FORECAST-BASED MODEL SELECTION IN THE PRESENCE OF STRUCTURAL BREAKS

Todd E. Clark
Michael W. McCracken

AUGUST 2002

RWP 02-05

Research Division
Federal Reserve Bank of Kansas City

Todd E. Clark is an assistant vice president and economist at the Federal Reserve Bank of Kansas City. Michael W. McCracken is an assistant professor of economics at the University of Missouri. The authors gratefully acknowledge the helpful comments of: Gregory Chow, David Rapach, Norman Swanson, and Jonathan Wright; seminar participants at Emory University, the University of Missouri, and the Federal Reserve Bank of Kansas City; and participants at the 2001 International Symposium on Forecasting, 2001 Midwest Econometrics Group Meetings, Winter 2002 Econometric Society Meetings, Federal Reserve System Committee on Macroeconomics, 2002 WEA Meetings, and 2002 NBER Summer Institute. Jushan Bai kindly provided computer programs for estimating the Bai-Perron break tests. A significant portion of this paper was completed while McCracken was employed at Louisiana State University; financial support is gratefully acknowledged. The views expressed are those of the authors and do not represent those of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

Clark email: todd.e.clark@kc.frb.org
McCracken email: mccrackenm@missouri.edu

Abstract

This paper presents analytical, Monte Carlo, and empirical evidence on the effects of structural breaks on tests for equal forecast accuracy and forecast encompassing. The forecasts are generated from two parametric, linear models that are nested under the null. The alternative hypotheses allow a causal relationship that is subject to breaks during the sample. With this framework, we show that in-sample explanatory power is readily found because the usual F-test will indicate causality if it existed for any portion of the sample. Out-of-sample predictive power can be harder to find because the results of out-of-sample tests are highly dependent on the timing of the predictive ability. Moreover, out-of-sample predictive power is harder to find with some tests than with others: the power of F-type tests of equal forecast accuracy and encompassing often dominates that of the more commonly-used t-type alternatives. Overall, out-of-sample tests are effective at revealing whether one variable has predictive power for another at the end of the sample. Based on these results and additional evidence from two empirical applications, we conclude that structural breaks can explain why researchers often find evidence of in-sample, but not out-of-sample, predictive content.

JEL classification: C53, C12, C52

Keywords: power, structural breaks, forecast evaluation, model selection

1. Introduction

It is now common knowledge that in-sample predictive ability need not imply out-of-sample predictive ability. For example, in the exchange rate literature spanning Meese and Rogoff (1983, 1988) through Kilian and Taylor (2001), study after study has concluded that models that fit well in-sample fail to improve over a simple random walk in out-of-sample forecast comparisons. Some of the differences between in-sample and out-of-sample results on predictive ability may be due to model instability.[1] Stock and Watson (1996, 1999) show that instability pervades a wide range of time series. Moreover, detailed studies of particular relationships find some instabilities. Examples include the Estrella, Rodrigues, and Schich (2000) and Stock and Watson (2001) analyses of the link from financial variables to real activity and the Paye and Timmermann (2002) and Rapach and Wohar (2002) studies of the relationship of stock returns to financial variables.[2] Clements and Hendry (1999) and Hendry (2000) argue that, because of structural shifts, simple autoregressions in differences or second differences forecast better than more structural models.

In light of these findings, this paper examines the behavior of out-of-sample forecast tests in the presence of structural breaks. In particular, using parametric linear regression models we present analytical, Monte Carlo, and empirical evidence on the effects of structural breaks on out-of-sample forecast tests. The out-of-sample tests include two forecast accuracy tests considered in McCracken (2000), two forecast encompassing tests considered in Clark and McCracken (2001a), and an out-of-sample causality test proposed by Chao, Corradi and

[1] Of course, structural breaks need not be the only explanation for the gulf between in-sample and out-of-sample evidence. Inoue and Kilian (2002a) argue that power differences can explain the gap.

[2] While Estrella, Rodrigues, and Schich (2000) find a probit model relating the probability of recession to the spread to be stable, they find a linear model relating growth to the spread is not.

Swanson (2001). For comparison, our analysis includes the standard in-sample F-test for exclusion restrictions, a metric frequently used in gauging predictive ability.

Our results appear to account for the "typical" finding that in-sample predictive ability fails to translate into out-of-sample predictive ability. The paper's theoretical analysis shows in-sample explanatory power is readily found because the usual F-test indicates Granger causality or predictive ability if it existed for any portion of the sample. Out-of-sample predictive power can be harder to find because the results of out-of-sample tests are highly dependent on the timing of the predictive ability – whether the predictive ability existed at the beginning or end of the sample, and where a break occurred relative to the start of the forecast sample. Moreover, out-of-sample predictive power is harder to find with some tests than with others: the power of F-type tests of equal forecast accuracy and encompassing often dominates that of the more commonly-used t-type alternatives. Overall, out-of-sample tests are effective at revealing whether one variable has predictive power for another at the end of the sample. All of these analytical findings are confirmed by our Monte Carlo results.

The two empirical applications we consider provide further, concrete evidence that structural shifts can account for the out-of-sample breakdown in predictive power encountered in so much empirical work. Using U.S. data, we examine how identified structural breaks affect the predictive power of (1) an interest rate spread for real GDP growth and (2) growth in nominal stock prices for growth in industrial production. Simulations of models estimated with historical data show the breaks identified in the mid-1980s would produce the basic pattern documented in the sample results: in-sample causality and even out-of-sample causality in 1971-85 forecasts, but not in 1986-2000 forecasts.

The analysis in this paper builds on two extant lines of work. The first is the literature on forecasts from nested models – McCracken (2000), Chao, Corradi, and Swanson (2001), Clark and McCracken (2001a,b), Gilbert (2001), and Inoue and Kilian (2002a) – which to this point has assumed stable models. For example, McCracken (2000) and Clark and McCracken (2001a,b) derive the limiting distributions of tests of forecast accuracy and encompassing under the narrowly defined null hypothesis that one of the two models being compared is nested within the other *for all t*. Here we consider a much broader set of alternatives that allow for either no nesting or nesting for only a portion of the sample. Doing so allows us to provide details on the ability of these tests to detect a range of empirically relevant alternatives.[3]

We also build on Rossi's (2001) development of optimal tests of the *joint* null hypothesis of no (Granger) causality and model stability. Rossi uses local asymptotic methods to derive a point optimal test of the joint hypothesis, noting that out-of-sample tests also serve, in effect, as tests of the joint null of no causality and model stability.[4] Simulations are then used to analyze the finite-sample efficacy of several tests under a range of alternatives that allow for deviations from the null of model stability and nesting. The tests include the newly derived test, a t-test for equal forecast MSE (labeled MSE-t below), and other 'optimal' tests for more narrowly defined alternative hypotheses (such as the Andrews and Ploberger (1994) and Nyblom (1989) tests for structural breaks).[5] Rossi finds that the newly derived optimal test and the t-test for equal MSE are the only ones to maintain power against all alternatives, although in any particular situation,

[3] Inoue and Kilian (2002b) compare the effectiveness of forecast MSE criteria and information criteria in selecting forecasting models, in environments with and without structural breaks.

[4] Rossi (2001) does not derive any results for out-of-sample tests.

[5] Rossi (2001) compares the MSE-t test against standard normal critical values. However, McCracken (2000) has shown that, with nested models, the test is not asymptotically standard normal but instead has an asymptotic representation that is a complex function of stochastic integrals of quadratics of Brownian Motion. The asymptotically valid critical values provided by McCracken differ substantially from standard normal critical values.

the out-of-sample test is dominated by the relevant optimal test. The results of our paper substantiate and extend some of Rossi's conjectures regarding the power of out-of-sample tests.

The remainder of the paper proceeds as follows. Section 2 provides the notation and assumptions used to derive the asymptotics. Section 3 presents six Lemmas, one associated with each of the test statistics considered. Each Lemma provides an asymptotic expansion that can be used to derive the limiting behavior of the respective test statistic. This section also examines a simple example designed to illuminate the complicated asymptotics. Section 4 reports Monte Carlo evidence on the ability of out-of-sample tests, relative to the standard in-sample test, to detect alternatives that allow for a break in the causal relation of interest. Section 5 presents our analysis of the effect of structural breaks on the predictive power of an interest rate spread and stock prices for output growth. Section 6 concludes. All proofs are presented in a technical appendix, Clark and McCracken (2002).

2. Environment

The sample of observations $\{y_t, x'_{2,t}\}_{t=1}^{T+1}$ includes a scalar random variable y_t to be predicted and a ($k_1 + k_2 = k\times1$) vector of predictors $x_{2,t} = (x'_{1,t}, x'_{22,t})'$. The sample is divided into in-sample and out-of-sample portions. The in-sample observations span 1 to **R**. Letting **P** denote the number of 1-step ahead predictions, the out-of-sample observations span R + 1 through R + P. The total number of observations in the sample is R + P = T + 1.

Forecasts of $y_{t+1,}$ t = R,…,T, are generated using two linear models of the form $x'_{i,t}\beta^*_i$, i = 1,2, each of which is estimated using OLS. Under the null, model 2 nests the restricted model 1 for all t = 1,..., T+1, and hence model 2 includes k_2 excess parameters. Without loss of generality, under the null hypothesis we define $\beta^*_2 = (\beta^{*'}_{1\,1\times k_1}, 0_{1\times k_2})'$. Under the alternative hypothesis, these restrictions are not necessarily true for all t and thus the data-generating process is allowed to take the very general form $y_{t+1} = x'_{2,t}\beta^*_{2,t} + u_{t+1}$ with $Ex_{2,t}u_{t+1} \equiv Eh_{2,t+1} = 0$ for all t.

The forecasts are allowed to be either *recursive*, *rolling* or *fixed* 1-step ahead predictions.[6] Under the recursive scheme, each model's parameters are estimated with added data as forecasting moves forward through time: for t = R,…,T, model i's prediction of y_{t+1}, $x'_{i,t}\hat{\beta}_{i,t}$, is created using the parameter estimate $\hat{\beta}_{i,t}$ based on data from 1 to t. The largest number of observations used to estimate the model parameters is then T = R + P – 1. Under the rolling scheme, forecasts are constructed similarly but each model's parameters are estimated using only a rolling window of the past R observations. We continue to denote these parameter estimates as $\hat{\beta}_{i,t}$ even though it might be more appropriate to add an additional subscript R to denote the window width. Under the fixed scheme the parameters are not revised as new data becomes available. The parameter vector is estimated once using the first R observations and then remains fixed as forecasting proceeds. We continue to denote these parameter estimates as $\hat{\beta}_{i,t}$ even though it is trivially true that $\hat{\beta}_{i,t} = \hat{\beta}_{i,R}$ for all $t \geq R$.

We denote the 1-step ahead forecast errors as $\hat{u}_{1,t+1} = y_{t+1}\text{-}x'_{1,t}\hat{\beta}_{1,t}$ and $\hat{u}_{2,t+1} = y_{t+1}\text{-}x'_{2,t}\hat{\beta}_{2,t}$ for models 1 and 2, respectively. We denote the in-sample regression residuals as $\hat{v}_{1,s+1} = y_{s+1}\text{-}x'_{1,s}\hat{\beta}_{1,T}$ and $\hat{v}_{2,s+1} = y_{s+1}\text{-}x'_{2,s}\hat{\beta}_{2,T}$ s = 1,..., T for models 1 and 2, respectively. Using these forecast errors and residuals we consider the following six test statistics, described in more detail in Section 3:

[6] The analytical results are easily generalized to multi-step forecasts from long-horizon regressions. Null asymptotics for these types of predictions are given in Clark and McCracken (2001b). We do not consider them here since asymptotically valid critical values are unavailable due to the non-pivotal nature of the null asymptotic distributions.

