In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?*

Atsushi Inoue[†]

North Carolina State University

Lutz Kilian[‡]

University of Michigan
European Central Bank
CEPR

July 11, 2002

Abstract

It has become a standard empirical finding that significant in-sample evidence of predictability does not guarantee significant out-of-sample predictability. This is often interpreted as an indication that in-sample evidence is likely to be spurious and should be discounted. In this paper we question this conventional wisdom. We show that in-sample and out-of-sample tests of predictability are equally reliable under the null hypothesis of no predictability, provided that appropriate critical values are used. We then compare the local asymptotic power of these tests. We show that in many cases of practical interest in-sample tests have higher power than out-of-sample tests. Our results provide an alternative explanation of the comparatively weak out-of-sample evidence of predictability. We conclude that results of in-sample tests of predictability will typically be more credible than results of out-of-sample tests.

*We thank seminar participants at CORE, the European Central Bank, the 2002 European Econometric Society Meeting, Exeter, Leuven, the 2002 NBER Summer Institute, Pittsburgh, Pompeu Fabra, Southampton, Tokyo Metropolitan, Tokyo, Warwick, Waseda, Yokohama National, York and especially Valentina Corradi, Todd Clark, Frank Diebold, Robert Engle, Scott Gilbert, Clive Granger, Alastair Hall, Kirstin Hubrich, Michael McCracken, Peter Reinhard Hansen, Barbara Rossi, Norman Swanson, and Ken West for helpful discussions. The views expressed in this paper do not necessarily reflect the opinion of the ECB or its staff.

[†]Department of Agricultural and Resource Economics, Box 8109, North Carolina State University, Raleigh, NC 27695-8109. E-mail: atsushi_inoue@ncsu.edu.

[‡]European Central Bank, DG Research, Kaiserstraße 29, D-60311 Frankfurt a.M., Germany. E-mail: lkilian@umich.edu.

1 Introduction

A common problem in empirical work is that of assessing the predictability of one time series variable, given information about another variable. This problem involves a comparison of the predictive content of two nested forecast models subject to estimation uncertainty. The restricted model serves as the benchmark.[1] Our objective is to determine whether a predictive relationship exists in population. This question is conceptually different from the objective of selecting the better forecast model among the models in question, because in the latter case it may be advantageous to select a misspecified model (see Inoue and Kilian 2002).

A leading example of the type of problem we have in mind are tests of the predictability of asset returns in empirical finance. For example, there is great interest in whether variables such as the dividend-price ratio or the earnings-price ratio help to predict future stock returns or excess returns (see Fama and French 1988, Campbell and Shiller 1988a,b, Goetzmann and Jorion 1993, 1995). Another example are tests of the predictive ability of technical trading rules or tests for calendar effects in stock returns (see Sullivan, Timmermann, and White 2001; Hansen 2001; White 2000).

A closely related problem arises in international finance. There is a large literature on testing the predictability of future changes in the nominal exchange rate based on current deviations of nominal exchange rates from macroeconomic fundamentals (see, e.g., Chinn and Meese 1995, Mark 1995, Berben and van Dijk 1998, Kilian 1999, Berkowitz and Giorgianni 2001, Faust, Rogers and Wright 2003, Kilian and Taylor 2003). Similar problems arise in testing whether forward rates predict future spot rates (see Clarida and Taylor 1997, Clarida, Sarno, Taylor and Valente 2003). In closely related work, Meese and Rogoff (1983) and Meese and Rose (1991) have tested the predictability of the level of spot exchange rates based on the level of macroeconomic fundamentals. Meese and Rogoff (1988) test whether real interest rate differentials help to forecast real exchange rates.

Other applications include tests of the predictive content of advertising for consumer spending (see Ashley, Granger and Schmalensee 1980), the predictive content of money for output (see Amato and Swanson 2001), the predictive content of output-gap measures for inflation (see Clark 2000), and the predictive content of asset prices for output and inflation (see Stock and Watson 2001).

Predictability tests can be conducted based on in-sample fit or they can be based on the out-of-sample fit obtained from a sequence of recursive or rolling regressions. In the former case, we use the full sample in fitting the models of interest. Examples of in-sample tests are standard $t-$tests or F-tests. In the latter case we attempt to mimic the data constraints faced by a real-time forecaster. Examples of out-of-sample tests are tests of equal predictive accuracy and tests of forecast encompassing.

