}$. The decision rule in this case is to reject H_0 at the 5% level if:

$$M_P > 4.26 \int_\Gamma \widehat{K}(\gamma, \gamma) \phi(\gamma) d\gamma.$$

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, $\sup_{\gamma \in \Gamma} |m_P(\gamma)|$.

One approach for obtaining data dependent but asymptotically correct critical values is to use the bootstrap. For example, White (2000) considers the case in which the statistic does not depend on γ, and establishes the validity of the Politis and Romano (1994) stationary bootstrap, when parameter estimation error vanishes asymptotically.[13] Additionally, the validity of the block bootstrap has been established for the case where test statistics are constructed using functionals of empirical processes, when parameter estimation error (PEE) does not vanish (i.e. see Corradi and Swanson (2001)). An alternative approach which avoids resampling, and which we consider here, is the conditional p-value approach of Hansen (1996). Recently Inoue (2001) has extended this approach to allow for non-martingale difference score functions. However, Inoue's approach does not take into account non vanishing PEE. For our purposes, it suffices to extend Inoue's result to the case of nonvanishing PEE.

Let ϵ_t, η_t be i.i.d. $N(0, 1/l)$ random variables, with $E(\epsilon_t \eta_s) = 0$, $\forall t, s$ and where l plays the role of the blocksize in a block bootstrap, or equivalently of the lag truncation parameter in the estimation

---

[12]In Lemma A we provide an almost sure result, so that we need $b_T \log T/T^{1/4} \to 0$ as $T \to \infty$.

[13]White's result has also been shown to be valid in the case of cointegrated variables (i.e. see Corradi, Swanson and Olivetti (2001)).

of the HAC covariance matrix. The appropriate "simulated" statistic is:

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

where

$$m_P^{*(1)}(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} \left(f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) + \widehat{\Pi} \widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T) \right) \qquad (7)$$

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

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

with $\widehat{\Pi}$, $\widehat{F}(\widehat{\beta}_T, \gamma)'$, $\widehat{B}(\widehat{\beta}_T)^{-1}$ defined as above.[14]

In order to motivate the use of this statistic, it is first worth considering the case in which we set $l = 1$ and $\Pi = 0$. In particular, consider the case where the block length is unity, and there is no parameter estimation error ($\Pi = 0$). This is the same framework used in Hansen (1996), and in this case we can write $m_P^*(\gamma) = m_P^{*(1)}(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} \epsilon_t \left(f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) \right)$, as $m_P^{*(2)}(\gamma) = 0$. Note that conditional on the sample, $m_P^*(\gamma)$ is a zero mean normal random variable, with variance equal to $\frac{1}{P} \sum_{t=R}^{T-1} \left(f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) \right)^2$, which converges almost surely to the second moment of $f'(u_{0,i+1}) g(z^i, \gamma)$, which in turn only coincides with the long run variance of $m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} f'(\widehat{u}_{0,t+1}) g(z^t, \gamma)$ (i.e. $K(\gamma, \gamma)$) when $f'(u_{0,i+1}) g(z^i, \gamma)$ is a martingale difference sequence and when there is no parameter estimation error. Now, consider a generic value for l. That is, consider $m_P^*(\gamma) = m_P^{*(1)}(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} \left(f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) \right)$. In this case, conditional on the sample, $m_P^*(\gamma)$ is a zero mean normal random variable, with variance equal to $\frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) \right)^2$, which converges almost surely to $S_{ff}(\gamma, \gamma)$, as defined in the statement of Theorem 1. However, $S_{ff}(\gamma, \gamma)$ only coincides with $K(\gamma, \gamma)$ when there is no parameter estimation error. This, in turn, suggests accounting for parameter estimation error by writing $m_P^*(\gamma) = m_P^{*(1)}(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} \left(f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) + \widehat{\Pi} \widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T) \right)$. Here, we capture data dependence and parameter estimation error. However, the variance of the simulated statistic in this case converges to some positive definite kernel which differs from $K(\gamma, \gamma)$ by one constant term. In particular, the second term in the definition of $K(\gamma, \gamma)$ is multiplied by Π^2 instead of 2Π. This explains why we need to define $m_P^{*(2)}(\gamma)$, which serves the desired role, given that ϵ_t and η_t are uncorrelated.

---

[14]Note that the results reported below also hold when $\widehat{F}(\widehat{\beta}_T, \gamma)'$, $\widehat{B}(\widehat{\beta}_T)^{-1}$ are replaced with $\widehat{F}(\widehat{\beta}_t, \gamma)'$, $\widehat{B}_t(\widehat{\beta}_t)^{-1}$.

To recap how we account for parameter estimation error in the simulated statistic, notice that conditionally on the sample, and for all samples except a set of measure zero,

$$\frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) \sim N\left(0, \frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} f'(\widehat{u}_{0,i+1}) g(z^i, \gamma)\right)^2\right)$$

and

$$\frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} f'(u_{0,i+1}) g(z^i, \gamma) \sim N\left(0, \frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} f'(u_{0,i+1}) g(z^i, \gamma)\right)^2\right).$$

Under the assumptions given above, and as $P \to \infty$, these two variance expressions converge to the same limit, for all samples except a set of measure zero. Thus, the effect of parameter estimation error in the simulated statistic would vanish if the second term were not added to $m_P^{*(1)}(\gamma)$. However, the second term in (7) only provides a consistent estimator of the contribution of parameter estimation error up to a constant depending on Π. To overcome this additional problem we have added $m_P^{*(2)}(\gamma)$ to $m_P^*(\gamma)$, where $m_P^{*(2)}(\gamma)$ is normally distributed, and is independent of $m_P^{*(1)}(\gamma)$, conditional on the sample and for all samples except a set of measure zero. This result is summarized in the following theorem:

Theorem 2: Let A1-A4 hold. Also, assume that as $P \to \infty$, $l \to \infty$ and $\frac{l(\log P)^{1/2}}{P^{1/4}} \to 0$. Then, conditional on the sample and for all samples except a set of measure zero, the following results hold: (i) Under H_0,

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

where d^* denotes convergence in distribution with respect to P^*, P^* is the probability law governing ϵ_t and η_t, conditional on the sample, Z is a Gaussian process with the same covariance as that given in Theorem 1, and $a.s. - \omega$ means conditional on the sample, and for all samples excpet a set of measure zero.

(ii) Under H_A, $M_P^* = O_{P^*}(l)$, $a.s. - \omega$.

Thus, 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. 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. For any independent draw of ϵ_t, η_t, $t = R, \ldots, T-1$, it thus suffices to compute M_P^*. By carrying out a large number of draws of ϵ_t, η_t, 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

12

is obtained is above (equal to or below) the $(1 - \alpha)th-$percentile. This rule provides a test with asymptotic size equal to α, and unit asymptotic power.

3 Monte Carlo Results

In this section we carry out a series of Monte Carlo experiments using data generated as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 functions of y_{t-1}. Thus, our focus is on (non)linear out-of-sample Granger causality testing. The functional forms that are specified under the alternative include: (i) exponential (*Power1,Power2*); (ii) linear (Power3,Power4); and (iii) self exciting threshold (*Power5,Power6*). In addition, *Power7-Power12* are the same as *Power1-Power6*, except that an MA(1) term is added. Notice that *Power1* and *Power2* include a nonlinear term that is similar in form to the test function, $g(\cdot)$. Also, *Power3* and *Power4* serve as linear causality benchmarks. In all experiments, we set $g(z^{t-1}, \gamma) = \exp(\sum_{i=1}^{2}(\gamma_i \tan^{-1}((z_{i,t-1} - \bar{z}_i)/2\hat{\sigma}_{z_i})))$, with $z_{1,t-1} = y_{t-1}$, $z_{2,t-1} = x_{t-1}$, and γ_1, γ_2 scalars. Additionally, define $\Gamma = [0.0, 5] \times [0.0, 5]$ (overall,

1000 grid points on this interval are evaluated, with the point $\{0,0\}$ being omitted). We consider two loss functions, namely: (i) quadratic loss and (ii) linex loss (i.e. $f(c) = \exp(ac) - ac - 1$). We set $a = 1$. Note that for quadratic loss, when the DGPs are as in *Size1 and Size2,* the best predictor is the conditional mean, i.e. $a_1 + a_2 x_t$; on the other hand, for the linex loss the best predictor is $E(x_{t+1}|x_t) + 0.5 Var(x_{t+1}|x_t)$,[15] that is, $a_1 + 0.5 + a_2 x_t$ for *Size1*, and $a_1 + 0.5(1 + a_3) + a_2 x_t$ for *Size2*. Also note that the $m-$estimator for the slope parameter converges to a_2, while the intercept estimators converges for $a_1 + 0.5$ and $a_1 + 0.5(1 + a_3)$, respectively for *Size1* and *Size2*. All results are based on 500 Monte Carlo replications, and samples of T=400, T=600, and T=1000 observations are used, although only results for the latter two samples are reported. In addition, the following parametrizations are used: $a_1 = 0.1$, $a_2 = \{0.2, 0.4\}$, and $a_3 = 0.3$. Finally, conditional p-values are constructed using 100 simulated statistics, and l is set equal to $\{30, 40, 50\}$ (see Inoue (2001) for further discussion of the choice of l in the current context).