$$\text{GC} = \text{T}\times\frac{\text{T}^{-1}\sum_{\text{s}=1}^{\text{T}}(\hat{\text{v}}_{1,\text{s}+1}^2-\hat{\text{v}}_{2,\text{s}+1}^2)}{\text{T}^{-1}\sum_{\text{s}=1}^{\text{T}}\hat{\text{v}}_{2,\text{s}+1}^2}, \qquad \text{MSE-t} = \text{P}^{1/2}\times\frac{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}(\hat{\text{u}}_{1,\text{t}+1}^2-\hat{\text{u}}_{2,\text{t}+1}^2)}{\sqrt{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}(\hat{\text{u}}_{1,\text{t}+1}^2-\hat{\text{u}}_{2,\text{t}+1}^2)^2}},$$

$$\text{MSE-F} = \text{P}\times\frac{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}(\hat{\text{u}}_{1,\text{t}+1}^2-\hat{\text{u}}_{2,\text{t}+1}^2)}{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\hat{\text{u}}_{2,\text{t}+1}^2}, \qquad \text{ENC-t} = \text{P}^{1/2}\times\frac{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}(\hat{\text{u}}_{1,\text{t}+1}^2-\hat{\text{u}}_{1,\text{t}+1}\hat{\text{u}}_{2,\text{t}+1})}{\sqrt{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}(\hat{\text{u}}_{1,\text{t}+1}^2-\hat{\text{u}}_{1,\text{t}+1}\hat{\text{u}}_{2,\text{t}+1})^2}},$$

$$\text{ENC-F} = \text{P}\times\frac{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}(\hat{\text{u}}_{1,\text{t}+1}^2-\hat{\text{u}}_{1,\text{t}+1}\hat{\text{u}}_{2,\text{t}+1})}{\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\hat{\text{u}}_{2,\text{t}+1}^2}, \qquad \text{CCS} = \text{P}\times\bar{\text{m}}'\hat{\Omega}^{-1}\bar{\text{m}},$$

where, for $\hat{\pi}$ = P/R and $\pi = \lim_{\text{P,R}\to\infty}\text{P/R}$,

$$\bar{\text{m}} = \text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\hat{\text{u}}_{1,\text{t}+1}\text{x}_{22,\text{t}}, \; \hat{\Omega} = \hat{\text{S}}_{\text{ff}} + (\hat{\lambda}_{\text{hh}} - 2\hat{\lambda}_{\text{fh}})\hat{\text{F}}\hat{\text{B}}_1\hat{\text{S}}_{\text{hh}}\hat{\text{B}}_1'\hat{\text{F}}', \; \hat{\text{F}} = -\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\text{x}_{22,\text{t}}\text{x}_{1,\text{t}}',$$

$$\hat{\text{S}}_{\text{hh}} = (\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\hat{\text{u}}_{1,\text{t}+1}^2)(\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\text{x}_{1,\text{t}}\text{x}_{1,\text{t}}'), \; \hat{\text{S}}_{\text{ff}} = (\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\hat{\text{u}}_{1,\text{t}+1}^2)(\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\text{x}_{22,\text{t}}\text{x}_{22,\text{t}}'),$$

$$\hat{\text{B}}_1 = (\text{P}^{-1}\sum_{\text{t}=\text{R}}^{\text{T}}\text{x}_{1,\text{t}}\text{x}_{1,\text{t}}')^{-1},$$

and

Scheme	$\hat{\lambda}_{\text{fh}}$	$\hat{\lambda}_{\text{hh}}$
Recursive	$1-\hat{\pi}^{-1}\ln(1+\hat{\pi})$	$2[1-\hat{\pi}^{-1}\ln(1+\hat{\pi})]$
Rolling, $\pi \le 1$	$\hat{\pi}/2$	$\hat{\pi}-\hat{\pi}^2/3$
Rolling, $1 < \pi < \infty$	$1-(2\hat{\pi})^{-1}$	$1-(3\hat{\pi})^{-1}$
Fixed	0	$\hat{\pi}$.

We use the moniker GC (Granger causality) to denote the textbook in-sample test that is asymptotically $\chi^2(\text{k}_2)$.[7] The null limiting distributions of the tests for equal MSE (MSE-F and

[7] To simplify presentation of the analytical results, we drop from the GC test formula the degrees of freedom adjustment commonly incorporated in tests of causality. In our Monte Carlo and empirical work, however, the computed GC statistics incorporate the usual degrees of freedom adjustment.

MSE-t) and forecast encompassing (ENC-F and ENC-t) are derived in McCracken (2000) and Clark and McCracken (2001a), respectively.[8] The null limiting distribution of the CCS test is derived by Chao, Corradi and Swanson (2001). To facilitate presentation we have modified the original CCS statistic by assuming homoskedastic forecast errors, which simplifies the weighting matrix Ω as discussed in West and McCracken (1998).[9]

Following Harvey, Leybourne, and Newbold (1998) and Clark and McCracken (2001a), among others, we treat the tests for equal MSE (MSE-F and MSE-t) and forecast encompassing (ENC-F and ENC-t) as one-sided tests. Clark and McCracken note that because the models are nested, the null hypothesis is that model 1's MSE is less than or equal to model 2's MSE, while the alternative is that model 1's MSE is greater than model 2's. The alternative is one-sided because, if the restrictions imposed on model 1 are not true, there is no reason to expect forecasts from model 1 to be superior to those from model 2. Harvey, Leybourne, and Newbold point out that, under the null that model 1 forecast encompasses 2, the covariance in the numerator of the encompassing tests will be less than or equal to 0. Under the alternative that model 2 contains added information, the covariance should be positive.

For the GC test as well as the fixed and recursive schemes let $B_i(t)$, $\bar{B}_i(t)$, $D_i(t)$, $\bar{D}_i(t)$ and $H_i(t)$, $i = 1,2$ and $t = R,\ldots,T$, denote $(t^{-1}\sum_{s=1}^{t-1} x_{i,s}x_{i,s}')^{-1}$, $(t^{-1}\sum_{s=1}^{t-1} Ex_{i,s}x_{i,s}')^{-1}$, $(t^{-1}\sum_{s=1}^{t-1} x_{i,s}x_{i,s}'\beta_{i,s}^{*})$, $(t^{-1}\sum_{s=1}^{t-1} Ex_{i,s}x_{i,s}'\beta_{i,s}^{*})$ and $(t^{-1}\sum_{s=1}^{t-1} x_{i,s}u_{s+1}) \equiv (t^{-1}\sum_{s=1}^{t-1} h_{i,s+1})$, respectively. For the rolling scheme let $B_i(t)$, $\bar{B}_i(t)$, $D_i(t)$, $\bar{D}_i(t)$ and $H_i(t)$, $i = 1,2$ and $t = R,\ldots,T$, denote $(R^{-1}\sum_{s=t-R+1}^{t-1} x_{i,s}x_{i,s}')^{-1}$, $(R^{-1}\sum_{s=t-R+1}^{t-1} Ex_{i,s}x_{i,s}')^{-1}$, $(R^{-1}\sum_{s=t-R+1}^{t-1} x_{i,s}x_{i,s}'\beta_{i,s}^{*})$, $(R^{-1}\sum_{s=t-R+1}^{t-1} Ex_{i,s}x_{i,s}'\beta_{i,s}^{*})$ and $(R^{-1}\sum_{s=t-R+1}^{t-1} x_{i,s}u_{s+1}) \equiv$

[8] In the theoretical analysis we use the above versions of the MSE-t and ENC-t that do not center the variances in the denominators around the mean (which is zero under the null) in order to simplify the analysis. But in our Monte Carlo and empirical work, we do in fact center the variances around the means, as is standard.

$(R^{-1}\sum_{s=t-R+1}^{t-1} h_{i,s+1})$, respectively. Finally, let J' and $\tilde{J}'$ denote the selection matrices $(I_{k_1\times k_1}, 0_{k_1\times k_2})'$ and $(0_{k_2\times k_1}, I_{k_2\times k_2})'$ and let [z] denote the integer component of the real number z.

Given the definitions and forecasting schemes described above, the following assumptions are used to derive the necessary asymptotic expansions in Lemmas 1 - 6.

Assumption 1: (a) The DGP satisfies $y_{t+1} = x_{2,t}'\beta_{2,t}^{*} + u_{t+1}$ with $Ex_{2,t}u_{t+1} \equiv Eh_{2,t+1} = 0$ for all t, (b) The nonstochastic sequence $\beta_{2,t}^{*}$ is uniformly bounded, (c) The parameters are estimated using OLS and hence satisfy $\hat{\beta}_{1,t} = B_1(t)J'D_2(t) + B_1(t)J'H_2(t)$ and $\hat{\beta}_{2,t} = B_2(t)D_2(t) + B_2(t)H_2(t)$, t = R,..., T.

Our first assumption is largely notational but is stated explicitly in order to clarify the relevant environment and insure that the parameters are estimated by OLS. That the parameter sequence is uniformly bounded insures that weighted averages of the parameters are finite and converge. Part (b) does exclude models that allow for random variation in the parameters. As a technical matter, our proofs do allow for stochastic parameters if the stochastic nature of the parameters is strongly exogenous to the process generating both y_t and $x_{2,t}$. In this case we would simply interpret our results as conditional on the given sequence of parameters.

Assumption 2: Define $U_t = [u_t, h_{2,t}', \text{vech}(x_{2,t-1}x_{2,t-1}' - Ex_{2,t-1}x_{2,t-1}')']'$. (a) For some r > 8, U_t is uniformly L^r bounded, (b) Both $\lim_{T\to\infty} T^{-1}\sum_{s=1}^{T} Eu_{s+1}^2$ and $\lim_{T\to\infty} P^{-1}\sum_{t=R}^{T} Eu_{t+1}^2$ are finite and

[9] A more general version that allows for heteroskedasticity could be used. What is important is that $\hat{\Omega}$, and its probability limit Ω, are positive definite.

positive, (c) For all t = R,...T, $\bar{B}_2^{-1}(t)$ is p.d., (d) For some r > d > 2, U_t is strong mixing with coefficients of size –rd/(r – d), (e) With $\tilde{U}_t$ denoting the vector of nonredundant elements of U_t, $\lim_{T\rightarrow\infty} T^{-1}E(\sum_{s=1}^{T}\tilde{U}_{s+1})(\sum_{s=1}^{T}\tilde{U}_{s+1})' = V < \infty$ is p.d.

Assumption 2 allows the application of an invariance principle and is sufficient for joint weak convergence of partial sums and averages of these partial sums to Brownian motion and integrals of Brownian motion. Assumption 2 is directly comparable to the assumptions in Hansen (1992) and hence we are able to apply his Theorems. Note that although we do not allow for stochastic processes with unit roots, we do not require that the process be covariance stationary. This is important since in practice it is common to use lags of the dependent variable to form predictions. If this is the case and if the parameter vector is not constant across time, $x_{2,t}$ is unlikely to be covariance stationary.[10]

Assumption 3: (a) $\lim_{P,R\rightarrow\infty} P/R = \pi$, $0 < \pi < \infty$, $\lambda = (1+\pi)^{-1}$, (b) $T^{1/2}(P/T-(1-\lambda)) = O(1)$.

Assumption 3 introduces the means by which the asymptotics are achieved. As in Ghysels and Hall (1990), West (1996), and White (2000) the limiting distribution results are derived by imposing a slightly stronger condition than simply that the sample size, T+1, becomes arbitrarily large. Here we impose the additional condition that the numbers of in-sample (R) and out-of-sample (P) observations become arbitrarily large at the same rate (i.e. $P/R \rightarrow \pi > 0$). Part (b) is a technical assumption that helps insure that certain remainder terms are bounded in probability.

3. Analytical Results

In this section we present six Lemmas, one associated with each of the test statistics. The Lemmas provide asymptotically valid expansions that can be used to describe how the GC, MSE-F, ENC-F, MSE-t, ENC-t, and CCS test statistics will behave in large samples under a particular alternative. As will become evident, these expansions can be cumbersome. Particularly for the out-of-sample tests, the expansions provide only limited detail.

It is for this reason that following each Lemma we provide a simplistic example. We follow Rossi (2001) and consider the case in which the null model is that y_{t+1} forms a zero-mean martingale difference sequence while the alternative model allows the unconditional mean to be non-zero. In particular, our data-generating process allows for one change point in the unconditional mean: from zero to $\alpha \neq 0$ or vice versa. Formally, if we let the break occur at time $[\lambda_B T]$ for some $\lambda_B \in (0, 1)$, then either: (i) $\alpha_t = \alpha \neq 0$, $1 \leq t \leq [\lambda_B T]$, and $\alpha_t = 0$, $[\lambda_B T] < t \leq T+1$; or (ii) $\alpha_t = 0$, $1 \leq t \leq [\lambda_B T]$, and $\alpha_t = \alpha \neq 0$, $[\lambda_B T] < t \leq T+1$. For brevity, we only derive the special case results using the recursive forecasting scheme. Results for the rolling and fixed schemes differ in the details but are qualitatively similar.

3.1 The GC Test

As a baseline for the out-of-sample tests we first derive the limiting behavior of the GC test under the general alternative presented in Assumption 1.

Lemma 1 (GC expansion): (a) $T^{-1}\sum_{s=1}^{T}\hat{v}_{2,s+1}^{2} \to_{p} \sigma_{v}^{2}$ a finite positive constant, (b) $\sum_{s=1}^{T}(\hat{v}_{1,s+1}^{2}-\hat{v}_{2,s+1}^{2}) = T\{[\bar{D}_2(T)]'[-J\bar{B}_1(T)J'+\bar{B}_2(T)][\bar{D}_2(T)]\} + O_p(T^{1/2})$.

[10] If lagged values of the dependent variable are used as predictors, they are included in $x_{2,t}$.

Part (a) of Lemma 1 simply states that asymptotically the scaling factor for the standard F-test is finite and non-zero. Part (b) provides the central expansion. As is well known the GC test diverges under the alternative at rate T and moreover, since $[\bar{D}_2(T)]'[-J\bar{B}_1(T)J'+\bar{B}_2(T)][\bar{D}_2(T)]$ is a symmetric quadratic form, it diverges to positive infinity.

As an example of what the expansion looks like consider the data-generating process described at the beginning of Section 3. Regardless of whether case (i) or case (ii) hold, $\bar{B}_1(T) = 0$ and $\bar{B}_2(T) = 1$. We therefore find that, in case (i), GC = $T[\alpha^2([\lambda_B T]/T)^2] + O_p(T^{1/2})$, while in case (ii), GC = $T[\alpha^2((T-[\lambda_B T])/T)^2] + O_p(T^{1/2})$. In both situations GC diverges to positive infinity at rate T. Moreover it is clear that the scale factor depends upon the magnitude of the deviation (from the null of no causality) through α^2 and the percentage of the sample in which $\alpha \neq 0$ (through either λ_B or $1 - \lambda_B$).