There is a folk wisdom among applied researchers that in-sample tests are biased in favor of detecting spurious predictability. This perception has led to a tendency to discount significant evidence in favor of predictability based on in-sample tests, if this evidence cannot also be

[1] This paper does not deal with forecast accuracy tests for nonnested models (see, e.g., West 1996). An example of nonnested comparisons are forecast accuracy tests involving alternative formulations of the Phillips curve (see Stock and Watson 1999).

supported by out-of-sample tests. For example, Ashley, Granger and Schmalensee (1980) insist that "a sound and natural approach" to testing predictive ability "must rely primarily on the out-of-sample forecasting performance" (p. 1149). They note that "the riskiness of basing conclusions about [Granger] causality... entirely on within-sample performance is reasonably clear" (p. 1156) and stress the likelihood of "spurious inferences ... when out of-sample verification is not employed" (p. 1165).

The purpose of this paper is to question this conventional wisdom. We note that strong in-sample evidence and weak out-of-sample evidence are not necessarily an indication that in-sample tests are not reliable. Any out-of-sample analysis based on sample-splitting involves a loss of information and hence lower power in small samples. As a result, an out-of-sample test may fail to detect predictability that exists in population, whereas the in-sample test correctly will detect it. This fact has been recently illustrated by Kilian and Taylor (2003) who provide empirical evidence that in small samples out-of-sample tests may have considerably lower power than in-sample tests, given the same test size. Examples such as this one are illustrative, but not dispositive. They underscore the need for a systematic investigation of the size and power properties of in-sample and out-of-sample tests of predictability.

In this paper, we use asymptotic theory to analyze more formally the trade-offs between in-sample tests and out-of-sample tests of predictability in terms of their size and power. We provide a formal definition of reliability in terms of the size distortion of a test. We link concerns about the reliability of predictive inference to data mining. We discuss the effect of data mining on the size of tests of predictability. We show that there are no systematic differences in the accuracy of in-sample and out-of-sample tests under the null hypothesis of no predictability, provided appropriate critical values are used. Our results overturn the conventional wisdom that out-of-sample test results are more reliable than in-sample test results. We then proceed with a comparison of the local asymptotic power of out-of-sample tests relative to in-sample tests. We show that in many cases of practical interest in-sample tests have higher power than out-of-sample tests. Our results provide an alternative explanation of the comparatively weak out-of-sample evidence of predictability in applied work.

We conclude that the empirical evidence on in-sample and out-of-sample tests of predictability needs to be reconsidered. In choosing between in-sample and out-of-sample evidence, applied researchers will need to give careful attention to the size and power properties of their predictability tests on a case-by-case basis. In practice, the finite-sample size and power properties of predictability tests may be approximated by bootstrap methods (see, e.g., Kilian 1999, Kilian and Taylor 2003). The properties of such bootstrap methods are an interesting field of study.

The remainder of the paper is organized as follows. In section 2, we discuss how data mining may affect the size of in-sample and out-of-sample tests of predictability. In section 3, we derive the asymptotic distributions of these tests under local alternatives. In section 4, we compare the power of in-sample and of out-of-sample tests against local alternatives. We distinguish between environments that are free of data mining and environments, in which data mining has occurred and the critical values have been adjusted accordingly. We conclude in section 5. The proofs are in the appendix.

2 Implications of Data Mining for the Reliability of Predictive Inference

2.1 Data Mining May Affect the Reliability of Predictability Tests

The literature is replete with warnings about unreliable in-sample inference. For example, Granger (1990, p. 3) writes: "One of the main worries about the present methods of model formulation is that the specification search procedure produces models that fit the data spuriously well, and also makes standard techniques of inference unreliable". It is important to be clear about what we mean by unreliable inference. In the context of predictive inference, the prevailing concern is that in-sample tests of predictability may spuriously indicate predictability when there is none. In this context, a predictability test would be considered unreliable if it has a tendency to reject the no predictability null hypothesis more often than it should at the chosen significance level. Formally, we define a test to be unreliable if its effective size exceeds its nominal size.