Our findings are summarized in Tables 1-2 (quadratic loss) and Table 3 (linex loss). The first two columns in the tables state the model type (e.g. *Size1*). In addition, sample sizes, l values, and versions of the $m_P(\gamma)$ statistic that are reported on are given in the tables. Finally, all numerical entries represent rejection frequencies, and correspond to the case where P is set equal to $0.5T$. Additional results for $T = 400$, and for $P = 0.4T$ and $P = 0.6T$ are qualitatively the same as those reported, and are available upon request from the authors. Results based on quadratic loss are fairly clear-cut, as is evidenced by inspection of Tables 1 and 2. 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). Also, the finite sample rejection frequency is high (usually above 0.60) when $T = 600$, and approaches unity quite rapidly, as evidenced by much higher rejection frequencies when T is increased to 1000. Overall, our findings are rather robust to the choice of the lag truncation parameter l. Results for the linex loss function are reported in Table 3. We report only the results for M_P^{sup}. It is immediate to notice that the findings for linex loss are not particularly encouraging. The rejection frequencies in the experiments carried out under H_0 are between $0.20 - 0.25$, and do not improve substantially when we increases the sample from 600 to 1000 observations. On the other hand, rejection frequencies under the alternative increase substantially with the sample size. However, observed rejection frequencies under the alternative of

---

[15]see e.g. Christoffersen and Diebold (1997).

between $0.5-0.7$ are not unusual for $T = 1000$, suggesting that very large samples are needed before these rejection frequencies will approach unity.[16] The unsatisfactory finite sample performance of the test in the linex loss case is mainly attributable to the fact that the $m-$estimator converges to its probability limit very slowly. This suggests that the use of other non-quadratic loss functions should be viewed with caution when samples are small. In summary, then, while the tests appear to perform very well under quadratic loss, somewhat larger samples are needed under linex loss in order to ensure reasonable finite sample test performance.[17]

4 Concluding Remarks

In this paper we have drawn from the literatures on predictive accuracy and consistent specification testing in order to develop a test for out of sample predictive accuracy which is consistent against generic (non)linear alternatives. In particular, we outline an out-of-sample variant of the Bierens ICM test whose limiting distribution is a functional of a Gaussian process with a covariance kernel that captures both the time series structure of the data as well as the contribution of parameter estimation error (PEE). As critical values are data dependent, we develop a conditional p-value procedure which extends the work of Inoue (2001) by allowing for non vanishing PEE. In a series of Monte Carlo experiments, we examine versions of the test that are valid in the case of quadratic and nonquadratic loss functions. In particular, we evaluate the finite sample performance of the test when forecasts are compared using mean square error and linex loss functions. Our focus in these experiments is on (non)linear out of sample Granger causality, and our findings show that the test has good finite sample properties, at least in the case of quadratic loss.

---

[16]This is not surprising, as the exponential term in the linex loss function blows up the variance of the statistic, and therefore slows down the speed with which the law of large numbers and the functional central limit theorem work.

[17]An interesting experiment which we leave to future research is the comparison of the finite sample properties of the test when critical values are constructed using our conditional p-value approach with the finite sample properties when critical values are constructed using the upper bounds approach of Bierens and Ploberger (1997).

5 Appendix

Recall from above that $f_t(\beta) = f(x_t - \beta_1 - \beta_2 x_{t-1})$, with $f'_t(\beta)$ and $f''_t(\beta)$ defined analogously; $f_{t+1}(\widehat{\beta}_t) = f(\widehat{u}_{0,t+1}) = f(x_{t+1} - \widetilde{x}'_t \widehat{\beta}_t)$, with $f'_{t+1}(\widehat{\beta}_t)$ and $f''_{t+1}(\widehat{\beta}_t)$ again defined analogously; and the operators $\nabla_\beta(\cdot)$, and $\nabla^2_\beta(\cdot)$ denote first and second derivatives with respect to β. In addition, let $\|A\| = (tr(A'A))^{1/2}$ denote the Euclidian norm of a matrix, A, and let $\|A\|_p = (E((tr(A'A))^{1/2})^p)^{1/p}$ denote the L^p-norm of a random matrix. For a vector b, $|b|$ denotes the vector whose elements are the absolute values of the elements of b. Further, $\rightarrow$ denotes the limit as $P \rightarrow \infty$ and $\gamma, \gamma', \gamma^+, \gamma_1$, and γ_2 are all generic elements of Γ. Also, $\Delta, \widetilde{\Delta}$ are constants, which may take different values depending on the context in which they are used.

Proof of Theorem 1: (i) We first show convergence in distribution pointwise in γ. By a second order Taylor expansion around β^*,

$$
\begin{aligned}
m_P(\gamma) &= \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} f'_{t+1}(\beta^*) g(z^t, \gamma) - \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta^*)' \nabla_\beta f'_{t+1}(\beta^*) g(z^t, \gamma) \\
&\quad + 0.5 \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} (\widehat{\beta}_t - \beta^*)' \nabla^2_\beta f'_{t+1}(\overline{\beta}_t) g(z^t, \gamma)(\widehat{\beta}_t - \beta^*),
\end{aligned} \tag{8}
$$

where $\overline{\beta}_t \in (\widehat{\beta}_t, \beta^*)$. Now, note that,

$$
(\widehat{\beta}_t - \beta^*) = \left(-\frac{1}{t} \sum_{j=2}^{t} \nabla^2_\beta f_j(\overline{\beta}_t) \right)^{-1} \left(\frac{1}{t} \sum_{j=2}^{t} \nabla_\beta f_j(\beta^*) \right) = \widehat{B}(\overline{\beta}_t)^{-1} H_t.
$$

Let $F(\beta, \gamma) = E\left(\nabla_\beta f'_{t+1}(\beta) g(z^t, \gamma)\right)$ and $(E(\widehat{B}(\beta)))^{-1} = B(\beta)^{-1}$. We begin by showing that the last term on the RHS of equation (8) is $o_p(1)$, uniformly in γ. Note that the absolute value of this term is majorized by,

$$
\left| \sup_{t \geq R} P^{1/4}(\widehat{\beta}_t - \beta^*)' \frac{1}{P} \sum_{t=R}^{T-1} \nabla^2_\beta f'_{t+1}(\overline{\beta}_t) g(z^t, \gamma) \sup_{t \geq R} P^{1/4}(\widehat{\beta}_t - \beta^*) \right|.
$$

By the Ranga-Rao uniform strong law of large numbers (USLLN) for stationary ergodic processes (see e.g. White (1994, pp. 351)), and, given the uniform negative definiteness of $B(\beta)$,

(a) $\widehat{B}(\overline{\beta}_t)^{-1} - B(\beta^*)^{-1} \overset{a.s.}{\rightarrow} 0$.

In addition, given A1, and by the same argument used in the proof of Lemma 3.1(i) and (ii) in Altissimo and Corradi (2001), it follows that,

(b) $\sup_{t \geq R} t^\theta H_t \overset{a.s.}{\rightarrow} 0$, for $\theta < 1/2$, and so

16

(c) $\sup_{t \geq R} t^\theta(\widehat{\beta}_t - \beta^*) = o_{a.s.}(1)$, for $\theta < 1/2$.