3.2 The MSE-F Test

Like the GC test, the F-type forecast accuracy test proposed by McCracken (2000) is comprised of two components. The numerator tests the null of equal forecast accuracy (when MSE is the measure of loss) while the denominator serves as a scale factor. We therefore have the following two-parted Lemma.

Lemma 2 (MSE-F expansion): (a) $P^{-1}\sum_{t=R}^{T}\hat{u}_{2,t+1}^2 \rightarrow_p \sigma_u^2$ a finite positive constant, (b)

$$\sum_{t=R}^{T}(\hat{u}_{1,t+1}^2-\hat{u}_{2,t+1}^2) = T(1-\lambda)\{2P^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)J'+\bar{B}_2(t)][Ex_{2,t}x_{2,t}'\beta_{2,t}^*] -$$

$$P^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)Ex_{1,t}x_{1,t}'\bar{B}_1(t)J'+\bar{B}_2(t)Ex_{2,t}x_{2,t}'\bar{B}_2(t)][\bar{D}_2(t)]\}+ O_p(T^{1/2}).$$

Once again, part (a) of this Lemma is necessary only in that it ensures that asymptotically the scale factor is finite and non-zero. The key is the expansion in part (b). It is immediately clear that the leading term is unlike that for the GC test. Although this term does diverge at rate T (presuming that the bracketed term {.} is non-zero), it need not be positive. Hence, for a test of the null hypothesis using any finite critical value from the upper tail of the null distribution, there exist alternatives that the MSE-F statistic will fail to detect.[11] Note also that the scale factor depends explicitly on the percentage of the sample used for forecast evaluation, $1 - \lambda$. Holding the bracketed term {.} constant and presuming it is positive, power increases in $1 - \lambda$.

Because it is difficult to glean too much information from the expansion in Lemma 2, we return to the simple example described at the beginning of this section. There is one difference here that was not relevant to the GC test: the precise value of the expansion depends upon the location of the forecast sample split relative to the location of the break. Therefore we derive results for both cases (i) and (ii) twice. We use (a) to denote that the break has occurred after the sample split and (b) to denote that the break has occurred before the sample split.

In case (ia) we find that MSE-F = $T[\,\alpha^2(([\lambda_B T]\text{-}R)/T) \;-\; \alpha^2(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^2([\lambda_B T]/T)^2)\,] + O_p(T^{1/2})$. Since $\alpha^2 > 0$, to determine the limiting behavior we need the sign of $(([\lambda_B T]\text{-}R)/T) \;-\; (T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^2([\lambda_B T]/T)^2)$. But for large enough T, $(([\lambda_B T]\text{-}R)/T) \;-\; (T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^2([\lambda_B T]/T)^2) \sim \lambda_B - \lambda - \lambda_B^2\int_{\lambda_B}^{1} s^{-2}ds \;=\; \lambda_B^2\text{-}\lambda$. If $\lambda_B^2\text{-}\lambda > 0$ the MSE-F test diverges to infinity, but if it is negative the MSE-F statistic diverges to negative infinity. In case (ib) we find that MSE-F = $-T[\,\alpha^2(T^{-1}\sum_{t=R}^{T}(T/t)^2([\lambda_B T]/T)^2)\,] + O_p(T^{1/2})$. Since this term is negative it must be the case that the MSE-F test diverges to negative infinity so long as $\alpha \neq 0$.

[11] The same can also be said if the lower tail of the null distribution is used instead. In this case, however, the primary class of alternatives being considered is unclear.

The case in which there is a break away from causality thus provides a situation in which we expect in-sample methods to indicate predictive ability (a large GC test) while having no out-of-sample predictive ability according to the MSE-F test. Note that for a fixed value of λ_B, this is more likely to happen when λ is large, or equivalently, when the number of out-of-sample observations (P) is small. Since this is often the case in empirical work it may be that structural breaks away from causality are what drives the common finding of in-sample, but not out-of-sample, predictive ability.

When the break is away from $\alpha = 0$ and toward $\alpha \neq 0$, the results for the simple example are much easier to interpret. Regardless of when the break occurs the test diverges to positive infinity. In particular, in case (iia) we find that MSE-F = $T[\,2\alpha^2(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)((t-[\lambda_B T])/T)) - \alpha^2(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^2((t-[\lambda_B T])/T)^2)\,]$. Since α^2 is bounded we need only derive the sign of $2(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)((t-[\lambda_B T])/T)) - (T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^2((t-[\lambda_B T])/T)^2)$. But for large enough T we obtain $2(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)((t-[\lambda_B T])/T)) - (T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^2((t-[\lambda_B T])/T)^2) \sim 2\int_{\lambda_B}^{1} s^{-1}(s-\lambda_B)ds - \int_{\lambda_B}^{1} s^{-2}(s-\lambda_B)^2 ds = (1-\lambda_B)^2$. In case (iib) we find that MSE-F = $T[\,2\alpha^2(T^{-1}\sum_{t=R}^{T}(T/t)((t-[\lambda_B T])/T)) - \alpha^2(T^{-1}\sum_{t=R}^{T}(T/t)^2((t-[\lambda_B T])/T)^2)\,]$. Since α^2 is bounded we need only derive the sign of $2(T^{-1}\sum_{t=R}^{T}(T/t)((t-[\lambda_B T])/T)) - (T^{-1}\sum_{t=R}^{T}(T/t)^2((t-[\lambda_B T])/T)^2)$. But for large enough T we have $2(T^{-1}\sum_{t=R}^{T}(T/t)((t-[\lambda_B T])/T)) - (T^{-1}\sum_{t=R}^{T}(T/t)^2((t-[\lambda_B T])/T)^2) \sim \int_{\lambda}^{1} s^{-1}(s-\lambda_B)ds - \int_{\lambda}^{1} s^{-2}(s-\lambda_B)^2 ds = 1-\lambda+\lambda_B^2-\lambda_B^2\lambda^{-1}$. In both cases (iia) and (iib) the leading term is positive and hence the MSE-F test diverges to infinity.

3.3 The ENC-F Test

Like the previous two tests, the forecast encompassing test proposed by Clark and

McCracken (2001a) is comprised of two components. The numerator tests the null of forecast encompassing while the denominator serves as a scale factor. A priori we expect the expansion of the numerator to be closely related to that for the MSE-F, because the two statistics are related by the identity $\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{2,t+1}^{2}) = \sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1}) - \sum_{t=R}^{T}(\hat{u}_{2,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1})$, or equivalently, MSE-F = ENC-F $-$ $\sum_{t=R}^{T}(\hat{u}_{2,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1})/(P^{-1}\sum_{t=R}^{T}\hat{u}_{2,t+1}^{2})$. Lemma 3 describes the limiting behavior of the ENC-F statistic.

Lemma 3 (ENC-F expansion): (a) $P^{-1}\sum_{t=R}^{T}\hat{u}_{2,t+1}^{2} \rightarrow_{p} \sigma_{u}^{2}$ a finite positive constant, (b) $\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1}) = T(1-\lambda)\{P^{-1}\sum_{t=R}^{T}[\bar{D}_{2}(t)]'[-J\bar{B}_{1}(t)J'+\bar{B}_{2}(t)][Ex_{2,t}x_{2,t}'\beta_{2,t}^{*}] - P^{-1}\sum_{t=R}^{T}[\bar{D}_{2}(t)]'[-J\bar{B}_{1}(t)Ex_{1,t}x_{1,t}'\bar{B}_{1}(t)J'+J\bar{B}_{1}(t)J'Ex_{2,t}x_{2,t}'\bar{B}_{2}(t)][\bar{D}_{2}(t)]\}+O_{p}(T^{1/2})$.

Part (a) simply repeats the first part of Lemma 2. Part (b) shows that the expansion for the ENC-F test is similar to the expansion of the MSE-F statistic. There are two differences. The first is that the lead term in the MSE-F expansion is double that for the ENC-F. The second is that the latter term in the MSE-F expansion contains the quadratic $B_{2}(t)x_{2,t}x_{2,t}'B_{2}(t)$ whereas that for the ENC-F contains the quadratic $JB_{1}(t)J'x_{2,t}x_{2,t}'B_{2}(t)$. We therefore reach many of the same conclusions here as we did before. The ENC-F statistic diverges at rate T (presuming that the bracketed term {.} is non-zero) and need not be positive. Hence for a test of the null hypothesis using any finite critical value from the upper tail of the null distribution, there exist alternatives that the ENC-F test will fail to detect. Also, as was the case for the MSE-F statistic, the scale factor depends explicitly on the percentage of the sample used for forecast evaluation. Hence, holding the bracketed term {.} constant and presuming it is positive, power increases in $1-\lambda$.

More is revealed when we consider the simple illustrative example. In case (ia) we find that ENC-F = $T[\alpha^2(([\lambda_B T]\text{-}R)/T)] + O_p(T^{1/2})$. Since $\alpha^2(([\lambda_B T]\text{-}R)/T)$ is positive the ENC-F test diverges to positive infinity. This result differs from that for the MSE-F. Recall there it was the case that the MSE-F test diverged to positive infinity if $\lambda_B^2\text{-}\lambda$ was positive. Here, since $\lambda_B - \lambda$ is positive by assumption the ENC-F test always diverges to positive infinity. This implies that the two statistics can lead to different conclusions in the same environment. In particular, if the break occurs 'soon after' the sample split, this is precisely what we expect.

In case (ib), the expansion simplifies to ENC-F = $0 + O_p(T^{1/2})$, making the ENC-F test's rate of divergence different from the MSE-F's. Some additional algebra yields an expansion that can be used to determine the rate of divergence: ENC-F = $T^{1/2}[\alpha(T^{-1/2}\sum_{t=R}^{T}([\lambda_B T]/t)y_{t+1})] + O_p(1)$. Because y_{t+1} forms a zero-mean martingale difference sequence for all $t \geq R > [\lambda_B T]$, we can apply a central limit theorem to obtain asymptotic normality of $\alpha(T^{-1/2}\sum_{t=R}^{T}([\lambda_B T]/t)y_{t+1})$.[12] We therefore have a situation in which, for large enough T, the ENC-F statistic will be larger than any finite positive constant with probability approaching 1/2 and less than any finite negative constant with probability approaching 1/2. In contrast, Section 3.2 shows that the MSE-F statistic diverges to negative infinity with probability one. Hence for this type of alternative we expect the ENC-F test to be more powerful than the MSE-F test, even though it is not consistent.

Cases (iia) and (iib) coincide with those for the MSE-F test in the sense that they both imply the ENC-F statistic diverges to positive infinity. Specifically, we find that ENC-F equals $T[\alpha^2(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)((t\text{-}[\lambda_B T])/T))] + O_p(T^{1/2})$ and $T[\alpha^2(T^{-1}\sum_{t=R}^{T}(T/t)((t\text{-}[\lambda_B T])/T))] + O_p(T^{1/2})$ in cases (iia) and (iib), respectively. Since both $\alpha^2(T^{-1}\sum_{t=R}^{T}(T/t)((t\text{-}[\lambda_B T])/T))$ and

[12] Wooldridge and White (1998) develop a CLT for dependent heterogeneous processes that can be applied here under slightly more detailed assumptions.

$\alpha^2(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)((t-[\lambda_B T])/T))$ are positive, the ENC-F test diverges to infinity.

3.4 The MSE-t Test

The numerator of the MSE-t test, a t-statistic for equal MSE developed by Diebold and Mariano (1995) and West (1996), is the same as the numerator of the MSE-F statistic. The denominator takes the form $(\sum_{t=R}^{T}(\hat{u}_{1,t+1}^2-\hat{u}_{2,t+1}^2)^2)^{1/2}$. So long as this term is non-zero with probability one we can take the ratio of the two to determine the limiting behavior of the MSE-t statistic. Since the square-root function is continuous we work with $\sum_{t=R}^{T}(\hat{u}_{1,t+1}^2-\hat{u}_{2,t+1}^2)^2$ first.