It is important to note that the mere inclusion of irrelevant variables, although it inflates in-sample fit, does not affect the reliability of in-sample tests of predictability. By construction, a t-test of predictability is designed to mimic the distribution of the test statistic under the null that the regressor is irrelevant. Similarly, as more and more irrelevant variables are included, the critical values of the F-test will increase to account for this fact. Thus, the possible inclusion of irrelevant variables has no effect on the asymptotic size of predictability tests. This point is important because it means that for a given nested forecast model comparison there is no reason to expect that in-sample tests offer any less protection against overfitting than do out-of-sample tests.

Rather Granger's concern is that in-sample inference may be rendered unreliable by specification searches that are not properly reflected in the choice of critical values. This "data mining" is said to occur when a researcher searches over alternative forecast models, but only reports results for the specification with the highest predictive content. For example, Granger (1990, p. 8) notes that: "with a limited amount of data available and a huge number of possible models there is always a possibility that, if enough models are fitted to the data, one will appear to fit very well, but in fact will not be useful". For example, data mining occurs when a researchers considers several alternative predictors, say the earnings-price ratio and the dividend-price ratio, but only reports results for the predictor that appears significant in the return regression using the standard critical values. This practice will cause the size of the test of predictability to be inflated, resulting in spurious rejections of the no-predictability null and thus overfitting relative to the true model even asymptotically. Note that it is not necessary for any one researcher to mine the data deliberately. It suffices that several researchers independently consider alternative predictors and only significant results are ultimately published. How severe the problem of data mining is depends on the context.

This discussion suggests that the properties of in-sample and out-of-sample tests of predictability will depend on whether data mining has taken place or not. We therefore will study the relative merits of in-sample and out-of-sample tests of predictability in two alternative environments: One environment that is free from data mining and corresponds to the standard assumptions used in empirical work, and another environment that is subject to systematic data mining along the lines described by Granger (1990).

2.2 Do Out-of-Sample Tests Protect Against Data Mining?

In the environment without data mining, standard critical values are adequate, and the choice between in-sample and out-of-sample tests of predictability reduces to the question of which test has higher power. We will therefore derive the asymptotic distributions of these tests under local alternatives in section 3.1 and compare their power in section 4.1.

If data mining is presumed to have occurred, in contrast, the properties of predictability tests are far less clear. There is a common perception that in the presence of data mining out-of-sample tests of predictability are more reliable than in-sample tests. It is not clear what the basis for this perception is, however. Standard critical values for both in-sample and out-of-sample tests are constructed under the presumption that no data mining has taken place. If we use these conventional critical values, neither in-sample nor out-of-sample tests will be robust against data mining and it is unclear how to rank the tests.

The lack of reliability of in-sample tests in the presence of data mining is immediately obvious. Similar problems also arise in out-of-sample inference. To see this consider the commonly used procedure of recursive predictive inference about predictability (see, e.g., Mark 1995, Kilian 1999, Faust, Rogers and Wright 2003). Recursive predictive inference means that the researcher estimates both the restricted and the unrestricted model on the first S observations of the sample and evaluates the fit of each model on observation $S+1$ for $S = R, R+1, R+2, ..., T-1$. Contrary to the conventional wisdom, this procedure offers no more protection from data mining than in-sample tests of predictability. The problem is that, when this exercise is completed, the researcher knows exactly the "out-of-sample"-performance of any given model and is free to experiment with alternative predictors prior to publication. Thus, out-of-sample inference is subject to exactly the same potential data mining problems as in-sample inference.[2]

2.3 How to Compare In-Sample and Out-of-Sample Tests in the Presence of Data Mining

Since both in-sample tests and out-of-sample tests of predictability, as currently used, are rendered unreliable by data mining, neither test can be recommended for applied work when data mining is a concern. An obvious solution to this problem is to adjust the critical values of both in-sample and out-of-sample tests to account for data mining. This proposal is in the spirit of recent work by White (2000) and by Hansen (2001), who proposed bootstrap methods for out-of-sample inference in the presence of data mining. Since White's theoretical results presume that the out-of-sample test statistic of interest has a Gaussian limit distribution, they cannot be applied to the test statistics of interest in our paper. In section 3.2. we propose a natural generalization of White's approach and derive the appropriate limit distributions for