Thus, given A3, $\sup_{t \geq R} P^{1/4}(\widehat{\beta}_t - \beta^*) = o_{a.s.}(1)$. Finally, given the size, smoothness and domination conditions in A1, and by the uniform strong law of large numbers for stationary strong mixing processes,

(d) $\sup_{\gamma \in \Gamma} \left(\frac{1}{P} \sum_{t=R}^{T-1} \nabla_\beta^2 f'_{t+1}(\overline{\beta}_t) g(z^t, \gamma) - E\left(\nabla_\beta^2 f'_{t+1}(\beta^*) g(z^t, \gamma) \right) \right) = o_{a.s.}(1)$.

Thus, the last term on the RHS of equation (8) is $o_p(1)$, uniformly in γ. The second term on the RHS of equation (8) can be treated by first writing it as,

$$\frac{1}{P^{1/2}} \sum_{t=R}^{T-1} F(\beta^*, \gamma)' B(\beta^*)^{-1} H_t + \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} (\nabla_\beta f'_{t+1}(\beta^*) g(z^t, \gamma) - F(\beta^*, \gamma))' B(\beta^*)^{-1} H_t$$

$$+ \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} F(\beta^*, \gamma)' (B_t(\overline{\beta}_t)^{-1} - B(\beta^*)^{-1}) H_t$$

$$+ \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} (\nabla_\beta f'_{t+1}(\beta^*) g(z^t, \gamma) - F(\beta^*, \gamma))' (B_t(\overline{\beta}_t)^{-1} - B(\beta^*)^{-1}) H_t. \qquad (9)$$

We begin by showing that the second, third and fourth terms on the RHS of equation (9) are $o_p(1)$, uniformly in γ. Given A1,

(e) $\sup_{\gamma \in \Gamma} \frac{1}{P} \sum_{t=R}^{T-1} (\nabla_\beta f'_t(\beta^*) g(z^{t-1}, \gamma) - F(\beta^*, \gamma)) \overset{a.s.}{\to} 0$.

Now, given (a),(b),(c), and (e) the second, third and fourth terms on the RHS of equation (9) are $o_p(1)$, pointwise in γ, by the same argument as that used in Lemma A4 of West (1996). In order to show that they are also $o_p(1)$ uniformly in γ, it remains to establish stochastic equicontinuity on Γ. We begin this part of the proof by considering the second term on the RHS of equation (9). Given the absolute regularity assumption and the domination condition in A1, stochastic equicontinuity can be shown, along the lines of Hansen (1996, proof of Theorem 1). In particular,

$$\sup_{\gamma \in \Gamma} \left\| (\nabla_\beta f'_{t+1}(\beta^*) g(z^t, \gamma) - F(\beta^*, \gamma))' B(\beta^*)^{-1} H_t \right\|_{2+\psi}$$

$$\leq \sup_{\gamma \in \Gamma} \left\| (b_{11} + b_{12})(\nabla_{\beta_1} f'_t(\beta^*) g(z^{t-1}, \gamma) - F_1(\beta^*, \gamma))(H_{1t} + H_{2t}) \right\|_{2+\psi}$$

$$+ \sup_{\gamma \in \Gamma} \left\| (b_{12} + b_{22})(\nabla_{\beta_2} f'_{t+1}(\beta^*) g(z^t, \gamma) - F_2(\beta^*, \gamma))(H_{1t} + H_{2t}) \right\|_{2+\psi}, \qquad (10)$$

where $F_i(\beta^*, \gamma)$ and H_{it} are the $i-th$ components of $F(\beta^*, \gamma)$ and H_t, respectively, and b_{ij} is the $ij-th$ element of $B(\beta^*)^{-1}$, for $i, j = 1, 2$. The second term on the RHS of equation (10) is majorized by,

$$\sup_{\gamma \in \Gamma} \left\| (b_{22} + b_{12})(\nabla_{\beta_2} f'_{t+1}(\beta^*) g(z^t, \gamma) - F_2(\beta^*, \gamma)) \right\|_{2(2+\psi)} \|H_{1t} + H_{2t}\|_{2(2+\psi)} < \infty,$$

17

because of A1(ii), and A1(iv), $\|H_{1t} + H_{2t}\|_{2(2+\psi)} < \infty$. Now, for any $\gamma_1, \gamma_2 \in \Gamma$, for $v > 1$,

$$\left\|(b_{22} + b_{12})(\nabla_{\beta_2} f'_{t+1}(\beta^*)((g(z^t, \gamma_1) - F_2(\beta^*, \gamma_1)) - (g(z^t, \gamma_2) - F_2(\beta^*, \gamma_2))))\right\|_{2v}$$

$$\leq \left\|(b_{22} + b_{12})(\nabla_{\beta_2} f'_{t+1}(\beta^*)(g(z^t, \gamma_1) - g(z^t, \gamma_2)))\right\|_{2v}$$

$$+ \left\|(b_{22} + b_{12})(\nabla_{\beta_2} f'_{t+1}(\beta^*)(F_2(\beta^*, \gamma_1)' - F_2(\beta^*, \gamma_2)))\right\|_{2v}$$

$$\leq |b_{22} + b_{12}|^{1/2v} \left\|\nabla_{\beta_2} f'_{t+1}(\beta^*)\nabla_\gamma g(z^t, \overline{\gamma})(\gamma_1 - \gamma_2)\right\|_{2v}$$

$$+ |b_{22} + b_{12}|^{1/2v} \left\|\nabla_{\beta_2} f'_{t+1}(\beta^*)\nabla_\gamma F_2(\beta^*, \overline{\gamma})(\gamma_1 - \gamma_2)\right\|_{2v}$$

$$\leq |b_{22} + b_{12}|^{1/2v} \sum_{i=1}^{d} \left\|(\nabla_{\beta_2} f'_{t+1}(\beta^*)\nabla_\gamma g(z^t, \overline{\gamma}))_i\right\|_{2v} |(\gamma_1 - \gamma_2)_i|$$

$$+ |b_{22} + b_{12}|^{1/2v} \sum_{i=1}^{d} \left\|(\nabla_{\beta_2} f'_{t+1}(\beta^*)\nabla_\gamma F_2(\beta^*, \overline{\gamma}))_i\right\|_{2v} |(\gamma_1 - \gamma_2)_i|$$

$$\leq \Delta \sum_{i=1}^{d} |(\gamma_1 - \gamma_2)_i| \leq \widetilde{\Delta} \|\gamma_1 - \gamma_2\|_2 \ ,$$

where $\overline{\gamma} \in (\gamma_1, \gamma_2)$, $(u)_i$ denotes the $i-$th component of u, and $\Delta, \widetilde{\Delta} < \infty$. The above statement follows given the domination condition in A1 and given the fact that $\nabla_\gamma F(\beta^*, \overline{\gamma}) = E\left(\nabla_\gamma \nabla_\beta f'_t(\beta^*)g(z^t, \overline{\gamma})\right)$ is uniformly bounded because of A1(ii),(iv),(v). The inequality above ensures the integrability of the bracketing number, and thus the second term in equation (9) is stochastic equicontinuous in Γ.

The last two terms in equation (9) can also be shown to be stochastic equicontinuouos, by the same arguments as those used above. Thus, the second term in equation (8) is equal to the first term in equation (9), up to an $o_p(1)$ term, uniformly in γ. This ensures that equation (8) can be written as,

$$m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} f'_{t+1}(\beta^*)g(z^t, \gamma) + \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} F(\beta^*, \gamma)'B(\beta^*)^{-1}H_t + o_p(1), \qquad (11)$$

where the $o_p(1)$ term holds uniformly in Γ. Further, it follows by the same arguments used in West (1996, Lemmas A5,A6, and A1), that $\forall \gamma \in \Gamma$,

$$m_P(\gamma) \xrightarrow{d} N(0, K(\gamma, \gamma)),$$

pointwise in γ, where

$$K(\gamma, \gamma) = S_{ff}(\gamma, \gamma) + 2\Pi(F(\beta^*, \gamma)'B(\beta^*)^{-1}S_{hh}B(\beta^*)^{-1}F(\beta^*, \gamma)) + 2\Pi(F(\beta^*, \gamma)'B(\beta^*)^{-1}S_{fh}(\gamma)),$$

with $S_{ff}(\gamma, \gamma)$, S_{hh}, and $S_{fh}(\gamma)$ defined as in the statement of the theorem, and with $\Pi = 1 - \pi^{-1} \ln(1 + \pi)$, for $\pi > 0$, and $\Pi = 0$ for $\pi = 0$.