Lemma 4 (MSE-t denominator expansion): $\sum_{t=R}^{T}(\hat{u}_{1,t+1}^2-\hat{u}_{2,t+1}^2)^2$ =

$$T(1-\lambda)\{4P^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)J'+\bar{B}_2(t)](Ey_{t+1}^2x_{2,t}x_{2,t}')[-J\bar{B}_1(t)J'+\bar{B}_2(t)][\bar{D}_2(t)]$$

$$+\ 4P^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)J'+\bar{B}_2(t)](Ey_{t+1}x_{2,t}\text{vec}(x_{2,t}x_{2,t}')')[J\bar{B}_1(t)J'\bar{D}_2(t))\otimes J\bar{B}_1(t)J'\bar{D}_2(t)]$$

$$-\ 4P^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)J'+\bar{B}_2(t)](Ey_{t+1}x_{2,t}\text{vec}(x_{2,t}x_{2,t}')')\ [\bar{B}_2(t)\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]$$

$$+\ P^{-1}\sum_{t=R}^{T}[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes J\bar{B}_1(t)J'\bar{D}_2(t)]'(E\text{vec}(x_{2,t}x_{2,t}')\text{vec}(x_{2,t}x_{2,t}')')\times$$

$$[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes J\bar{B}_1(t)J'\bar{D}_2(t)]$$

$$+\ P^{-1}\sum_{t=R}^{T}[\bar{B}_2(t)\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]'(E\text{vec}(x_{2,t}x_{2,t}')\text{vec}(x_{2,t}x_{2,t}')')\ [\bar{B}_2(t)\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]$$

$$-\ 2P^{-1}\sum_{t=R}^{T}[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes J\bar{B}_1(t)J'\bar{D}_2(t)]'(E\text{vec}(x_{2,t}x_{2,t}')\text{vec}(x_{2,t}x_{2,t}')')\times$$

$$[\bar{B}_2(t)\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]\}+O_p(T^{1/2}).$$

The squared denominator term for the MSE-t statistic is extremely complicated. The most important facts of note are that it is positive semi-definite by construction and that when it is positive, it diverges to positive infinity at rate T. Because, however, the denominator is actually the square root of this term, we can determine only that the actual denominator $(\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{2,t+1}^{2})^{2})^{1/2}$ is of order $T^{1/2}$. This is important for determining how the MSE-t test will behave under any particular alternative. For example, if under a particular environment the numerator diverges at rate T while the denominator diverges at rate $T^{1/2}$, we expect the MSE-t statistic to diverge at rate $T^{1/2}$ as well. In other words, there exist alternatives under which the power of the MSE-t test will be strictly dominated by that of the MSE-F for large enough sample sizes.[13] This relationship is supported by the simulation evidence reported in Clark and McCracken (2001a), for an environment with no structural breaks.

While we can't be certain that the denominator of the MSE-t statistic always diverges at rate $T^{1/2}$, it does in our simple illustrative example. Substitution and algebra reveals that the T-order term in the expansion of Lemma 4 takes the value $T\{4\alpha^{2}T^{-1}\sum_{t=R}^{[\lambda_{B}T]}(y_{t+1}-\alpha)^{2} + \alpha^{4}(([\lambda_{B}T]-R)/T) + 4\alpha^{2}([\lambda_{B}T]/T)^{2}(T^{-1}\sum_{t=[\lambda_{B}T]+1}^{T}(T/t)^{2}y_{t+1}^{2}) + \alpha^{4}([\lambda_{B}T]/T)^{4}(T^{-1}\sum_{t=[\lambda_{B}T]+1}^{T}(T/t)^{4})\}$,

$T\{4\alpha^{2}([\lambda_{B}T]/T)^{2}(T^{-1}\sum_{t=R}^{T}(T/t)^{2}y_{t+1}^{2}) + \alpha^{4}([\lambda_{B}T]/T)^{4}(T^{-1}\sum_{t=R}^{T}(T/t)^{4})\}$,

$T\{4\alpha^{2}T^{-1}\sum_{t=[\lambda_{B}T]+1}^{T}(T/t)^{2}((t-[\lambda_{B}T])/T)^{2}(y_{t+1}-\alpha)^{2} + 4\alpha^{4}T^{-1}\sum_{t=[\lambda_{B}T]+1}^{T}(T/t)^{2}((t-[\lambda_{B}T])/T)^{2} - 4\alpha^{4}T^{-1}\sum_{t=[\lambda_{B}T]+1}^{T}(T/t)^{3}((t-[\lambda_{B}T])/T)^{3} + \alpha^{4}T^{-1}\sum_{t=[\lambda_{B}T]+1}^{T}(T/t)^{4}((t-[\lambda_{B}T])/T)^{4}\}$ and

$T\{4\alpha^{2}T^{-1}\sum_{t=R}^{T}(T/t)^{2}((t-[\lambda_{B}T])/T)^{2}(y_{t+1}-\alpha)^{2} + 4\alpha^{4}T^{-1}\sum_{t=R}^{T}(T/t)^{2}((t-[\lambda_{B}T])/T)^{2} - 4\alpha^{4}T^{-1}\sum_{t=R}^{T}(T/t)^{3}((t-[\lambda_{B}T])/T)^{3} + \alpha^{4}T^{-1}\sum_{t=R}^{T}(T/t)^{4}((t-[\lambda_{B}T])/T)^{4}\}$ respectively for cases (ia), (ib), (iia) and (iib). In all cases, the denominator term diverges at rate $T^{1/2}$ and hence the MSE-F statistic will often have power that asymptotically dominates that of the MSE-t statistic.

[13] As pointed out to us by Peter Hansen, the asymptotic power difference between the F-type and t-type test (of either equal MSE or forecast encompassing) could be eliminated by simply squaring the t-type test. Doing so, however, would create problems of interpretation. As noted above, the sensible tests are one-sided: equal MSE, for example, should only be rejected if the difference in MSE is positive. Squaring the MSE-t test would produce inappropriate rejections – ones occurring when the difference in MSE is negative.

3.5 The ENC-t Test

The numerator of the ENC-t test, a t-statistic for forecast encompassing proposed by Harvey, Leybourne, and Newbold (1998), is the same as the numerator of the ENC-F statistic.[14] The denominator takes the form $(\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1})^{2})^{1/2}$. So long as this term is non-zero with probability one we can take the ratio of the two to determine the limiting behavior of the ENC-t statistic. Once again first consider the argument $\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1})^{2}$ of the square-root function.

Lemma 5 (ENC-t denominator expansion): $\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}\text{-}\hat{u}_{1,t+1}\hat{u}_{2,t+1})^{2}$ =

$$\mathbf{T}\{\,T^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)J'+\bar{B}_2(t)](Ey_{t+1}^{2}x_{2,t}x_{2,t}')[-J\bar{B}_1(t)J'+\bar{B}_2(t)][\bar{D}_2(t)]$$

$$+\ 2T^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)J'+\bar{B}_2(t)](Ey_{t+1}x_{2,t}\text{vec}(x_{1,t}x_{1,t}')')\ [\bar{B}_1(t)J'\bar{D}_2(t)\otimes\bar{B}_1(t)J'\bar{D}_2(t)]$$

$$-\ 2T^{-1}\sum_{t=R}^{T}[\bar{D}_2(t)]'[-J\bar{B}_1(t)J'+\bar{B}_2(t)](Ey_{t+1}x_{2,t}\text{vec}(x_{1,t}x_{2,t}')')\ [\bar{B}_1(t)J'\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]$$

$$+\ T^{-1}\sum_{t=R}^{T}[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes J\bar{B}_1(t)J'\bar{D}_2(t)]'(E\text{vec}(x_{2,t}x_{2,t}')\text{vec}(x_{2,t}x_{2,t}')')\times$$

$$[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes J\bar{B}_1(t)J'\bar{D}_2(t)]$$

$$+\ T^{-1}\sum_{t=R}^{T}[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]'(E\text{vec}(x_{2,t}x_{2,t}')\text{vec}(x_{2,t}x_{2,t}')')\times$$

$$[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]$$

$$-\ 2T^{-1}\sum_{t=R}^{T}[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes J\bar{B}_1(t)J'\bar{D}_2(t)]'(E\text{vec}(x_{2,t}x_{2,t}')\text{vec}(x_{2,t}x_{2,t}')')\times$$

$$[J\bar{B}_1(t)J'\bar{D}_2(t)\otimes\bar{B}_2(t)\bar{D}_2(t)]\,\}+O_p(T^{1/2}).$$

[14] West (2001) develops the limiting distribution of the test for forecasts from non-nested models with estimated parameters.

This very complicated expansion shows that the term $\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1})^{2}$ is positive semi-definite by construction and that when it is positive, it diverges to positive infinity at rate T. This implies that $(\sum_{t=R}^{T}(\hat{u}_{1,t+1}^{2}-\hat{u}_{1,t+1}\hat{u}_{2,t+1})^{2})^{1/2}$ is of order $T^{1/2}$. Thus, as was the case with the MSE tests, there exist alternatives under which the power of the ENC-t test will be strictly worse than that of the ENC-F statistic. That is, for any finite critical value, so long as the numerator is positive we can always choose T large enough so that, under the alternative, the probability of the ENC-t statistic rejecting the null is no greater than that of the ENC-F test. This relationship is supported by the simulation evidence reported in Clark and McCracken (2001a), for an environment with no structural breaks.

While we can't be certain that the denominator of the ENC-t statistic always diverges at rate $T^{1/2}$, it does in our simple illustrative example. Substitution and algebra reveals that the T-order term in the expansion of Lemma 5 takes the value $T\{\alpha^{2}T^{-1}\sum_{t=R}^{[\lambda_B T]}(y_{t+1}-\alpha)^{2} + \alpha^{4}(([\lambda_B T]-R)/T) + \alpha^{2}([\lambda_B T]/T)^{2}(T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^{2}y_{t+1}^{2})\}$, $T\{\alpha^{2}([\lambda_B T]/T)^{2}(T^{-1}\sum_{t=R}^{T}(T/t)^{2}y_{t+1}^{2})\}$, $T\{\alpha^{2}T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^{2}((t-[\lambda_B T])/T)^{2}(y_{t+1}-\alpha)^{2} + \alpha^{4}T^{-1}\sum_{t=[\lambda_B T]+1}^{T}(T/t)^{2}((t-[\lambda_B T])/T)^{2}\}$ and $T\{\alpha^{2}T^{-1}\sum_{t=R}^{T}(T/t)^{2}((t-[\lambda_B T])/T)^{2}(y_{t+1}-\alpha)^{2} + \alpha^{4}T^{-1}\sum_{t=R}^{T}(T/t)^{2}((t-[\lambda_B T])/T)^{2}\}$ respectively for cases (ia), (ib), (iia) and (iib). In all cases, the denominator term diverges at rate $T^{1/2}$, implying the ENC-F test's power will often dominate that of the ENC-t statistic.

3.6 The CCS Test

The structure of the CCS statistic, proposed by Chao, Corradi and Swanson (2001) as an out-of-sample causality test, differs from that of the other out-of-sample tests.[15] The CCS test, like the standard in-sample GC statistic, is positive semi-definite by construction. Hence we know that if the statistic diverges, it does so to positive infinity with probability one. Even so, it is not immediately clear under what deviations from the null the CCS statistic will do so. We consider this question in the following two-parted Lemma.

Lemma 6 (CCS expansion): (a) $\hat{\Omega} \to_p \Omega$ a finite positive definite matrix, (b) $P\times\bar{m}'\hat{\Omega}^{-1}\bar{m}$ = $T(1-\lambda)\{[P^{-1}\sum_{t=R}^{T} Ex_{2,t}x'_{2,t}(\beta^*_{2,t}\text{-}J\bar{B}_1(t)J'\bar{D}_2(t))]'\tilde{J}\Omega^{-1}\tilde{J}'[P^{-1}\sum_{t=R}^{T} Ex_{2,t}x'_{2,t}(\beta^*_{2,t}\text{-}J\bar{B}_1(t)J'\bar{D}_2(t))]\}+ O_p(T^{1/2})$.

Part (a) of Lemma 6 is important only in that it ensures that the weighting matrix in the quadratic form has a finite positive definite limit. Since this is the case we immediately know from part (b) that the statistic diverges to positive infinity at rate T if for all sufficiently large sample sizes T, $\tilde{J}'P^{-1}\sum_{t=R}^{T} Ex_{2,t}x'_{2,t}(\beta^*_{2,t}\text{-}J\bar{B}_1(t)J'\bar{D}_2(t))$ is non-zero. Note that, as was the case for all the other out-of-sample tests, power increases in $1 - \lambda$.

The expansion in Lemma 6 (b) is a bit simpler to interpret than many of the other expansions. The term $\tilde{J}'P^{-1}\sum_{t=R}^{T} Ex_{2,t}x'_{2,t}(\beta^*_{2,t}\text{-}J\bar{B}_1(t)J'\bar{D}_2(t))$ is a weighted average (with weights $\tilde{J}'Ex_{2,t}x'_{2,t}$) of deviations of the true parameter sequence $\beta^*_{2,t}$ from the population-level parameter estimates $J\bar{B}_1(t)J'\bar{D}_2(t)$. If a causal linkage exists, so that $\beta^*_{22,t} \neq 0$, the CCS statistic typically diverges to positive infinity. Moreover, since structural breaks imply deviations of the true parameter sequence $\beta^*_{2,t}$ from the population-level parameter estimates $J\bar{B}_1(t)J'\bar{D}_2(t)$, the CCS statistic will generally diverge to positive infinity in the presence of structural breaks.

For more detail we return to the simple illustrative example. If we define $\Omega = \lim_{T\to\infty}$

[15] The CCS test is similar in spirit to a regression-based encompassing test proposed by Chong and Hendry (1986) and discussed in detail by West and McCracken (1998).

$P^{-1}\sum_{t=R}^{T} Ey_{t+1}^2$, the example is particularly simple for the CCS statistic. In cases (ia), (iia) and (iib) we find that the CCS statistic equals $T[\alpha^2(T/P)(([\lambda_B T]-R)/T)^2]/\Omega + O_p(T^{1/2})$, $T[\alpha^2(T/P)((T-[\lambda_B T])/T)^2]/\Omega + O_p(T^{1/2})$ and $T[\alpha^2(P/T)]/\Omega + O_p(T^{1/2})$ respectively. Each of these is positive definite and diverges at rate T. In case (ib), however, the statistic does not diverge at all but is instead bounded in probability. More precisely, since CCS = $(P^{-1/2}\sum_{t=R}^{T} y_{t+1})^2/\hat{\Omega}$ and y_{t+1} forms a zero-mean martingale difference sequence over the time frame of the summation, CCS $\to_d \chi^2(1)$. Like the ENC-F statistic, but unlike each of the other statistics considered, in case (ib) the CCS statistic does not diverge at rate T.