[2]McCracken (2001) studies out-of-sample inference involving forecast models that in turn were selected based on some inconsistent model selection procedure. His methodology, however, presumes that no respecification of the forecast model occurs after the out-of-sample test is conducted. Thus, he rules out data mining of the form described here.

our out-of-sample tests. This allows us to construct critical values that are robust against data mining. We also derive the limit distribution of the in-sample tests of predictability under the assumption that data mining has taken place. This allows us to construct appropriate data-mining robust critical values for the in-sample t- and F-tests. Our analysis is a natural extension of work in classical statistics on the testing of multiple hypotheses (see, e.g., Anderson 1994, Dasgupta and Spurrier 1997; Royen 1984).[3]

Although such robust critical values for predictability tests have not been used to date in empirical work, clearly both in-sample and out-of-sample tests of predictability based on these robust critical values will be reliable and free of size distortions at least asymptotically. Thus, the choice between in-sample and out-of-sample inference again reduces to a question about relative power. We will therefore derive the asymptotic distributions of these tests under local alternatives in section 3.2 and compare their power in section 4.2.

3 The Asymptotic Distributions of Tests of Predictability under Local Alternatives

3.1 Environments without Data Mining

We begin with the results for environments that are free from data mining. Consider

$$y_t \;=\; \gamma' x_t + u_t \;=\; \alpha' v_t + \beta' w_t + u_t, \tag{1}$$

where $x_t = (v_t', w_t')'$, v_t and w_t are m, l and k dimensional vectors of regressors, respectively, $\{u_t\}$ is a sequence of martingale differences. Throughout the rest of the paper we assume the following standard conditions hold:

Assumption 1.

(a) $E(u_t|x_t, u_{t-1}, x_{t-1}, u_{t-2}, ...) = 0$ a.s. for all t.

(b) $E(u_t^2|x_t, u_{t-1}, x_{t-1}, u_{t-2}, ...) = \sigma^2$ a.s. for all t.

(c) $E(u_t^4|x_t, u_{t-1}, x_{t-1}, u_{t-2}, ...) < K_4 < \infty$ a.s. for all t, where K_4 is some constant.

(d) $\{x_t\}$ is strictly stationary and ergodic with $E(x_t x_t')$ positive definite and has four finite moments.

Assumptions 1(a)(b)(c) are used for the functional central limit theorem (e.g., Stock, 1994, Theorem 1). Assumption 1(d) rules out some important and interesting cases, such as the near-cointegrated case analyzed by Rossi (2001a) (also see Corradi, Swanson and Olivetti 2001). To make our point, however, these simple assumptions suffice.

We are interested in testing

$$H_0: \;\beta \;=\; 0_{k\times 1}$$

[3]A similar framework has also been used by Hansen (2000) who proposed bootstrap inference for the distribution of R^2 in the presence of data mining.

against the two-sided alternative

$$H_1 : \beta \neq 0_{k\times 1}.$$

For some applications one is interested in testing for H_0: $\beta_1 = 0_{1\times 1}$against a more specific one-sided alternative

$$H_2 : \beta_1 > 0_{1\times 1}.$$

with $k = 1$. Both H_1 and H_2 are potentially relevant for empirical work, but, in practice, more often than not H_2 is the economically interesting alternative hypothesis.

For each alternative hypothesis we consider a sequence of local alternatives. That is,

$$H_1' : \beta = T^{-1/2}c$$

and

$$H_2' : \beta_1 = T^{-1/2}c_1,$$

respectively, where c is a k-dimensional nonzero vector and c_1 is a positive number.

We will define in-sample and out-of-sample test statistics. First we consider in-sample tests. For testing against H_1, we consider an in-sample F test statistic

$$S_1 = \frac{\sum_{t=1}^{T}(\hat{u}_{0t}^2 - \hat{u}_{1t}^2)}{\hat{\sigma}^2} \tag{2}$$

where $\{\hat{u}_{1t}\}$ are the unrestricted OLS residuals, $\{\hat{u}_{0t}\}$ are the restricted OLS residuals under H_1 and $\hat{\sigma}^2 = (1/T)\sum_{t=1}^{T}\hat{u}_{1t}^2$. For testing against H_2, we consider the in-sample t test statistic

$$S_2 = \sqrt{T}\hat{\beta}_1/\hat{\sigma}_{\beta_1} \tag{3}$$

where $\hat{\sigma}_{\beta_1}$ is the $(l+1)\times(l+1)$ element of $\hat{\sigma}^2((1/T)\sum_{t=1}^{T}x_t x_t')^{-1}$. The following proposition provides the asymptotic distributions of S_1 and S_2 under a sequence of local alternatives.