Finally, the convergence of the finite dimensional distribution of $(m_P(\gamma_1), \ldots, m_P(\gamma_k))$, follows as a direct consequence of the Cramer-Wold device. In addition, given A1(v), the smoothness of g, and the domination condition in A1, the stochastic equicontinuity over Γ of the two terms on the RHS of equation (11) follows by the same arguments as those used above. Thus, as a straightfoward application of the continuous mapping theorem,

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

where Z is a Gaussian process with covariance kernel given by,

$$
\begin{aligned}
K(\gamma_1, \gamma_2) &= S_{ff}(\gamma_1, \gamma_2) + 2\Pi(F(\beta^*, \gamma_1)'B(\beta^*)^{-1} S_{hh} B(\beta^*)^{-1} F(\beta^*, \gamma_2)) \\
&\quad + \Pi(F(\beta^*, \gamma_1)'B(\beta^*)^{-1} S_{fh}(\gamma_2)) + \Pi(F(\beta^*, \gamma_2)'B(\beta^*)^{-1} S_{fh}(\gamma_1)).
\end{aligned}
\tag{12}
$$

(ii) Follows immediately, by noting that given the same arguments as those used in the first part of this proof,

$$
m_P(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} (f'_{t+1}(\beta^*) g(z^t, \gamma) - E\left(f'_{t+1}(\beta^*) g(z^t, \gamma) \right))
$$

$$
+ P^{1/2} E\left(f'_{t+1}(\beta^*) g(z^t, \gamma) \right) + \frac{1}{P^{1/2}} \sum_{t=R}^{T-1} F(\beta^*, \gamma)'B(\beta^*)^{-1} H_t + o_p(1),
$$

with the $o_p(1)$ terms holding uniformly in γ. The first and third term term above are bounded in probability as they converge in distribution, while the second term diverges at rate $P^{1/2}$, and so at rate P when squared.

Before moving to the proof of Theorem 2 we need the following Lemma. Lemma A will be used in the proof of Theorem below in order to show that the simulated statistic, conditional on the sample, converges to a Gaussian process with the same covariance kernel as the actual statistic.

Lemma A: Let A1, A2, A3 and A4 hold. If $l \to \infty$ and $l\sqrt{\log P}/P^{1/4} \to 0$, as $P \to \infty$, then the following statements hold, uniformly in γ_1 and γ_2:

(i) $\dfrac{1}{P} \sum_{t=R+l}^{T-l} f'_{t+1}(\widehat{\beta}_t) g(z^t, \gamma_1) f'_{t+1}(\widehat{\beta}_t) g(z^t, \gamma_2) + \dfrac{2}{P} \sum_{t=R+l}^{T-l} \sum_{j=t+1}^{t+l-1} f'_{t+1}(\widehat{\beta}_t) g(z^t, \gamma_1) f'_{j+1}(\widehat{\beta}_j) g(z^j, \gamma_2)$

$\xrightarrow{a.s.} S_{ff}(\gamma_1, \gamma_2),$

(ii) $\dfrac{1}{P}\displaystyle\sum_{t=R+l}^{T-l} F(\widehat{\beta}_T,\gamma_1)'B(\widehat{\beta}_T)^{-1}\nabla_\beta f_t(\widehat{\beta}_T)\nabla_\beta f_t(\widehat{\beta}_T)'B(\widehat{\beta}_T)^{-1}F(\widehat{\beta}_T,\gamma_2)$

$+\dfrac{2}{P}\displaystyle\sum_{t=R+l}^{T-l}\sum_{j=t+1}^{t+l-1} F(\widehat{\beta}_T,\gamma_1)'B(\widehat{\beta}_T)^{-1}\nabla_\beta f_t(\widehat{\beta}_T)\nabla_\beta f_j(\widehat{\beta}_T)'B(\widehat{\beta}_T)^{-1}F(\widehat{\beta}_T,\gamma_2)$

$\stackrel{a.s.}{\rightarrow} F(\beta^*,\gamma_1)'B(\beta^*)^{-1}S_{hh}B(\beta^*)^{-1}F(\beta^*,\gamma_2),$

and

(iii) $\dfrac{1}{P}\displaystyle\sum_{t=R+l}^{T-l} f'_{t+1}(\widehat{\beta}_t)g(z^t,\gamma_1)\nabla_\beta f_t(\widehat{\beta}_T)'B(\widehat{\beta}_T)^{-1}F(\widehat{\beta}_T,\gamma_2)$

$+\dfrac{2}{P}\displaystyle\sum_{t=R+l}^{T-l}\sum_{j=t+1}^{t+l-1} f'_{t+1}(\widehat{\beta}_t)g(z^t,\gamma_1)\nabla_\beta f_j(\widehat{\beta}_T)'B(\widehat{\beta}_T)^{-1}F(\widehat{\beta}_T,\gamma_2)$

$\stackrel{a.s.}{\rightarrow} S_{fh}(\gamma_1)B(\beta^*)^{-1}F(\beta^*,\gamma_2).$

Proof of Lemma A: (i) First note that

$$\dfrac{2}{P}\displaystyle\sum_{t=R+l}^{T-l}\sum_{j=t+1}^{t+l-1} f'_{t+1}(\widehat{\beta}_t)g(z^t,\gamma)f'_{j+1}(\widehat{\beta}_j)g(z^j,\gamma^+)$$

$$-\dfrac{2}{P}\displaystyle\sum_{j=1}^{l}\sum_{t=R+j+l}^{T-1} f'_{t+1}(\widehat{\beta}_t)g(z^t,\gamma)f'_{t-j+1}(\widehat{\beta}_{t-j})g(z^{t-j},\gamma^+) = o_{a.s.}(1),$$

as the two differ by at most $2l$ terms and $l/P = o(1)$. We begin by showing that

$$\dfrac{2}{P}\displaystyle\sum_{j=1}^{l}\sum_{t=R+j+l}^{T-1} f'_{t+1}(\widehat{\beta}_t)g(z^t,\gamma)f'_{t-j+1}(\widehat{\beta}_{t-j})g(z^{t-j},\gamma^+)$$

$$-\dfrac{2}{P}\displaystyle\sum_{j=1}^{l}\sum_{t=R+j+l}^{T-1} f'_{t+1}(\beta^*)g(z^t,\gamma)f'_{t-j+1}(\beta^*)g(z^{t-j},\gamma^+) = o_{a.s.}(1), \tag{13}$$

where the $o_{a.s.}(1)$ term holds uniformly in γ,γ^+. Via a mean value expansion around β^*, the LHS of (13) can be written as,

$$\dfrac{2}{P}\displaystyle\sum_{j=1}^{l}\sum_{t=R+j+l}^{T-1} f'_{t+1}(\beta^*)g(z^t,\gamma)\nabla_\beta f'_{t-j+1}(\overline{\beta}_{t-j})'g(z^{t-j},\gamma^+)(\widehat{\beta}_{t-j}-\beta^*)$$

$$+\dfrac{2}{P}\displaystyle\sum_{j=1}^{l}\sum_{t=R+j+l}^{T-1} (\widehat{\beta}_t-\beta^*)'\nabla_\beta f'_{t+1}(\overline{\beta}_t)g(z^{t-1},\gamma)\nabla_\beta f'_{t-j+1}(\overline{\beta}_{t-j})'g(z^{t-j},\gamma^+)(\widehat{\beta}_{t-j}-\beta^*)$$

$$+\dfrac{2}{P}\displaystyle\sum_{j=1}^{l}\sum_{t=R+j+l}^{T-1} (\widehat{\beta}_t-\beta^*)'\nabla_\beta f'_{t+1}(\overline{\beta}_t)g(z^{t-1},\gamma)f'_{t-j+1}(\beta^*)g(z^{t-j},\gamma^+). \tag{14}$$