4. Monte Carlo Results

We use Monte Carlo simulations of two simple data-generating processes to evaluate the small-sample properties of the above out-of-sample forecast and in-sample causality tests in the presence of structural breaks.[16] In these experiments, the DGP relates the predictand y to a regressor x, with the coefficient on x subject to a structural break either toward or away from the null value of zero. The forecast tests compare predictions from an unrestricted model that includes x and a restricted model that does not. This section focuses on gauging the performance of the tests by the probability of selecting the unrestricted model over the restricted. For comparison, we add Rossi's (2001) optimal test for the joint null of no causality and stability, the Exp-W statistic, to the battery of tests considered.

[16] Simulations of a more complicated or realistic DGP based on a VAR(2) fit to quarterly growth rates of nominal GDP and M2 produced qualitatively similar results.

4.1 Experiment Design

The first DGP, denoted DGP-1, corresponds to the simple example considered in Section 3 and the basic DGP in Rossi (2001):

$$y_t = 0.5\,d_t + u_t, \tag{1}$$

where d_t is a dummy variable for the structural break and u_t is an i.i.d. standard normal random variable. In this case, the unrestricted forecasting model (model 2) takes the form $y_t = \gamma_0 + u_{1,t}$, while the restricted model (model 1) forecast is 0 for all t.

The second design, DGP-2, corresponds to one of the bivariate DGPs used by Clark and McCracken (2001a):

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0.3 & 0.5d_t \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{y,t} \\ u_{x,t} \end{pmatrix}, \tag{2}$$

where d_t is a break dummy variable and the innovations are i.i.d. standard normal random variables. In the DGP-2 experiments, we compare the predictive ability of the models $y_t = \gamma_0 + \gamma_1 y_{t-1} + u_{1,t}$ (model 1) and $y_t = \gamma_0 + \gamma_1 y_{t-1} + \gamma_2 x_{t-1} + u_{2,t}$ (model 2).

As in Section 3, we consider two classes of structural breaks in the coefficient of interest: (i) from a non-zero value for $t \le [\lambda_B T]$ to zero for $t > [\lambda_B T]$, where $[\lambda_B T]$ is the break point; and (ii) from a zero value for $t \le [\lambda_B T]$ to a non-zero value for $t > [\lambda_B T]$. The dummy variable d_t in equations (1) and (2) is defined to impose the structural break under consideration in each experiment. In case (i), the restricted (nested) model is well-specified <u>after</u> a break but misspecified before a break. In case (ii), the unrestricted (nesting) model is well-specified <u>after</u> a break but overparameterized before a break. For comparison, we report size and power results for the case of no break. In all cases, for each test we report the percentage of simulations –

based on 10,000 replications – in which the test rejects the null of no causality, equal MSE, or forecast encompassing and thereby selects the unrestricted model.

In the simulations, we generate data sets of R + P = 200 sample observations (plus any initial observations necessitated by the lag structure of the DGP) and then consider a range of break points and sample splits for forecasting.[17] Breaks are specified as occurring at five different points in the sample (a given experiment has only one break): observations 25, 50, 100, 150, or 175.[18] We report results for two different divisions of the sample into in-sample and out-of-sample portions: $P/R = 33/167 \approx 0.2$ and $P/R = 133/67 \approx 2.0$ (results for another split, $P/R = 75/125 \approx 0.6$, are qualitatively similar).[19] The forecasts are generated recursively, with the model estimates updated using added data as forecasting moves forward in time. In results not reported in the interest of brevity, using forecasts based on the rolling and fixed schemes described by West and McCracken (1998) yields qualitatively similar results.

We compare the simulated test statistics against critical values that would be valid under the null hypothesis, using a significance level of 5%. In the case of the GC test, the critical values are taken from the χ^2 distribution. Asymptotic critical values for the tests of equal MSE and forecast encompassing are taken from McCracken (2000) and Clark and McCracken (2001a), respectively. We compare the CCS statistic against its (null) asymptotic χ^2 distribution. Asymptotic critical values for the Exp-W statistic are from Rossi (2001).

[17] Any initial observations necessitated by the lag structure of the model are generated from draws of the unconditional normal distribution implied by the model.

[18] These break points correspond to λ_B taking values 0.125, 0.25, 0.50, 0.75, and 0.875.

[19] We generate the data such that the same draws from the normal distribution are used regardless of the timing of the break or forecast sample split. In particular, we generate a total of four different and independent data sets of (10,000) innovations, for the following experiments: (1) DGP-1, case (i); (2) DGP-2, case (i); (3) DGP-1, case (ii); and (4) DGP-2, case (ii).

4.2 Simulation Results

Our simulations indicate that, when no structural break occurs, the tests generally have good size and power properties. Moreover the results follow some simple patterns that are consistent with the findings of Clark and McCracken (2001a) and Rossi (2001). For example, the rows labeled “none” in Table 1’s results all show that, when no structural break occurs and the DGP satisfies (for all t) the null imposed by the restricted model, each test rejects the null in about 5 percent of the simulations. Similarly, the no-break rows labeled “none” in Table 2’s results indicate that when the DGP corresponds to the unrestricted model (for all t), the GC and ENC-F tests reject the null in virtually all of the simulations. The remainder of this section focuses on results in which a break occurs.

4.2.1 Breaks away from causality (case (i))

Table 1’s results for simulations in which a break away from causality occurs generally conform with Section 3’s analytical results. Theory implies that when a break away from the unrestricted model and toward the restricted model occurs – cases (ia) and (ib) – the out-of-sample tests are less likely than the GC test to select the unrestricted model. Our finite-sample simulations confirm this analytical result. For example, as shown in the DGP-2 results, if P/R = 0.2 and the break occurs at observation 100, the GC test selects the unrestricted model in 95.7 percent of the simulations while the MSE-F and ENC-F tests select it with frequencies of only 11.6 and 34.1 percent, respectively!

Theory also implies that the behavior of the out-of-sample tests may depend on whether the break has occurred after (case (ia)) or before (case (ib)) the sample split. More specifically, Section 3’s analytical results show that the ENC-F statistic diverges to positive infinity with probability one in case (ia), but is equally likely to diverge to plus or minus infinity in case (ib).

Accordingly, the probability of selecting the unrestricted model should rise substantially when the date of the break moves forward enough to shift the case from (ib) to (ia) – an implication borne out by the simulation results. The DGP-1 results with P/R = 2, for example, show that the percentage of trials in which the ENC-F test selects the unrestricted model soars from 17.3 percent to 88.8 percent with the shift from case (ib) to (ia).

The analytical results also imply that the MSE-F test should register a surge in probability, but not precisely at the same point as the ENC-F. In the simple analytical example, the increase will reach its peak only when the break point moves far enough past the forecast split point to make $\lambda_B^2 - \lambda$ positive. In the simulation results for P/R = 2, $\lambda_B^2 - \lambda$ becomes positive between the break points of 100 and 150. With the move of the break date from observation 100 to 150, the percentage of trials in which the MSE-F test selects the unrestricted model soars from 15.3 to 83.8 percent in the case of DGP-1 and from 18.0 to 86.2 percent in the case of DGP-2.

Finally, the simulation results for the MSE-t, ENC-t, and CCS tests also appear consistent with Section 3's analytical implications. Theory indicates the MSE-t and ENC-t tests diverge at a slower rate than their MSE-F and ENC-F counterparts. In the simulations, the –t tests are generally less likely than their –F counterparts to select the unrestricted model. For instance, as shown in the DGP-2 results, if P/R = 0.2 and the break occurs at observation 100, the MSE-t and ENC-t tests select the unrestricted model with a frequency of 2.7 and 9.9 percent, respectively, compared to 11.6 and 34.1 percent for the MSE-F and ENC-F tests. The analytical results for the simple example also imply the CCS statistic has a limiting $\chi^2(1)$ distribution in case (ib) and diverges to positive infinity in case (ia). Accordingly, in the case (ib) simulations, we find that the CCS test selects the unrestricted model in only about 5 percent of the draws.[20] But as the timing of the break moves from case (ib) to (ia), the CCS test's probability rises from 4.4 percent with the break at observation 50 to 94.8 percent with the break at observation 150, in the case of DGP-2 with P/R = 2.

4.2.2 Breaks toward causality (case (ii))

Section 3's analytical results are further corroborated by Table 2's results for simulations in which a break toward causality occurs. Asymptotic theory implies that when a break away from the restricted model toward the unrestricted occurs – cases (iia) and (iib) – the GC, MSE-F, and ENC-F statistics always diverge to positive infinity. In large samples, then, each test is equally likely to choose the unrestricted model. The simulation results bear this out. For example, as shown in the DGP-2 results, if P/R = 0.2 and the break occurs at observation 100, the GC, MSE-F, and ENC-F tests select the unrestricted model in 95.5, 96.0, and 97.1 percent of the simulations.

The analytical results also imply the MSE-t and ENC-t tests diverge at a slower rate than their MSE-F and ENC-F counterparts (in cases (iia) and (iib) as well as cases (ia) and (ib)). The simulation results indicate that, when P/R = 0.2, the –t tests are generally less likely than their –F counterparts to select the unrestricted model. For instance, as shown in the DGP-1 results, if the break occurs at observation 100, the MSE-t and ENC-t tests select the unrestricted model with a frequency of 85.8 and 92.7 percent, respectively, compared to 95.0 and 96.5 percent for the MSE-F and ENC-F tests.

[20] For DGP-1, the reported CCS results under case (ib) are numerically the same because, for this model, the restricted model's forecast error is exactly the same regardless of the breakpoint. Because the forecast is always zero, the forecast error is simply equal to y. For DGP-1, the test is then formed by projecting the forecast error on a constant.

Finally, the simulations yield a finite-sample result that seems intuitive: when the break toward causality occurs late in the sample rather than in the early or middle portions, the probabilities of selecting the unrestricted model fall substantially. With DGP-2 and P/R = 0.2, for example, the frequency with which the ENC-F statistic selects the unrestricted model drops from 66.4 percent when the break occurs at observation 150 to 28.6 percent when the break occurs at observation 175.

5. Applications

As noted in the introduction and in Rossi (2001), a substantial body of empirical research in macroeconomics and finance seems to indicate: (1) evidence of predictive ability is found much more readily in-sample than out-of-sample, and (2) most macroeconomic and financial time series suffer structural breaks. The analytical and Monte Carlo results presented above suggest that structural breaks in causal relationships may account for the difficulty in finding evidence of out-of-sample predictive ability. In this section we consider two applications to examine whether, in practice, breaks can account for these typical findings on predictability. The first application relates real GDP growth to an interest rate spread, a relationship addressed by a large number of studies, ranging from Estrella and Hardouvelis (1991) to Hamilton and Kim (2002). The second relates growth in industrial production (IP) to growth in nominal stock prices, a model considered by Stock and Watson (2001), among others.

We first show that the "usual" pattern is borne out for both of these applications. We then simulate estimated models to assess how the breaks identified in each application may affect in-sample and out-of-sample tests of predictive content. In particular, we simulate DGPs that impose no breaks and DGPs that allow for breaks. With these experiments, we first evaluate the probability each test selects the unrestricted model. These simulations allow us to assess whether breaks can generate changes in power that could lead to the pattern evident in the sample estimates. We then compare the sample test statistics against the simulated distributions to further determine whether the sample results could more reasonably result from a model with or without breaks.

5.1 Data and model selection details

In both applications, the raw data are quarterly, spanning 1953:Q1 to 2000:Q4.[21] The interest rate spread is defined as the yield on 10-year government notes less the yield on 1-year notes, in annualized percentage points. Nominal stock prices are measured using the S&P 500 index. For all variables except real GDP, the quarterly data are simple time averages of monthly data. Annualized growth rates of real GDP, industrial production, and stock prices were formed by multiplying log differences by 400.

The basic model specifications were selected by applying the SIC to full-sample model estimates, allowing different lag orders in each equation of a bivariate system as well as different lag lengths for each variable in each equation. In both applications, the SIC-determined unrestricted model for the predictand relates it to a constant, one lag of the predictand, and one lag of the causal variable. After differencing and allowing a maximum lag length of four, the estimation sample is 1954:Q2 to 2000:Q4.

5.2 Evidence of predictive content

As shown in the lower panels of Tables 3 and 4, full sample tests indicate the interest rate spread and growth in stock prices have significant explanatory power for growth in GDP and

[21] Data on real GDP, interest rates (yields at constant maturities), and the S&P 500 were obtained from the FAME database. Data on industrial production were obtained from the website of the Federal Reserve's Board of Governors.

industrial production, respectively. In both applications, the simple Granger causality test strongly rejects the null of no predictive power, as does Rossi's (2001) Exp-W statistic.