Proposition 1. Suppose that Assumption 1 holds. Under a sequence of local alternatives of the form H_1' and H_2', respectively,

$$S_1 \xrightarrow{d} \chi^2_{\delta_1}(k) \tag{4}$$

$$S_2 \xrightarrow{d} N(\delta_2, 1), \tag{5}$$

where $\chi^2_{c_1}(k)$ is a noncentral χ^2 random variable with degree of freedom k and noncentral parameters

$$\delta_1 = \frac{1}{\sigma^2}c'\{E(w_t w_t') - E(w_t v_t')[E(v_t v_t')]^{-1}E(v_t w_t')\}c$$

and

$$\delta_2 = e_{l+1}'(1/\sigma_{\beta_1})[E(x_t x_t')]^{-1}E(x_t w_t')c_1,$$

where e_{l+1} is an $(l+1)$-dimensional column vector, whose last element is one and other elements are all zero, and $\sigma^2_{\beta_1}$ is the $(l+1) \times (l+1)$ element of $\sigma^2[E(x_t x_t')]^{-1}$.

The proof of Proposition 1 is straightforward and thus is omitted. The null limit distribution emerges as a special case of this result with $c = \mathbf{0}$ and $c_1 = 0$, respectively.

Next, we consider the recursive test. We fit the model by OLS on the first S observations and evaluate the fit (loss) on observation $S+1$, for $S = R, R+1, R+2, ..., T-1$. The recursive OLS estimators are defined by $\hat{\gamma}_t = [\hat{\alpha}_t', \hat{\beta}_t']' = (\sum_{s=1}^t x_s x_s')^{-1} \sum_{s=1}^t x_s y_s$, and $\tilde{\alpha}_t = (\sum_{s=t}^t v_s v_s')^{-1} \sum_{s=1}^t v_s y_s$. For the split-sample test, we fit the model on the first S observations and evaluate the fit on the remaining $T-S$ observations. Thus for the sample-split test $S = R$. The split-sample OLS estimator is defined by $\hat{\gamma}_R = [\hat{\alpha}_R', \hat{\beta}_R']' = (\sum_{t=1}^R x_t x_t')^{-1} \sum_{t=1}^R x_t y_t$, and $\tilde{\alpha}_R = (\sum_{t=1}^R v_t v_t')^{-1} \sum_{t=1}^R v_t y_t$. Specifically, we consider the split-sample and recursive version of the F statistic:

$$S_3 = \frac{\sum_{t=R+1}^T (\tilde{u}_{0t}^2 - \tilde{u}_{1t}^2)}{\hat{\sigma}^2} \tag{6}$$

$$S_4 = \frac{\sum_{t=R+1}^T (\bar{u}_{0t}^2 - \bar{u}_{1t}^2)}{\hat{\sigma}^2} \tag{7}$$

where $\tilde{u}_{0t} = y_t - \hat{\alpha}_R' v_t$, $\tilde{u}_{1t} = y_t - \hat{\gamma}_R' x_t$, $\bar{u}_{0t} = y_t - \hat{\alpha}_{t-1}' v_t$, $\bar{u}_{1t} = y_t - \hat{\gamma}_{t-1}' x_t$. McCracken (1999) calls (6) the modified OOS-F statistics. Gilbert (2001) also consider this statistic. Thus we will refer to (6) as the Gilbert-McCracken (GM) test statistic. The GM test statistic is similar to the DM test statistic of Diebold and Mariano (1995) in that it is based on the loss differential. The key difference is that the GM test uses a different normalization designed to account for parameter estimation uncertainty in the forecast model, whereas the DM test is designed for forecast models with known parameters.

Assumption 2. $R/T \to \pi \in (0,1)$ as $T \to \infty$.