20

We now show that the first term in (14) is $o_{a.s.}(1)$, uniformly in γ, γ^+. The second and third terms can be treated in an analogous way, given A4(ii)-(iii). Now,

$$\sup_{t\geq R}\sqrt{\frac{t}{\log\log t}}|\widehat{\beta}_t - \beta^*| \leq \sup_{t\geq R}\left(\frac{1}{t}\sum_{j=2}^{t}\nabla_\beta^2 f_j(\overline{\beta}_t)\right)^{-1}\sup_{t\geq R}\left|\frac{1}{\sqrt{t\log\log t}}\sum_{j=2}^{t}\nabla_\beta f_j(\beta^*)\right|$$

$$= O_{a.s.}(1)\sup_{t\geq R}\left|\frac{1}{\sqrt{t\log\log t}}\sum_{j=2}^{t}\nabla_\beta f_j(\beta^*)\right| = O_{a.s.}(1),$$

This follows from the strong invariance principle for strong mixing processes (Eberlain (1986,)), given A1 and A2, as in this case, $\sup_{t\geq R}\left|\frac{1}{\sqrt{t\log\log t}}\sum_{j=2}^{t}\nabla_\beta f_j(\beta^*)\right| = O_{a.s.}(1)$. By the same argument as that used in the proof of Lemma 3.1(iii) in Altissimo and Corradi (2001), and given A4(i), for all $k = 1, 2$,

$$\Pr\left(\left|\frac{2\sqrt{R\log\log R}}{P}\sup_{\beta\times\gamma\times\gamma^+\in B\times\Gamma\times\Gamma}\sum_{j=1}^{l}\sum_{t=R+j+l}^{t+l-1}f'_t(\beta^*)g(z^{t-1},\gamma)\nabla_\beta^k f'_{t-j}(\overline{\beta}_{t-j})g(z^{t-j-1},\gamma^+)\right| > \varepsilon\right)$$

$$\leq \frac{l^4(\log\log R)^2}{\varepsilon^4 R^2}E\left(\sup_{\beta\times\gamma\times\gamma^+\in B\times\Gamma\times\Gamma}f'_t(\beta^*)g(z^{t-1},\gamma)\nabla_\beta^k f'_{t-j}(\overline{\beta}_{t-j})g(z^{t-j-1},\gamma^+)\right)^4$$

$$\leq \frac{l^4(\log\log R)^2}{\varepsilon^4 R^2}\Delta.$$

Given A3, and given that $l\sqrt{\log P}/P^{1/4} \to 0$, as $P \to \infty$, $\frac{l^4(\log\log R)^2}{R^2} = o(P^{-1})$. Thus, by the first Borel Cantelli Lemma, the first term in (14) converges to zero almost surely, uniformly in γ, γ^+. As mentioned above, the second and third terms in (14) can be treated in an analogous way. For fixed γ and γ^+, given A1,

$$\frac{1}{P}\sum_{t=R+l}^{T-2l}f'_{t+1}(\beta^*)g(z^t,\gamma_1)f'_{t+1}(\beta^*)g(z^t,\gamma_2) + \frac{2}{P}\sum_{j=1}^{l}\sum_{t=R+j+l}^{t+l-1}f'_t(\beta^*)g(z^{t-1},\gamma)f'_{t-j}(\beta^*)g(z^{t-j-1},\gamma^+)$$

$$\overset{a.s.}{\to} S_{ff}(\gamma,\gamma^+),$$

by Theorem 1 in de Jong (2000). In order to obtain almost sure convergence, uniformly in γ, we need to show strong stochastic equicontinuity in Γ. Given the mixing conditions in A1(i), and the domination conditions in A1(ii) and in A1(iv)-(v), the result follows from Lemma 3.1(iii) in Altissimo and Corradi (2001)[18], which extends the pointwise result of de Jong (2000) to a uniform result.

In addition, (ii) and (iii) follow by the same arguments as those used to show (i).

---

[18]Note that Assumption A5 in that paper is trivially satisfied because of the strict stationarity of (x_t, z^t).

Proof of Theorem 2: (i) We begin by considering $m_P^{*(1)}(\gamma)$. Thereafter, $m_P^{*(2)}(\gamma)$ is treated analogously. Hereafter, $E^*(\cdot)$, $Var^*(\cdot)$, and $Cov^*(\cdot)$ denote mean, variance, and covariance, conditional on the sample. Finally $a.s. - \omega$ means conditional on the sample and for all samples but a set of measure zero. Note first that, conditional on the sample, $m_P^{*(1)}(\gamma)$ is a zero mean normal random variable with variance equal to:

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

and

$$\begin{aligned}
Cov^*(m_P^{*(1)}(\gamma_1), m_P^{*(1)}(\gamma_2)) &= \frac{1}{Pl}\sum_{t=R}^{T-l}\left(\sum_{i=t}^{t+l-1}\left(f'(\widehat{u}_{0,i+1})g(z^i,\gamma_1) + \widehat{\Pi}\widehat{F}(\widehat{\beta}_T,\gamma_1)'\widehat{B}(\widehat{\beta}_T)^{-1}\nabla_\beta f_i(\widehat{\beta}_T)\right)\right) \\
&\quad \times\left(\sum_{i=t}^{t+l-1}\left(f'(\widehat{u}_{0,i+1})g(z^i,\gamma_2) + \widehat{\Pi}\widehat{F}(\widehat{\beta}_T,\gamma_2)'\widehat{B}(\widehat{\beta}_T)^{-1}\nabla_\beta f_i(\widehat{\beta}_T)\right)\right)
\end{aligned}$$

We need to show that, $a.s. - \omega$,

$$\begin{aligned}
&Cov^*(m_P^{*(1)}(\gamma_1), m_P^{*(1)}(\gamma_2)) \\
&\rightarrow \quad S_{ff}(\gamma_1,\gamma_2) + \Pi^2 F(\beta^*,\gamma_1)'B(\beta^*)^{-1}S_{hh}B(\beta^*)^{-1}F(\beta^*,\gamma_2) \\
&\quad +\Pi F(\beta^*,\gamma_2)'B(\beta^*)^{-1}S_{fh}(\gamma_1) + \Pi F(\beta^*,\gamma_1)'B(\beta^*)^{-1}S_{fh}(\gamma_2) \quad\quad (15)
\end{aligned}$$

uniformly in γ_1, and γ_2. Given (a)-(e) in the proof of Theorem (1) and given Lemma A above, it follows that, $a.s. - \omega$,

$$\begin{aligned}
&\frac{1}{Pl}\sum_{t=R}^{T-l}\left(\sum_{i=t}^{t+l-1}\left(f'(\widehat{u}_{0,i+1})g(z^i,\gamma_1) + \widehat{\Pi}\widehat{F}(\widehat{\beta}_T,\gamma_1)'\widehat{B}(\widehat{\beta}_T)^{-1}\nabla_\beta f_i(\widehat{\beta}_T)\right)\right) \\
&\quad \times\left(\sum_{i=t}^{t+l-1}\left(f'(\widehat{u}_{0,i+1})g(z^i,\gamma_2) + \widehat{\Pi}\widehat{F}(\widehat{\beta}_T,\gamma_2)'\widehat{B}(\widehat{\beta}_T)^{-1}\nabla_\beta f_i(\widehat{\beta}_T)\right)\right) \\
&= \frac{1}{Pl}\sum_{t=R}^{T-l}\left(\sum_{i=t}^{t+l-1}\left(f'(u_{0,i+1})g(z^i,\gamma_1) + \Pi F(\beta^*,\gamma_1)'B(\beta^*)^{-1}\nabla_\beta f_i(\beta^*)\right)\right) \\
&\quad \times\left(\sum_{i=t}^{t+l-1}\left(f'(u_{0,i+1})g(z^i,\gamma_2) + \Pi F(\beta^*,\gamma_2)'B(\beta^*)^{-1}\nabla_\beta f_i(\beta^*)\right)\right)(1 + o(1)),
\end{aligned}$$