To determine whether the in-sample explanatory power of the spread and stock prices translates into out-of-sample predictive power, we follow Stock and Watson (2001) in considering recursive out-of-sample forecasts that begin in the early 1970s and by dividing the out-of-sample period in half. In particular, we consider forecast performance over 1971-85 and 1986-2000. With this sample split, P/R = 60/67 = 0.896 for 1971-85 and P/R = 60/127 = 0.472 for 1986-2000.[22] For growth in both GDP and IP, 1-step ahead forecasts are generated using two models. The unrestricted model, Model 2, relates the predictand to one lag of the predictand and one lag of the causal variable. The restricted model, Model 1, is an AR(1).

As reported in Table 3, the out-of-sample tests indicate the spread has significant predictive power for GDP growth in the first forecasting period, but little or no predictive power in the second. For 1971-85, the unrestricted model's MSE is considerably lower than the restricted model's, and all of the forecast tests reject the null of no predictive content. For 1986-2000, however, the spread ceases to have any predictive power for GDP growth, with all of the forecast tests and even a simple GC test applied to just 1986-2000 data failing to reject the null.

Table 4's IP-stock price results paint a generally, but not completely, similar story. Once again, the out-of-sample tests indicate that stock prices have significant predictive power for output growth in the first forecasting period. For 1971-85, the unrestricted model's MSE is considerably lower than the restricted model's, and all of the forecast tests reject the null of no predictive content. For 1986-2000, however, the tests for equal MSE indicate growth in stock prices fails to improve the accuracy of forecasts of IP growth (and in fact produces forecasts with

[22] For these sample splits, we compare the tests for equal MSE and encompassing against asymptotic critical values associated with $\pi = 0.8$ and $\pi = 0.4$, respectively.

much higher MSE). Yet the forecast encompassing, CCS and simple GC tests applied to 1986-2000 data suggest stock prices have some predictive content.

5.3 Evidence of structural breaks

The above patterns in predictive ability could be the result of structural breaks in the causal relationships of interest. Because applying the multiple break test methodology of Bai and Perron (1998, 2000) indicates the number of coefficient breaks in each equation is limited to one, we generally base our inferences on Andrews' (1993) extremum Wald test for a single break date. Break tests are computed for individual coefficients, sets of coefficients, and the residual variance, using White's (1980) heteroskedasticity-robust variance estimator. We compare the test statistics against Andrews' asymptotic critical values and report *p*-values computed with Hansen's (1997) asymptotic approximation.[23] For some residual variances, however, Bai and Perron's tests indicate multiple breaks, in which case inference is based on their tests. While our primary concern is with structural breaks in the causal relationship between output growth and either the spread or stock prices, the simulation approach used below dictates that we test the stability of equations for not only output growth but also the spread and stock prices. We test each equation separately.

According to our test results, both applications are marked by a break in the causal relationship of interest and in the residual variance of output growth. For the GDP growth equation, the break test results reported in Table 5 indicate one significant shift in the coefficient on the lagged spread and one shift in the residual variance. The estimated date of the coefficient shift is 1984:Q2, consistent with Stock and Watson's (2001) summary assessment of the

[23] Generating *p*-values with Hansen's (2000) fixed regressor (heteroskedastic) bootstrap yields similar results.

evidence.[24] In results not shown, when the model is reestimated imposing this break, all the other regression coefficients appear stable. The estimated date of the break in the residual variance (allowing the break in the spread coefficient) is 1982:Q4. While McConnell and Perez-Quiros (2000) and Stock and Watson (2002) date the break in the residual variance of an AR model for GDP growth a bit later, the confidence interval around our estimate is large enough to cover the later estimates.[25] The estimated GDP growth equation with break takes the form (Appendix 1 reports estimates of the residual variance)

$$\Delta \ln \mathrm{GDP}_t = \underset{(.440)}{1.854} + \underset{(.075)}{0.245}\Delta \ln \mathrm{GDP}_{t-1} + \underset{(.395)}{2.038}\mathrm{spread}_{t-1} \times d_t + \underset{(.235)}{0.347}\mathrm{spread}_{t-1} \times (1 - d_t) + u_{y,t}$$

where $d_t = 1$ for all $t \leq 1984{:}\mathrm{Q}2$ and zero otherwise. Standard errors are in parentheses.

For the IP growth equation, the break tests point to one shift in the coefficient on lagged stock price growth and two breaks in the residual variance. As shown in Table 6, of the individual regression coefficients, only the break in the stock price coefficient is significant, with an estimated date of 1984:Q1. Applying Bai and Perron's (1998, 2000) tests for multiple break points to the residual variance of the model (allowing a 1984:Q1 shift in the stock price coefficient) identifies two breaks in the error variance, at 1961:Q2 and 1981:Q4. The estimated IP growth equation with break takes the form (see Appendix 1 for estimates of the residual variance)

[24] Results based on the real-time GDP data described in Croushore and Stark (2001) display a similar breakdown in the predictive content of the spread.

[25] A number of studies have now documented a clear decline in the volatility of the U.S. economy. See Stock and Watson (2002) for a review of the evidence.

$$\Delta \ln \mathrm{IP}_t = \underset{(.622)}{1.365} + \underset{(.078)}{0.397}\Delta \ln \mathrm{IP}_{t-1} + \underset{(.027)}{0.149}\Delta \ln \mathrm{SP500}_{t-1} \times \mathrm{d}_t + \underset{(.018)}{0.039}\Delta \ln \mathrm{SP500}_{t-1} \times (1-\mathrm{d}_t) + \mathrm{u}_{y,t}$$

where $\mathrm{d}_t = 1$ for all $t \leq 1984{:}\mathrm{Q}1$ and zero otherwise. Standard errors are in parentheses.

Applying break tests to equations for the interest rate spread and growth in stock prices yields mixed results. Table 5's results for the spread equation point to one break in the set of coefficients on the lagged spread variables. Although none of the individual coefficients experience statistically significant breaks, the set of coefficients on the lagged spread terms does, in 1984:Q4.[26] Multiple break tests applied to the residual variance of the model allowing the 1984:Q4 break in the spread coefficients indicate somewhere between one and three breaks, depending on the particulars of the test settings. Visually examining a plot of the squared residuals suggests that while there may be as many as three breaks, simply imposing two breaks – one in the early 1970s and another in the mid-1980s – serves to capture most of the heterogeneity. Bai and Perron (1998, 2000) estimates of two break dates put the shifts at 1973:Q2 and 1982:Q1. To facilitate the simulation analysis below, the second break is bumped up to the date of the GDP equation's variance break, 1982:Q4 (which is within the estimated confidence interval). In contrast, Table 6's results for the stock price growth equation indicate the regression coefficients and residual variance are stable. Appendix 1 reports estimates of the spread equation and the error variance matrices with the estimated breaks.

5.4 Ability of breaks to account for sample results

To gauge whether the documented structural breaks can account for the in-sample and out-of-sample test results, we conduct Monte Carlo simulations of estimated DGPs with and without breaks. We first use the resulting simulated test statistics to assess the probability each test would reject the null hypothesis of no predictive content, by comparing the simulated statistics against 5% asymptotic critical values. In the sample results, full-sample tests and out-of-sample tests for 1971-85 indicate each causal variable of interest has significant predictive power, but the out-of-sample tests for 1986-2000 generally do not. If structural breaks can account for this pattern, the simulations of DGPs with breaks – but not simulations of stable DGPs – should show that out-of-sample test rejection rates fall off sharply in 1986-2000 compared to the first forecasting subsample.

For the designs with and without breaks, we then compare the sample test statistics against the simulated distributions of test statistics, reporting the percentage of simulated test statistics exceeded by the sample test statistic. If structural breaks can account for the sample results, the sample test statistics should be unusual compared to the "no-break" distributions of tests but not the distributions based on simulations allowing the estimated breaks. This basic approach of examining whether sample results are more consistent with one model or another has been used by Rudebusch (1993) and Kuo and Mikkola (1999, 2001) in examining the evidence of unit roots versus trend stationarity.

In these experiments, we use DGPs taken from the model estimates presented above. With stability imposed, the DGPs for the GDP-spread and IP-stock market examples correspond to the model estimates reported in Tables 5 and 6, respectively. The DGPs that allow breaks are taken from the model estimates reported in Section 5.3 and Appendix 1. Although the reported "stable model" results are based on DGPs that impose stability in not only the regression coefficients but also the residual variance matrix, simulations of DGPs with breaks in the residual variance but stable regression coefficients produce results to those reported for the "stable model" case.

[26] When the model is reestimated imposing this break, the other regression coefficients appear stable.

In each experiment, data for the full estimation period of 1954:Q2 to 2000:Q4 are generated by drawing i.i.d. normal innovations with covariance equal to the appropriate sample covariance matrix and then constructing artificial data on the predictand and causal variable using the autoregressive structure of the simulated model.[27] In the models with breaks, the covariance matrix of the innovations depends on the time period for which data are being sampled. For each draw of artificial data, we construct recursive, 1-step ahead forecasts from the unrestricted and restricted models corresponding to those used in obtaining the sample results described in Section 5.2. As in the sample estimates, we use the forecasts to construct summary and test statistics in 1971-85 and 1986-2000 subsamples.

Simulations of models without breaks indicate that, for the GDP-spread and IP-stock market applications, stable models are unlikely to generate the sample results presented in Section 5.2. The upper left panels of Tables 7 and 8 show that, in the absence of a structural break, the powers of the tests – the simulated probabilities of rejecting the null hypothesis of no causality, equal MSE, or encompassing – are about the same in 1971-85 and 1986-2000. For instance, in Table 7's GDP-spread results, the MSE-F test rejects the null of equal accuracy with a frequency of 74.5 percent over 1971-85 and a frequency of 79.7 percent over 1986-2000. Moreover, the lower left panels of Tables 7 and 8 show that, in the absence of a break, the 1971-85 sample test statistics are larger than almost all the simulated test statistics, while the 1986-2000 sample tests are smaller than almost all the simulated statistics. For example, in the GDP-spread case, the probabilities of the sample MSE-F and ENC-F statistics exceeding their simulated counterparts are more than 93 percent for 1971-85 but essentially 0 percent for 1986-2000. In the IP-stock price example, the percentages of the simulated distributions exceeded by the sample tests are

[27] The initial observations necessitated by the lag structure of the model are drawn from the unconditional normal distribution implied by the model structure that applies at the start of the sample.

generally less extreme, but the pattern of a sharp drop from 1971-85 to 1986-2000 remains. On balance, the sample statistics from the two applications seem to be unusual compared to the distribution they would likely have if the underlying DGPs were stable.

Simulations of DGPs allowing structural breaks indicate the breaks can account for the pattern of sample results. The upper right panels of Tables 7 and 8 show that, with a structural break, the powers of the forecast tests are much lower over 1986-2000 than 1971-85, with the drop-off for the MSE and CCS tests exceeding those for the encompassing tests. For example, in Table 7's GDP-spread results, the MSE-F test's power drops from 96.4 percent for 1971-85 to 1.1 percent for 1986-2000. In the same application, the ENC-F test's power falls from 99.9 percent to 59.3 percent. While the same basic pattern is evident in Table 8's results for the IP-stock market application, the fall in the power of the tests is generally less dramatic because, in the DGP for this example, the structural break greatly reduces, but does not eliminate, the causal variable's explanatory power (see the IP growth equation in Section 5.3). Moreover, the lower right panels of the tables show that, once a break is taken into account, the sample test statistics no longer fall in the tails of their simulated distributions. For example, in the GDP-spread application, the sample values of the MSE-F test for 1971-85 and 1986-2000 exceed 32.3 and 47.0, respectively, of the test's simulated distribution in each period; for the ENC-F test, the corresponding probabilities are 48.4 percent for 1971-85 and 27.4 percent for 1986-2000.

6. Conclusions

In this paper we first derive asymptotically valid expansions associated with each of six test statistics that can be used to select a model for the purposes of forecasting. The statistics we consider include the standard in-sample F-test of (Granger) causality as well as five out-of-sample tests of equal forecast accuracy and encompassing. Using these asymptotic expansions we are able to provide a description of the behavior of these tests under a range of possible alternatives. In particular we are able to provide some information on how they will behave when breaks are involved under the alternative.

We find that the MSE-F, ENC-F, MSE-t and ENC-t tests have the ability to diverge to either plus or minus infinity depending on the particular alternative. Since the upper tail of the null distribution is normally used in drawing inferences, the tests will not be able to detect certain types of alternatives. As a result, out-of-sample tests will generally be more likely than the in-sample F-test to select a restricted model when given the choice between a restricted and unrestricted one. We also find that certain out-of-sample tests dominate others in terms of power. In particular, for large enough T, the F-type out-of-sample tests, the MSE-F and ENC-F statistics, will generally have greater power than their more commonly-used t-type counterparts, the MSE-t and ENC-t tests.

We then proceed to conduct Monte Carlo simulations to examine how the analytical differences among the tests translate into finite-sample performance, and find that our simulation results corroborate our analytical findings. For example, when a break away from causality (toward the restricted model) occurs, the out-of-sample tests are less likely than an in-sample causality test to select the unrestricted model.