Proposition 2. Suppose that Assumptions 1 and 2 holds. Under a sequence of local alternatives of the form H_1'

$$S_3 \xrightarrow{d} \frac{1}{\pi}(W(1) - W(\pi) + (1-\pi)\delta)'(W(\pi) + \pi\delta) - \frac{1-\pi}{2\pi^2}(W(\pi) + \pi\delta)'(W(\pi) + \pi\delta) \tag{8}$$

$$\begin{aligned} S_4 \quad \xrightarrow{d} \quad & \int_\pi^1 \frac{1}{r} W'(r) dW(r) + \delta' \int_\pi^1 \frac{1}{r} W(r) dr + \delta'(W(1) - W(\pi)) + (1-\pi)\delta'\delta \\ & -\frac{1}{2} \int_\pi^1 \frac{1}{r^2}(W(r) + r\delta)'(W(r) + r\delta) dr \end{aligned} \tag{9}$$

where $W(\cdot)$ is a k-dimensional standard Brownian motion, $\delta = (1/\sigma) L' E(x_t x_t')^{-1/2} E(x_t w_t') c$, L

is a $l \times k$ matrix that satisfies $LL' = Q$ and $L'L = I_k$, and

$$Q = [E(x_t x_t')]^{\frac{1}{2}} \left\{ (E(x_t x_t'))^{-1} - \left[\begin{array}{cc} (E(v_t v_t'))^{-1} & 0_{l\times k} \\ 0_{k\times l} & 0_{k\times k} \end{array} \right] \right\} [E(x_t x_t')]^{\frac{1}{2}}.$$

(8) and (9) are due to McCracken (1999, Theorem 4.1). The null limit distribution is a special case of (8) and (9) with $c = \mathbf{0}$.

An alternative approach to comparing the forecast accuracy of two models is the encompassing test, which involves running the regressions:

$$\begin{aligned} y_t &= \alpha' v_t + u_{0t} \\ y_t &= \alpha' v_t + \beta' w_t + u_{1t} \end{aligned}$$

and testing the null hypothesis

$$E(u_{0t}^2) - E(u_{0t}u_{1t}) = 0$$

against the alternative

$$E(u_{0t}^2) - E(u_{0t}u_{1t}) > 0.$$

Clark and McCracken (2001a) consider the encompassing tests (10) and (11). Their test for nested forecast models differ from the test in Harvey et al. (1994) in that it allows for parameter estimation uncertainty. West (2001) considers the encompassing test for nonnested models with estimation uncertainty. The split-sample and recursive versions of the encompassing test are

$$S_5 = \frac{(T-R)^{-1/2}\sum_{t=R+1}^{T}(\tilde{u}_{0t}^2 - \tilde{u}_{0t}\tilde{u}_{1t})}{\hat{V}^{1/2}(\tilde{u}_{0t}^2 - \tilde{u}_{0t}\tilde{u}_{1t})}, \tag{10}$$

$$S_6 = \frac{(T-R)^{-1/2}\sum_{t=R+1}^{T}(\bar{u}_{0t}^2 - \bar{u}_{0t}\bar{u}_{1t})}{\hat{V}^{1/2}(\bar{u}_{0t}^2 - \bar{u}_{0t}\bar{u}_{1t})}, \tag{11}$$

where $\hat{V}(\tilde{u}_{0t}^2 - \tilde{u}_{0t}\tilde{u}_{1t})$ is the sample variance of $\{\tilde{u}_{0t}^2 - \tilde{u}_{0t}\tilde{u}_{1t}\}_{t=R+1}^{T}$.

Proposition 3. Suppose that Assumptions 1 and 2 hold. Under a sequence of local alternatives of the form H_1'

$$S_5 \xrightarrow{d} \frac{(W(\pi)+\pi\delta)'[W(1)-W(\pi)+(1-\pi)\delta]}{\{(1-\pi)(W(\pi)+\pi\delta)'(W(\pi)+\pi\delta)\}^{1/2}}, \tag{12}$$

$$\begin{aligned} S_6 \quad \xrightarrow{d} \quad & \left[\int_\pi^1 \frac{1}{r}W'(r)dW(r) + \delta'\int_\pi^1 \frac{1}{r}W(r)dr + \delta'(W(1)-W(\pi)) + (1-\pi)\delta'\delta\right] \\ & / \left[\int_\pi^1 \frac{1}{r^2}(W(r)+r\delta)'(W(r)+r\delta)dr\right]^{1/2} \end{aligned} \tag{13}$$

Clark and McCracken (2001a, Theorem 3.1) derive the null limit distribution of the encompassing test.[4] The null limit distribution is obtained from (12) and (13) by setting $c = \mathbf{0}$.