where the $o(1)$ term is uniform in γ_1 and γ_2. Now, $a.s. - \omega$,

$$
\frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} \left(f'(u_{0,i+1}) g(z^i, \gamma_1) + \Pi F(\beta^*, \gamma_1)' B(\beta^*)^{-1} \nabla_\beta f_i(\beta^*) \right) \right)
$$
$$
\times \left(\sum_{i=t}^{t+l-1} \left(f'(u_{0,i+1}) g(z^i, \gamma_2) + \Pi F(\beta^*, \gamma_2)' B(\beta^*)^{-1} \nabla_\beta f_i(\beta^*) \right) \right)
$$
$$
= \frac{1}{P} \sum_{t=R+l}^{T-l} \left(f'(u_{0,t+1}) g(z^t, \gamma_1) + \Pi F(\beta^*, \gamma_1)' B(\beta^*)^{-1} \nabla_\beta f_t(\beta^*) \right)
$$
$$
\times \left(f'_{t+1}(\beta^*) g(z^t, \gamma_2) + \Pi F(\beta^*, \gamma_2)' B(\beta^*)^{-1} \nabla_\beta f_t(\beta^*) \right)
$$

$$
+ \frac{1}{P} \sum_{t=R+l}^{T-l} \sum_{i=t+1}^{t+l-1} \left(f'(u_{0,t+1}) g(z^t, \gamma_1) + \Pi F(\beta^*, \gamma_1)' B(\beta^*)^{-1} \nabla_\beta f_t(\beta^*) \right)
$$
$$
\times \left(f'(u_{0,i+1}) g(z^i, \gamma_2) + \Pi F(\beta^*, \gamma_2)' B(\beta^*)^{-1} \nabla_\beta f_i(\beta^*) \right) + o(1).
$$

From Lemma A(i)-(iii), it follows that $Cov^*(m_P^{*(1)}(\gamma_1), m_P^{*(1)}(\gamma_2))$ converges uniformly in γ_1 and γ_2, $a.s - \omega$ to the expression in (15). Similarly, we can show that

$$
Cov^*(m_P^{*(2)}(\gamma_1), m_P^{*(2)}(\gamma_2)) = (2\widehat{\Pi} - \widehat{\Pi}^2) \frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} \widehat{F}(\widehat{\beta}_T, \gamma_1)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T) \right)
$$
$$
\times \left(\sum_{i=t}^{t+l-1} \widehat{F}(\widehat{\beta}_T, \gamma_2)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T) \right)
$$
$$
\to (\Pi - \Pi^2)^{1/2} F(\beta^*, \gamma_1)' B(\beta^*)^{-1} S_{hh} B(\beta^*)^{-1} F(\beta^*, \gamma_2), \ a.s. - \omega,
$$

uniformly in γ_1, γ_2. As $E(\epsilon_t \eta_s) = 0$, $\forall t, s$ $Cov^*(m_P^{*(1)}(\gamma_1), m_P^{*(2)}(\gamma_2)) = 0$, $\forall \gamma_1, \gamma_2$, and so $a.s - \omega$,

$$
Cov^*(m_P^*(\gamma_1), m_P^*(\gamma_2)) \to S_{ff}(\gamma_1, \gamma_2) + 2\Pi(F(\beta^*, \gamma_1)' B(\beta^*)^{-1} S_{hh} B(\beta^*)^{-1} F(\beta^*, \gamma_2))
$$

$$
+ \Pi(F(\beta^*, \gamma_1)' B(\beta^*)^{-1} S_{fh}(\gamma_2)) + \Pi(F(\beta^*, \gamma_2)' B(\beta^*)^{-1} S_{fh}(\gamma_1)) = K(\gamma_1, \gamma_2)
$$

where $K(\gamma_1, \gamma_2)$ is defined as in the statement of Theorem 1. Thus, for any given γ_1 and γ_2,

$$
(m_P^*(\gamma_1), m_P^*(\gamma_2))' \xrightarrow{d} N \left(0, \left(\begin{array}{cc} K(\gamma_1, \gamma_1) & K(\gamma_1, \gamma_2) \\ K(\gamma_1, \gamma_2) & K(\gamma_2, \gamma_2) \end{array} \right) \right), \ a.s. - \omega
$$

Hereafter let,

$$
\widehat{S}_{1,i+1}(\gamma) = f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) + \widehat{\Pi} \widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T)
$$

and

$$
\widehat{S}_{2,i}(\gamma) = (2\widehat{\Pi} - \widehat{\Pi}^2)^{1/2} \widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T),
$$

23

so that

$$m_p^*(\gamma) = \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} \widehat{S}_{1,i+1}(\gamma) + \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \eta_t \sum_{i=t}^{t+l-1} \widehat{S}_{2,i}(\gamma).$$

In order to show that under the null, $m_P^*(\gamma)$ converges to a Gaussian process with the same covariance structure as $m_P(\gamma)$, we need to show that $a.s. - \omega$,

$$\text{(ia)} \ \limsup_{P \to \infty} E^* \left(\sup_{\gamma \in \Gamma} \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \epsilon_t \sum_{i=t}^{t+l-1} \widehat{S}_{1,i}(\gamma) \right)^2 < \infty,$$

and

$$\text{(ib)} \ \limsup_{P \to \infty} E^* \left(\sup_{\gamma \in \Gamma} \frac{1}{P^{1/2}} \sum_{t=R}^{T-l} \eta_t \sum_{i=t}^{t+l-1} \widehat{S}_{2,i}(\gamma) \right)^2 < \infty.$$

Also, we require that, $a.s. - \omega$,

$$\text{(ii)} \ \lim_{P \to \infty} E^* \big(\big(\big(\tfrac{1}{P^{1/2}} \sum_{t=R}^{T-l} \big(\epsilon_t \sum_{i=t}^{t+l-1} \widehat{S}_{1,i+1}(\gamma_1) + \eta_t \sum_{i=t}^{t+l-1} \widehat{S}_{2,i+1}(\gamma_1) \big) \big)$$
$$- \big(\tfrac{1}{P^{1/2}} \sum_{t=R}^{T-l} \big(\epsilon_t \sum_{i=t}^{t+l-1} \widehat{S}_{1,i+1}(\gamma_2) + \eta_t \sum_{i=t}^{t+l-1} \widehat{S}_{2,i+1}(\gamma_2) \big) \big) \big)^2 \big)^{1/2}$$
$$= \Psi(\gamma_1, \gamma_2) \leq \Delta \|\gamma_1 - \gamma_2\|_2.$$

Now, it suffices to show (ia) and (ii), as (ib) follows directly from (ia). Note first that (ia) can be written as:

$$\limsup_{P \to \infty} \sup_{\gamma \in \Gamma} \frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} f'(\widehat{u}_{0,i+1}) g(z^i, \gamma) + \widehat{\Pi} \widehat{F}(\widehat{\beta}_T, \gamma)' \widehat{B}(\widehat{\beta}_T)^{-1} \nabla_\beta f_i(\widehat{\beta}_T) \right)^2,$$

which is finite, $a.s. - \omega$, by Lemma A. Now, note that (ii) can be written as:

$$\left(\frac{1}{Pl} \sum_{t=R}^{T-l} \left(\left(\sum_{i=t}^{t+l-1} (\widehat{S}_{1i}(\gamma_1) + \widehat{S}_{2i}(\gamma_2)) \right) - \left(\sum_{i=t}^{t+l-1} (\widehat{S}_{1i}(\gamma_1) + \widehat{S}_{2i}(\gamma_2)) \right) \right)^2 \right)^{1/2}$$

$$= \ (Var^*(m_P^*(\gamma_1)) + Var^*(m_P^*(\gamma_2)) - 2Cov^*(m_P^*(\gamma_1), m_P^*(\gamma_2)))^{1/2}$$

$$\to \ (K(\gamma_1, \gamma_1) + K(\gamma_2, \gamma_2) - K(\gamma_1, \gamma_2) - K(\gamma_1, \gamma_2))^{1/2},$$

Now the square on the right hand side above, it majorized by

$$|K(\gamma_1, \gamma_1) - K(\gamma_1, \gamma_2)| + |K(\gamma_2, \gamma_2) - K(\gamma_1, \gamma_2)|. \tag{16}$$