Finally, we take up two applications, examining how identified structural breaks affect the predictive power of (1) an interest rate spread for real GDP growth and (2) growth in nominal stock prices for growth in industrial production. Simulations of models estimated with historical data show the breaks identified in the mid-1980s would produce the basic pattern documented in the sample results: in-sample causality and even out-of-sample causality in 1971-85 forecasts, but generally not in 1986-2000 forecasts. These empirical applications seem to provide concrete evidence that structural shifts can account for the out-of-sample breakdown in predictive power encountered in so much empirical work.

Overall, out-of-sample tests appear to be effective at revealing whether one variable has predictive power for another at the end of the sample. In other words, forecast tests seem to have power for identifying the true model at the end of the sample – an objective of particular relevance for forecasting. Clearly, however, if the objective is to identify whether one variable had predictive power for another at any point in history – as in Inoue and Kilian (2002a) – in-sample tests will often be more powerful than out-of-sample tests.

Given the apparent prevalence of model instability, an important outstanding question – beyond the scope of this paper – is, what forecast methods work well in the face of instability? One approach, considered in such studies as Stock and Watson (1996), Canova (2001), and Marcellino (2002), is to allow time-varying parameters. Alternatively, based on a wide battery of results, Stock and Watson (2001) have suggested that forecast combination – a particular form of shrinkage – may be a way of overcoming instabilities. Other shrinkage approaches could also be useful. Finally, Pesaran and Timmermann (2002) have proposed a two-step method of working backward in time to identify the most recent break and then using just the post-break data to estimate a model and forecast.

Appendix 1: Additional Detail on Estimated Models with Breaks

In this appendix we report some additional detail on the models with breaks discussed in Section 5.3: estimates of the equation for the spread with a break allowed and estimates of the residual variance-covariance matrix that allow the identified breaks in variances.

The spread equation in the DGP used for simulations of the GDP-spread example allowing structural shifts takes the form

$$\begin{aligned}
\text{spread}_t &= \underset{(.053)}{0.167} - \underset{(.009)}{0.032}\Delta \ln \text{GDP}_{t-1} \\
&+ [\underset{(.100)}{1.006}\text{spread}_{t-1} - \underset{(.191)}{0.443}\text{spread}_{t-2} + \underset{(.194)}{0.545}\text{spread}_{t-3} - \underset{(.089)}{0.231}\text{spread}_{t-4}] \times \text{d}_t \\
&+ [\underset{(.104)}{1.554}\text{spread}_{t-1} - \underset{(.172)}{0.704}\text{spread}_{t-2} + \underset{(.143)}{0.112}\text{spread}_{t-3} - \underset{(.081)}{0.032}\text{spread}_{t-4}] \times (1-\text{d}_t) \ + \text{u}_{x,t}
\end{aligned}$$

where $\text{d}_t = 1$ for all $\text{t} \le 1984\text{:Q4}$ and zero otherwise. Standard errors are in parentheses.

The error covariance matrix used in the GDP-spread example with breaks is

$$\Sigma = \begin{pmatrix} 15.222 & \\ -.282 & .066 \end{pmatrix} | \, t \le 73:2, \quad \begin{pmatrix} 15.222 & \\ -.612 & .486 \end{pmatrix} | \, 73:2 < t \le 82:4, \quad \begin{pmatrix} 3.621 & \\ -.022 & .059 \end{pmatrix} | \, t > 82:4 .$$

Note that, rather than separately dating breaks in the covariance between the equation residuals, we simply allow the covariance term to shift at the variance break dates.

Finally, the error covariance matrix used in the IP-stock market example with breaks is

$$\Sigma = \begin{pmatrix} 107.361 & \\ 25.715 & 477.567 \end{pmatrix} | \, t \le 61:2, \quad \begin{pmatrix} 41.501 & \\ 8.485 & 477.567 \end{pmatrix} | \, 61:2 < t \le 81:4, \quad \begin{pmatrix} 8.141 & \\ -10.182 & 477.567 \end{pmatrix} | \, t > 81:4 .$$

References

Andrews, D.W.K., (1993): "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-56.

Andrews, D.W.K. and W. Ploberger, (1994): "Optimal Tests When a Nuisance Parameter is Present Only Under the Alternative," *Econometrica*, 62, 1383-1414.

Bai, J. and P. Perron, (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47-78.

Bai, J. and P. Perron, (2000): "Computation and Analysis of Multiple Structural-Change Models," manuscript, Boston University.

Canova, F., (2001): "G-7 Inflation Forecasts," manuscript, Universitat Pompeu Fabra.

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

Chong, Y.Y. and D.F. Hendry, (1986): "Econometric Evaluation of Linear Macro-Economic Models," *Review of Economic Studies*, 53, 671-690.

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

Clark, T.E. and M.W. McCracken, (2001b): "Evaluating Long Horizon Forecasts," manuscript, Federal Reserve Bank of Kansas City and University of Missouri-Columbia.

Clark, T.E. and M.W. McCracken, (2002): "Not-for-publication appendix to 'Forecast-Based Model Selection in the Presence of Structural Breaks', manuscript, Federal Reserve Bank of Kansas City (available from www.kc.frb.org/Econres/staff/tec.htm).

Clements, M.P. and D.F. Hendry, (1999): Forecasting Non-Stationary Economic Time Series, (MIT Press).

Croushore, D. and T. Stark, (2001): "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics*, 105, 111-30.

Diebold, F.X. and R.S. Mariano, (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-63.

Estrella, A. and G.A. Hardouvelis, (1991): "The Term Structure as a Predictor of Real Economic Activity," *Journal of Finance*, 46, 555-76.

Estrella, A., A.P. Rodrigues and S. Schich, (2000): "How Stable is the Predictive Power of the Yield Curve? Evidence from Germany and the United States," Federal Reserve Bank of New York Staff Report # 113.

Ghysels, E. and A. Hall, (1990): "A Test for Structural Stability of Euler Conditions Parameters Estimated via the Generalized Method of Moments Estimator," *International Economic Review*, 31, 355-64.

Gilbert, S. (2001): "Sampling Schemes and Tests of Regression Models," manuscript, Southern Illinois University-Carbondale.

Hamilton, J.D. and D.H. Kim, (2002): "A Re-Examination of the Predictability of Economic Activity Using the Yield Spread," *Journal of Money, Credit, and Banking*, 34, 340-60.

Hansen, B.E., (1992): "Convergence to Stochastic Integrals for Dependent Heterogeneous Processes," *Econometric Theory*, 8, 489-500.

Hansen, B.E., (1997): "Approximate Asymptotic *P* Values for Structural-Change Models," *Journal of Business and Economic Statistics*, 15, 60-67.

Hansen, B.E., (2000): "Testing for Structural Change in Conditional Models," *Journal of Econometrics*, 97, 93-116.

Harvey, D.I., S.J. Leybourne and P. Newbold, (1998): "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics*, 16, 254-59.

Hendry, D.F., (2000): "On Detectable and Non-Detectable Structural Change," *Structural Change and Economic Dynamics*, 11, 45-65.

Inoue, A. and L. Kilian, (2002a): "In-Sample or Out-of-Sample Tests of Predictability? Which One Should We Use?," manuscript, North Carolina State University.

Inoue, A. and L. Kilian, (2002b): "On the Selection of Forecasting Models," manuscript, European Central Bank.

Kilian, L. and M. Taylor, (2001): "Why is it so Difficult to Beat the Random Walk Forecast of Exchange Rates?," *Journal of International Economics*, forthcoming.

Kuo, B.S. and A. Mikkola, (1999): "Re-examining Long-Run Purchasing Power Parity," *Journal of International Money and Finance*, 17, 251-66.

Kuo, B.S. and A. Mikkola, (2001): "How Sure Are We about Purchasing Power Parity? Panel Evidence with the Null of Stationary Real Exchange Rates," *Journal of Money, Credit, and Banking*, 33, 767-89.

Marcellino, M., (2002): "Instability and Non-Linearity in the EMU," IGIER Working Paper n.211.

McConnell, M.M. and G. Perez-Quiros, (2000): "Output Fluctuations in the United States: What Has Changed Since the Early 1980's?," *American Economic Review*, 90, 1464-76.

McCracken, M.W., (2000), "Asymptotics for Out-of-Sample Tests of Causality," manuscript, University of Missouri-Columbia.

Meese, R.A. and K. Rogoff, (1983), "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?," *Journal of International Economics*, 14, 3-24.

Meese, R.A. and K. Rogoff, (1988): "Was It Real? The Exchange Rate-Interest Differential Relation Over the Modern Floating-Rate Period," *Journal of Finance*, 43, 933-948.

Nyblom, J., (1989): "Testing for the Constancy of Parameters Over Time," *Journal of the American Statistical Association*, 84, 223-230.

Paye, B.S. and A. Timmermann, (2002): "How Stable Are Financial Prediction Models? Evidence from US and International Stock Market Data," manuscript, UCSD.

Pesaran, M.H. and A. Timmermann, (2002): "Market Timing and Return Prediction under Model Instability," *Journal of Empirical Finance*, forthcoming.

Rapach, D.E. and M.E. Wohar, (2002): "Structural Change and the Predictability of Stock Returns," manuscript, Seattle University.

Rossi, B., (2001): "Optimal Tests for Nested Model Selection with Underlying Parameter Instability," manuscript, Duke University.

Rudebusch, G.D. (1993): "The Uncertain Unit Root in Real GNP," *American Economic Review*, 83, 264-71.

Stock, J.H. and M.W. Watson, (1996): "Evidence on Structural Stability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14, 11-30.

Stock, J.H. and M.W. Watson, (1999): "Business Cycle Fluctuations in U.S. Macroeconomic Time Series," in Handbook of Macroeconomics, Volume 1, J. Taylor and M. Woodford., eds., (North Holland).

Stock, J.H. and M.W. Watson, (2001): "Forecasting Output and Inflation: The Role of Asset Prices," NBER Working Paper No. 8180.

Stock, J.H. and M.W. Watson, (2002): "Has the Business Cycle Changed and Why?," manuscript, Harvard University.

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

84.

West, K.D., (2001): "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics*, 19, 29-33.

West, K.D. and M.W. McCracken, (1998): "Regression-Based Tests of Predictive Ability," *International Economic Review*, 39, 817-40.

White, H., (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-38.

White, H., (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 1067-84.

Wooldridge, J.M. and H. White, (1998): "Central Limit Theorems for Dependent, Heterogeneous Processes with Trending Moments," in Topics in Econometric Theory: The Selected Works of Halbert White, H. White ed., (Edward Elgar, Cheltenham).

Table 1: Monte Carlo Results, Break Away From Causality

DGP-1

Break point	*Case*	*GC*	*EXP-W*	*MSE-F*	*MSE-T*	*ENC-F*	*ENC-T*	*CCS*
Forecast sample split: P/R = 33/167								
none		0.049	0.056	0.060	0.065	0.056	0.066	0.044
25	*(ib)*	0.137	0.382	0.082	0.061	0.091	0.071	0.044
50	*(ib)*	0.406	0.831	0.116	0.051	0.168	0.081	0.044
100	*(ib)*	0.937	0.992	0.115	0.024	0.294	0.089	0.044
150	*(ib)*	0.999	1.000	0.071	0.008	0.353	0.092	0.044
175	*(ia)*	1.000	1.000	0.185	0.032	0.648	0.254	0.095
Forecast sample split: P/R = 133/67								
none		0.049	0.056	0.048	0.049	0.044	0.049	0.043
25	*(ib)*	0.137	0.382	0.041	0.029	0.092	0.060	0.043
50	*(ib)*	0.406	0.831	0.021	0.009	0.173	0.071	0.043
100	*(ia)*	0.937	0.992	0.153	0.056	0.888	0.647	0.279
150	*(ia)*	0.999	1.000	0.838	0.649	0.999	0.996	0.942
175	*(ia)*	1.000	1.000	0.977	0.921	1.000	1.000	0.997

DGP-2

Break point	*Case*	*GC*	*EXP-W*	*MSE-F*	*MSE-T*	*ENC-F*	*ENC-T*	*CCS*
Forecast sample split: P/R = 33/167								
none		0.051	0.061	0.059	0.072	0.056	0.072	0.046
25	*(ib)*	0.173	0.480	0.084	0.068	0.100	0.077	0.043
50	*(ib)*	0.480	0.890	0.119	0.055	0.194	0.086	0.042
100	*(ib)*	0.957	0.996	0.116	0.027	0.341	0.099	0.039
150	*(ib)*	1.000	1.000	0.071	0.010	0.421	0.109	0.037
175	*(ia)*	1.000	1.000	0.195	0.037	0.709	0.293	0.096
Forecast sample split: P/R = 133/67								
none		0.051	0.061	0.049	0.052	0.050	0.054	0.052
25	*(ib)*	0.173	0.480	0.044	0.031	0.118	0.069	0.048
50	*(ib)*	0.480	0.890	0.024	0.009	0.240	0.099	0.044
100	*(ia)*	0.957	0.996	0.180	0.066	0.932	0.727	0.302
150	*(ia)*	1.000	1.000	0.862	0.682	1.000	0.997	0.948
175	*(ia)*	1.000	1.000	0.983	0.937	1.000	1.000	0.996

Notes:
1. DGP-1 and DGP-2 are defined in equations (1) and (2) in Section 4.1. In these experiments, the dummy d_t in the DGP has value 1 up through the break point and 0 thereafter (in the 'none' row, however, the dummy is 0 throughout the sample). Accordingly, in these experiments, the coefficient on the causal variable is non-zero up to the break point and zero thereafter.
2. The coefficient break occurs at observation 25, 50, 100, 150, or 175. Case (ia), defined in Section 3, refers to the situation in which the coefficient of interest shifts from a non-zero value to zero, with the shift occurring after forecasting begins. In case (ib), the break is also from a non-zero value to zero, but the shift occurs before forecasting begins.
3. The total sample of 200 observations is divided into the two in-sample (R) and out-of-sample (P) splits indicated in the table.
4. All test statistics except Exp-W are defined in Section 2. The Exp-W statistic is described in the introduction to Section 4. In each simulation, the tests are compared against 5% critical values, from the sources detailed in Section 4.1.
5. The table reports the fraction of 10,000 simulations in which the null of no causality, equal accuracy, or forecast encompassing is rejected.