From equation (12), we see that

$$|K(\gamma_1, \gamma_1) - K(\gamma_1, \gamma_2)| \leq \tag{17}$$

$$|S_{ff}(\gamma_1, \gamma_1) - S_{ff}(\gamma_1, \gamma_2)| + \left| 2\Pi (F(\beta^*, \gamma_1)' B(\beta^*)^{-1} S_{hh} B(\beta^*)^{-1} (F(\beta^*, \gamma_1) - F(\beta^*, \gamma_2)) \right|$$

$$+ \left| \Pi(F(\beta^*,\gamma_1)'B(\beta^*)^{-1}(S_{fh}(\gamma_1) - S_{fh}(\gamma_2))) \right| + \left| \Pi((F(\beta^*,\gamma_1)' - F(\beta^*,\gamma_2)')B(\beta^*)^{-1}S_{fh}(\gamma_1))) \right|.$$

The first term on the RHS of (17) is equal to

$$\left| \sum_{j=-\infty}^{\infty} E(f_q'(\beta^*)g(z^{q-1},\gamma_1)f_{q+j}'(\beta^*)(g(z^{q+j-1},\gamma_1) - g(z^{q+j-1},\gamma_2))) \right|$$

$$= \left| \sum_{j=-\infty}^{\infty} E(f_q'(\beta^*)g(z^{q-1},\gamma_1)f_{q+j}'(\beta^*)(\nabla_\gamma g(z^{q+j-1},\overline{\gamma})'(\gamma_1 - \gamma_2))) \right|$$

$$\leq \sum_{i=1}^{d} \left| \sum_{j=-\infty}^{\infty} E((f_q'(\beta^*)g(z^{q-1},\gamma_1)f_{q+j}'(\beta^*)(\nabla_\gamma g(z^{q+j-1},\overline{\gamma}))))_i(\gamma_1 - \gamma_2)_i \right|$$

$$\leq \Delta \left| \sum_{j=-\infty}^{\infty} E((f_q'(\beta^*)g(z^{q-1},\gamma_1)f_{q+j}'(\beta^*)(\nabla_\gamma g(z^{q+j-1},\overline{\gamma}))))_i \right|_{2v} \sum_{i=1}^{d} |(\gamma_1 - \gamma_2)_i|$$

$$\leq \widetilde{\Delta} \|\gamma_1 - \gamma_2\|_2, \text{ for } \overline{\gamma} \in (\gamma_1, \gamma_2) \text{ and for } v > 1,$$

given the size condition in A1(i) and given A1(iv). As for the second term on the RHS of (17), it is equal to

$$\left| 2\Pi(F(\beta^*,\gamma_1)'B(\beta^*)^{-1}S_{hh}B(\beta^*)^{-1}\nabla_\gamma F(\beta^*,\overline{\gamma})'(\gamma_1 - \gamma_2) \right|$$

$$\leq \sum_{i=1}^{d} \left| 2\Pi(F(\beta^*,\gamma_1)'B(\beta^*)^{-1}S_{hh}B(\beta^*)^{-1}(\nabla_\gamma F(\beta^*,\overline{\gamma}))_i(\gamma_1 - \gamma_2)_i \right|$$

$$\leq \widetilde{\Delta} \|\gamma_1 - \gamma_2\|_2, \text{ and } \overline{\gamma} \in (\gamma_1, \gamma_2)$$

as because of A1(v),

$$\nabla_\gamma F(\beta^*,\overline{\gamma}) = \nabla_\gamma E\left(\nabla_\beta f_t'(\beta^*)g(z^t,\overline{\gamma})\right) = E\left(\nabla_\gamma \nabla_\beta f_t'(\beta^*)g(z^t,\overline{\gamma})\right)$$

is finite, uniformly in Γ. The third as fourth terms on the RHS of (17), as well as the second term on the RHS of (16) can be treated in an analogous way. Thus (ii) follows. Now, (ii) ensures that the covering number condition in Pollard (1990, Theorem 10.6) is satisfied (see also Hansen (1996, proof of Theorem 2)). This, together with (ia)-(ib) and the convergence of the finite dimensional distribution ensures that $m_P^*(\gamma)$ is stochastic equicontinuouos over Γ, $a.s. -\omega$. The desired outcome then follows from the continuous mapping theorem.

(ii) Under the alternative $E(f_{t+1}'(\beta^*)g(z^t,\gamma)) = \mu_\gamma \neq 0$, for all $\gamma \in \Gamma$. Now, $a.s. - \omega$,

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

$$= \frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} \left((f'(\widehat{\beta}_i)g(z^{q+i-1},\gamma) - \mu_\gamma) + \Pi\widehat{F}(\widehat{\beta}_i,\gamma)'B_i(\widehat{\beta}_i)^{-1}\nabla_\beta f(\widehat{\beta}_i) + \mu_\gamma \right) \right)^2$$

$$= O(1) + \frac{1}{Pl} \sum_{t=R}^{T-l} \left(\sum_{i=t}^{t+l-1} \mu_\gamma \right)^2 = O(l), \ a.s. - \omega,$$

uniformly in γ. By the same argument used in Part (i), we can show that $l^{-1/2}m_P^*(\gamma)$ is stochastic equicontinuous on Γ, conditionally on the sample and for all samples except of measure zero. Thus $\sup_{\gamma \in \Gamma} l^{-(1+\eta)}m_P^*(\gamma) = o_{P^*}(1)$, $\forall \eta > 0$, and so $l^{-1/2}m_P^*(\gamma) = O_{P^*}(1)$ uniformly in γ. The desired result then follows.

6 References

Altissimo F. and V. Corradi, (2001), Bounds for Inference in the Presence of Nuisance Parameters Unidentified Under the Null, University of Exeter, Discussion Paper 00/13..

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,* forthcoming.

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

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

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

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 D.F. Hendry, (2001), Forecasting with Difference and Trend Stationary Models, *The Econometrics Journal,* forthcoming.

Corradi, V. and N.R. Swanson, (2001), Bootstrapping Conditional Distributions Tests Under Dynamic Misspecification, Mimeo, University of Exeter.

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

de Jong, R.M., (1996), The Bierens Test Under Data Dependence, *Journal of Econometrics,* 72, 1-32.

de Jong, R.M., (2000): A Strong Consistency Proof for Heteroskedasticity and Autocorrelation Consistent Covariance Estimators, *Econometric Theory, 16, 262-269.*

Diebold, F.X. and R.S. Mariano, (1995), Comparing Predictive Accuracy, *Journal of Business and*

Economic Statistics, 13, 253-263.

Doukhan, P., (1995), *Mixing: Properties and Examples,* Lecture Notes in Statistics, Springer and Verlag, Berlin.

Eberlain, E., (1986), On Strong Invariance Principles under Dependence Assumptions, *Annals of Probability,* 14, 260-270.

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

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

Granger, C.W.J. and P. Newbold, (1986), *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.

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

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

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.

Luukkonen, R., P. Saikkonen and T. Teräsvirta, (1988), Testing Linearity Against Smooth Transition Autoregressive Models, *Biometrika*, 75, 491-499.

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

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

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

Pollard, D., (1990), *Empirical Processes: Theory and Applications,* CBMS, Conference Series in Probability and Statistics, v.2, Hayward, CA, Institute of Mathematical Statistics.

Ramsey, J.B., (1969), Tests for Specification Errors in Classical Linear Least Squares Regression Analysis, *Journal of the Royal Statistical Society, Series B,* 31, 350-371.

Rothman, P., D.J.C. van Dijk and P.H. Franses, (2000), A Multivariate STAR Analysis of the Relationship Between Money and Output, *Macroeconomic Dynamics*, forthcoming.

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

Sullivan, R., A. Timmermann and H. White, (1999), Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns, *Journal of Finance,* 54, 1647-1691.

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.

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

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

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

White, H., (1994), *Estimation, Inference and Specification Analysis,* Cambridge University Press, Cambridge.

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

Wolkonski, V.A. and Y.A. Rozanov, (1959), Some Limit Theorems for Random Functions, Part I,

Theory of Probability and its Applications, 6, 186-198.