Table 2: Monte Carlo Results, Break Toward Causality

DGP-1

Break point	*Case*	*GC*	*EXP-W*	*MSE-F*	*MSE-T*	*ENC-F*	*ENC-T*	*CCS*
Forecast sample split: P/R = 33/167								
none		1.000	1.000	0.896	0.629	0.994	0.936	0.797
25	*(iib)*	1.000	1.000	0.922	0.699	0.994	0.935	0.797
50	*(iib)*	0.999	1.000	0.943	0.762	0.992	0.934	0.797
100	*(iib)*	0.936	0.993	0.950	0.858	0.965	0.927	0.797
150	*(iib)*	0.405	0.836	0.680	0.752	0.624	0.752	0.797
175	*(iia)*	0.138	0.394	0.323	0.444	0.268	0.433	0.543
Forecast sample split: P/R = 133/67								
none		1.000	1.000	0.999	0.993	1.000	1.000	1.000
25	*(iib)*	1.000	1.000	1.000	0.999	1.000	1.000	1.000
50	*(iib)*	0.999	1.000	1.000	1.000	0.999	0.999	1.000
100	*(iia)*	0.936	0.993	0.929	0.933	0.885	0.928	0.991
150	*(iia)*	0.405	0.836	0.464	0.540	0.303	0.483	0.557
175	*(iia)*	0.138	0.394	0.157	0.204	0.097	0.167	0.175

DGP-2

Break point	*Case*	*GC*	*EXP-W*	*MSE-F*	*MSE-T*	*ENC-F*	*ENC-T*	*CCS*
Forecast sample split: P/R = 33/167								
none		1.000	1.000	0.919	0.678	0.996	0.948	0.812
25	*(iib)*	1.000	1.000	0.941	0.750	0.995	0.946	0.818
50	*(iib)*	1.000	1.000	0.957	0.810	0.993	0.944	0.823
100	*(iib)*	0.955	0.996	0.960	0.887	0.971	0.936	0.834
150	*(iib)*	0.478	0.893	0.712	0.785	0.664	0.777	0.845
175	*(iia)*	0.172	0.478	0.344	0.461	0.286	0.443	0.619
Forecast sample split: P/R = 133/67								
none		1.000	1.000	0.999	0.997	1.000	1.000	1.000
25	*(iib)*	1.000	1.000	1.000	0.999	1.000	1.000	1.000
50	*(iib)*	1.000	1.000	0.999	0.999	0.999	0.999	1.000
100	*(iia)*	0.955	0.996	0.944	0.945	0.916	0.941	0.994
150	*(iia)*	0.478	0.893	0.514	0.574	0.358	0.513	0.634
175	*(iia)*	0.172	0.478	0.187	0.228	0.116	0.187	0.228

Notes:
1. DGP-1 and DGP-2 are defined in equations (1) and (2) in Section 4.1. In these experiments, the dummy d_t in the DGP has value 0 up through the break point and 1 thereafter (in the 'none' row, however, the dummy is 1 throughout the sample). Accordingly, in these experiments, the coefficient on the causal variable is zero up to the break point and non-zero thereafter.
2. The coefficient break occurs at observation 25, 50, 100, 150, or 175. Case (iia), defined in Section 3, refers to the situation in which the coefficient of interest shifts from zero to a non-zero value, with the shift occurring after forecasting begins. In case (iib), the break is also from zero to a non-zero value, but the shift occurs before forecasting begins.
3. See notes 4-6 for Table 1.

Table 3: Sample Statistics for the GDP-Spread Example

	Forecast sample statistics		
	1971-85		*1986-2000*
Model 1 MSE	19.313		4.011
Model 2 MSE	16.088		6.439
MSE-F	12.026	**	-22.629
MSE-T	1.193	**	-2.813
ENC-F	18.647	**	-0.448
ENC-T	3.595	**	-0.131
CCS	9.075	**	0.019
GC	17.520	**	0.098
	Full sample statistics		
GC	16.474	**	
EXP-W	14.194	**	

Notes:
1. In this example, the predictand is annualized growth in GDP (in percentage points). The causal variable is the spread between the yields on 10-year and 1-year government securities.
2. The 1-step ahead forecasts underlying the out-of-sample tests are generated recursively using the models described in Section 5.2. Model 1 and Model 2 refer to the restricted and unrestricted models, respectively. The full-sample causality statistics reported at the bottom of the table are based on the unrestricted model described in Section 5.2, estimated over the period 1954:Q2—2000:Q4.
3. All test statistics except Exp-W are defined in Section 2. The Exp-W statistic is described in the introduction to Section 4. The tests are compared against critical values from the sources detailed in Section 4.1. The symbols * and ** denote statistical significance at the 10 and 5 percent levels, respectively.

Table 4: Sample Statistics for the IP-Stock Market Example

	Forecast sample statistics			
	1971-85		*1986-2000*	
Model 1 MSE	60.151		7.691	
Model 2 MSE	50.800		11.777	
MSE-F	11.045	**	-20.814	
MSE-T	1.114	**	-1.432	
ENC-F	14.910	**	10.311	**
ENC-T	2.326	**	1.940	**
CCS	10.022	**	3.600	*
GC	12.466	**	7.303	**
	Full sample statistics			
GC	30.152	**		
EXP-W	18.927	**		

Notes:

1. In this example, the predictand is annualized growth in industrial production (in percentage points). The causal variable is annualized growth in the S&P 500.
2. See notes 2-3 for Table 3.

Table 5: Model Estimates and Break Test Results for the GDP-Spread Example

	Dependent variable: $\Delta \ln GDP_t$			*Dependent variable:* $spread_t$		
regressand	*coefficients*	*(s.e.)*	*break tests (p-values)*	*coefficients*	*(s.e.)*	*break tests (p-values)*
constant	1.672	(.441)	7.29 (.09) *	.167	(.058)	7.18 (.10) *
$\Delta \ln GDP_{t-1}$	.282	(.075)	5.21 (.23)	-.031	(.007)	2.25 (.72)
$spread_{t-1}$	1.076	(.282)	19.75 (.00) **	1.106	(.092)	5.05 (.24)
$spread_{t-2}$				-.482	(.180)	5.79 (.18)
$spread_{t-3}$				.518	(.183)	5.03 (.24)
$spread_{t-4}$				-.242	(.077)	2.60 (.64)
SEE	3.456			.408		
$\bar{R}^2$	.167			.822		
sets of coefficients and residual variance						
spread			19.75 (.00) **			26.25 (.00) **
all coefs.			20.53 (.00) **			26.77 (.00) **
resid. var.			26.92 (.00) **			14.98 (.00) **

Notes:

1. GDP growth is measured in annualized percentage points; the factor of 400 by which the log growth rate is scaled is ignored for simplicity. The spread is the gap between the yields on 10-year and 1-year government securities.
2. The SIC was used to determine the lag orders of the GDP and spread equations, allowing different lag lengths for each variable in each equation.
3. The sample period of estimation is 1954:Q2-2000:Q4.
4. The break test is Andrews's (1993) extremum Wald statistic, computed using a sample trim of $\pi_0 = .15$. The test statistics are compared against the asymptotic critical values provided by Andrews. The symbols * and ** denote statistical significance at the 10 and 5 percent levels, respectively. The reported asymptotic *p*-values are computed using Hansen's (1997) approximation.
5. All standard errors and test statistics are based on White's (1980) heteroskedasticity-robust variance estimator.

Table 6: Model Estimates and Break Test Results for the IP-Stock Market Example

	Dependent variable: $\Delta \ln IP_t$			*Dependent variable:* $\Delta \ln SP500_t$		
regressand	*coefficients*	*(s.e.)*	*break tests (p-values)*	*coefficients*	*(s.e.)*	*break tests (p-values)*
constant	1.115	(.647)	4.76 (.27)	5.709	(1.824)	3.38 (.48)
$\Delta \ln IP_{t-1}$	.414	(.081)	2.11 (.76)			
$\Delta \ln SP500_{t-1}$	.112	(.023)	15.71 (.00) **	.315	(.072)	5.47 (.20)
SEE	6.372			21.971		
$\overline{R}^2$	.303			.093		
sets of coefficients and residual variance						
all coefs.			19.62 (.00) **			9.97 (.10)
resid. var.			24.47 (.00) **			5.96 (.16)

Notes:

1. The growth rates of industrial production (IP) and nominal stock prices (SP500) are measured in annualized percentage points; the factor of 400 by which the log growth rates are scaled is ignored for simplicity.
2. The SIC was used to determine the lag orders of the IP and SP500 equations, allowing different lag lengths for each variable in each equation.
3. See notes 3-5 of Table 5.

Table 7: Simulation Results for the GDP-Spread Example

	Stable model		**Model with break**	
	Probability of tests rejecting null			
	Forecast sample statistics		*Forecast sample statistics*	
	1971-85	*1986-2000*	*1971-85*	*1986-2000*
MSE-F	.745	.797	.964	.011
MSE-T	.554	.524	.849	.001
ENC-F	.918	.958	.999	.593
ENC-T	.817	.836	.992	.246
CCS	.467	.457	.961	.096
GC	.676	.671	.985	.313
	Full sample statistics		*Full sample statistics*	
GC	.992		.987	
EXP-W	.980		.999	
	Percent of Simulated Test Statistics that Sample Statistics Exceed			
	Forecast sample statistics		*Forecast sample statistics*	
	1971-85	*1986-2000*	*1971-85*	*1986-2000*
Model 1 MSE	.990	.000	.692	.538
Model 2 MSE	.948	.001	.804	.552
MSE-F	.932	.000	.323	.470
MSE-T	.598	.000	.263	.400
ENC-F	.995	.003	.484	.274
ENC-T	.963	.006	.699	.273
CCS	.914	.016	.269	.087
GC	.970	.012	.453	.081
	Full sample statistics		*Full sample statistics*	
GC	.465		.419	
EXP-W	.874		.353	

Notes:
1. The *stable model* results on the left side of the table use the equations given in Table 5 as the DGP. The *model with break* results on the right side of the table use the GDP growth equation in Section 5.3 and the spread equation in the appendix as the DGP. In each simulation, artificial data for 1954-2000 are generated using draws of innovations from the normal distribution.
2. In each simulation, the 1-step ahead forecasts underlying the out-of-sample tests are generated recursively using the models described in Section 5.2, over the separate (artificial) samples of 1971-85 and 1986-2000. The full-sample causality statistics reported at the bottom of each panel are based on the unrestricted model described in Section 5.2, estimated over the (artificial) period 1954:Q2—2000:Q4.
3. The upper panel of the table reports the fraction of 10,000 simulations in which the null hypothesis of no causality, equal accuracy, or forecast encompassing is rejected, using 5% critical values from the sources detailed in Section 4.1. All test statistics except Exp-W are defined in Section 2. The Exp-W statistic is described in the introduction to Section 4.

Table 8: Simulation Results for the IP-Stock Market Example

	Stable model		Model with break	
	Probability of tests rejecting null			
	Forecast sample statistics		*Forecast sample statistics*	
	1971-85	*1986-2000*	*1971-85*	*1986-2000*
MSE-F	.888	.899	.944	.106
MSE-T	.733	.692	.819	.025
ENC-F	.983	.992	.998	.977
ENC-T	.939	.947	.991	.837
CCS	.787	.784	.933	.391
GC	.852	.855	.973	.666
	Full sample statistics		*Full sample statistics*	
GC	1.000		1.000	
EXP-W	.999		.999	
	Percent of Simulated Test Statistics that Sample Statistics Exceed			
	Forecast sample statistics		*Forecast sample statistics*	
	1971-85	*1986-2000*	*1971-85*	*1986-2000*
Model 1 MSE	.917	.000	.953	.127
Model 2 MSE	.886	.000	.978	.371
MSE-F	.638	.000	.335	.262
MSE-T	.362	.001	.253	.478
ENC-F	.821	.581	.373	.391
ENC-T	.306	.168	.100	.360
CCS	.703	.199	.412	.588
GC	.652	.366	.309	.613
	Full sample statistics		*Full sample statistics*	
GC	.495		.554	
EXP-W	.682		.543	

Notes:
1. The *stable model* results on the left side of the table use the equations given in Table 6 as the DGP. The *model with break* results on the right side of the table use the IP growth equation in Section 5.3 and the stock price equation in Table 6 as the DGP. In each simulation, artificial data for 1954-2000 are generated using draws of innovations from the normal distribution.
2. See note 3 of Table 7.