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

Table 1: Monte Carlo Rejection Frequencies Based on Quadratic Loss, T=600 *

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	0.110	0.125	0.120	0.125	0.140	0.125	0.145	0.160	0.155
Size2	0.110	0.120	0.125	0.125	0.140	0.125	0.140	0.160	0.150
Power1	0.720	0.635	0.800	0.705	0.610	0.795	0.725	0.635	0.780
Power2	0.890	0.815	0.935	0.880	0.810	0.920	0.870	0.780	0.910
Power3	0.915	0.860	0.945	0.910	0.835	0.945	0.890	0.805	0.935
Power4	0.930	0.880	0.975	0.930	0.885	0.980	0.930	0.875	0.965
Power5	0.790	0.705	0.870	0.760	0.710	0.835	0.775	0.710	0.835
Power6	0.805	0.760	0.875	0.825	0.760	0.895	0.835	0.755	0.880
Power7	0.725	0.630	0.785	0.700	0.620	0.785	0.730	0.635	0.785
Power8	0.895	0.825	0.930	0.885	0.810	0.930	0.870	0.785	0.925
Power9	0.920	0.845	0.945	0.900	0.840	0.945	0.895	0.815	0.940
Power10	0.930	0.875	0.970	0.930	0.875	0.975	0.940	0.870	0.970
Power11	0.780	0.715	0.865	0.760	0.690	0.840	0.785	0.700	0.830
Power12	0.805	0.755	0.870	0.825	0.760	0.890	0.810	0.750	0.890
Panel B: $a_2 = 0.4$									
Size1	0.100	0.110	0.120	0.130	0.145	0.140	0.150	0.165	0.155
Size2	0.105	0.125	0.120	0.140	0.145	0.115	0.150	0.155	0.145
Power1	0.695	0.635	0.775	0.710	0.635	0.775	0.715	0.640	0.765
Power2	0.865	0.820	0.935	0.860	0.810	0.920	0.855	0.780	0.910
Power3	0.895	0.830	0.940	0.885	0.820	0.930	0.865	0.810	0.925
Power4	0.910	0.865	0.950	0.920	0.870	0.960	0.910	0.870	0.950
Power5	0.795	0.700	0.855	0.770	0.685	0.835	0.760	0.715	0.820
Power6	0.810	0.765	0.875	0.810	0.770	0.865	0.800	0.740	0.870
Power7	0.715	0.625	0.785	0.710	0.645	0.770	0.715	0.630	0.770
Power8	0.860	0.810	0.945	0.840	0.805	0.925	0.855	0.790	0.905
Power9	0.895	0.840	0.950	0.880	0.820	0.940	0.865	0.810	0.930
Power10	0.910	0.870	0.960	0.915	0.865	0.960	0.900	0.865	0.950
Power11	0.790	0.695	0.865	0.760	0.665	0.830	0.775	0.700	0.820
Power12	0.815	0.765	0.870	0.795	0.770	0.870	0.805	0.750	0.865

* 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 so that all forecasts used in test construction are real-time. See above for further details.

Table 2: Monte Carlo Rejection Frequencies Based on Quadratic Loss, T=1000 [*]

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	0.130	0.115	0.140	0.135	0.135	0.130	0.130	0.145	0.125
Size2	0.135	0.130	0.150	0.150	0.135	0.150	0.140	0.140	0.150
Power1	0.940	0.865	0.965	0.925	0.855	0.955	0.910	0.840	0.955
Power2	0.985	0.970	0.990	0.970	0.960	0.990	0.965	0.935	0.990
Power3	0.985	0.960	1.000	0.975	0.960	0.990	0.970	0.945	0.990
Power4	0.985	0.975	1.000	0.985	0.975	1.000	0.985	0.960	0.995
Power5	0.950	0.905	0.980	0.940	0.890	0.960	0.925	0.890	0.955
Power6	0.965	0.940	0.975	0.965	0.920	0.970	0.935	0.915	0.970
Power7	0.930	0.845	0.965	0.930	0.865	0.955	0.910	0.845	0.955
Power8	0.985	0.970	0.995	0.970	0.950	0.980	0.960	0.940	0.990
Power9	0.985	0.960	1.000	0.975	0.955	0.995	0.965	0.945	0.985
Power10	0.985	0.975	1.000	0.985	0.975	1.000	0.985	0.965	0.995
Power11	0.945	0.895	0.980	0.945	0.885	0.965	0.925	0.880	0.955
Power12	0.965	0.940	0.975	0.960	0.930	0.975	0.945	0.910	0.970
Panel B: $a_2 = 0.4$									
Size1	0.135	0.125	0.145	0.145	0.140	0.140	0.140	0.135	0.135
Size2	0.130	0.135	0.125	0.135	0.160	0.135	0.155	0.170	0.135
Power1	0.915	0.825	0.965	0.930	0.830	0.960	0.895	0.825	0.955
Power2	0.985	0.955	0.990	0.970	0.945	0.990	0.955	0.930	0.990
Power3	0.985	0.955	1.000	0.965	0.950	0.990	0.960	0.935	0.990
Power4	0.985	0.975	1.000	0.980	0.970	0.995	0.985	0.975	0.995
Power5	0.935	0.905	0.975	0.935	0.885	0.955	0.890	0.850	0.955
Power6	0.955	0.905	0.980	0.950	0.910	0.975	0.925	0.890	0.970
Power7	0.910	0.825	0.970	0.915	0.820	0.955	0.885	0.830	0.950
Power8	0.985	0.955	0.990	0.965	0.940	0.990	0.960	0.925	0.990
Power9	0.985	0.955	1.000	0.970	0.960	0.990	0.965	0.930	0.990
Power10	0.985	0.975	1.000	0.980	0.970	0.995	0.985	0.965	0.990
Power11	0.940	0.875	0.975	0.940	0.870	0.955	0.905	0.850	0.950
Power12	0.945	0.910	0.980	0.945	0.905	0.975	0.930	0.890	0.965

[*] See notes to Table 1.

Table 3: Monte Carlo Rejection Frequencies Using M_P^{sup} and Based on Linex Loss*

Model	T=600			T=1000		
	l=30	l=40	l=50	l=30	l=40	l=50
Panel A: $a_2 = 0.2$						
Size1	0.235	0.230	0.225	0.240	0.225	0.210
Size2	0.225	0.230	0.235	0.245	0.250	0.250
Power1	0.300	0.280	0.295	0.540	0.530	0.520
Power2	0.265	0.260	0.250	0.575	0.565	0.560
Power3	0.495	0.505	0.460	0.685	0.680	0.705
Power4	0.360	0.390	0.345	0.475	0.530	0.545
Power5	0.395	0.375	0.375	0.635	0.615	0.605
Power6	0.270	0.315	0.285	0.390	0.390	0.445
Power7	0.315	0.300	0.305	0.525	0.550	0.530
Power8	0.270	0.275	0.220	0.570	0.590	0.550
Power9	0.530	0.500	0.485	0.715	0.700	0.690
Power10	0.345	0.395	0.360	0.480	0.525	0.525
Power11	0.390	0.415	0.390	0.625	0.635	0.615
Power12	0.270	0.305	0.265	0.395	0.420	0.450
Panel B: $a_2 = 0.4$						
Size1	0.230	0.245	0.235	0.260	0.245	0.240
Size2	0.245	0.230	0.240	0.255	0.270	0.250
Power1	0.290	0.270	0.275	0.500	0.540	0.515
Power2	0.215	0.175	0.180	0.480	0.455	0.460
Power3	0.480	0.480	0.460	0.635	0.680	0.665
Power4	0.315	0.350	0.330	0.450	0.470	0.455
Power5	0.360	0.395	0.380	0.560	0.600	0.575
Power6	0.255	0.280	0.270	0.355	0.350	0.355
Power7	0.280	0.270	0.260	0.515	0.535	0.510
Power8	0.190	0.155	0.150	0.455	0.450	0.430
Power9	0.485	0.485	0.475	0.665	0.670	0.665
Power10	0.325	0.335	0.305	0.450	0.470	0.460
Power11	0.355	0.380	0.365	0.570	0.585	0.590
Power12	0.255	0.285	0.280	0.340	0.350	0.370

* See notes to Table 1.