3.2 Environments with Data Mining

As discussed in section 2, there are many situations in empirical work, in which data mining must be presumed to have occurred. The presence of data mining affects the distribution of in-sample and out-of-sample tests of predictability both under the null and under the alternative. In this section we derive suitable critical values for predictability tests that account for data mining and we derive their local asymptotic power. We formalize data mining as follows. For $j = 1, 2, ..., M$, let $w_{j,t}$ denote a k_j-dimensional subvector of w_t where $1 \leq k_j \leq k$. Let $x_{j,t} = [v_t' \; w_{j,t}']'$. Define Q_j and L_j as in Proposition 2 with x_t replaced by $x_{j,t}$ and k replaced by k_j. Suppose that one is interested in whether any one of M models

$$y_t = \alpha' v_t + \beta_j' w_{j,t} + u_{j,t} \quad j = 1, 2, ..., M$$

has predictive power that is superior to the benchmark model

$$y_t = \alpha' v_t + u_{0,t}.$$

Formally the null hypothesis can be written as

$$H_0 : \; \beta_j = 0 \;\; \forall j$$

and the alternative hypothesis as

$$H_1 : \quad \beta_j \neq 0 \text{ for some } j.$$

Under the null hypothesis we have

$$\max_{j \in \{1,...,M\}} E(u_{0,t}^2) - E(u_{j,t}^2) = 0,$$

whereas under the alternative hypothesis we have

$$\max_{j \in \{1,...,M\}} E(u_{0,t}^2) - E(u_{j,t}^2) > 0.$$

This suggests the following in-sample test statistic:

$$S_7 \;=\; \max_{j \in \{1,...,M\}} \frac{\sum_{t=1}^{T} \hat{u}_{0,t}^2 - \hat{u}_{j,t}^2}{\hat{\sigma}_j^2}. \tag{14}$$

[4] Theorem 3.1 of Clark and McCracken (2001) reports the null limit distribution of the recursive case only. The null limit distribution of the split-sample case follows from the intermediate results in Clark and McCracken (2000).

The statistic S_7 is the data-mining proof version of S_1.

Let

$$\Omega = \begin{bmatrix} \Omega_{11} & \cdots & \Omega_{M1} \\ \vdots & \Omega_{ij} & \vdots \\ \Omega_{1M} & \cdots & \Omega_{MM} \end{bmatrix},$$
$$d = \begin{bmatrix} d_1 & d_2 & \cdots & d_M \end{bmatrix}',$$

where

$$\Omega_{ij} = L_i' E(x_{i,t}x_{i,t}')^{-1/2} E(x_{i,t}x_{j,t}) E(x_{j,t}x_{j,t}')^{-1/2} L_j,$$
$$d_j = (1/\sigma) L_j' [E(x_{j,t}x_{j,t}')]^{-1/2} E(x_{j,t}w_{j,t}') c(\zeta_j),$$

ζ_j denotes a k-dimensional selection vector whose jth element is 1 if $w_{j,t}$ includes the jth element of w_t and is zero otherwise. The notation $c(\zeta)$ stands for a subset of c where the selection vector $\zeta \in \times_{i=1}^{k}\{0,1\}$ and $1 \leq k_j = \zeta'\zeta \leq k$. For example, if $\zeta = [0,1,0,1,0]$ then $c(\zeta) = [c_2, c_4]'$.

Assumption 3. Ω is positive definite.

Proposition 4. Suppose that Assumptions 1 and 3 hold. Under the sequence of local alternatives of the form H_1',

$$S_7 \xrightarrow{d} \max_{j \in \{1,\ldots,M\}} \chi^2_{d_j}(k_j) \tag{15}$$

where $\chi^2_{d_1}(k_1) = \sum_{i=1}^{k_1} u_i^2$, $\chi^2_{d_j}(k_j) = \sum_{i=k_1+\cdots+k_{j-1}+1}^{k_1+\cdots+k_j} u_i^2$ for $j = 2, ..., M$, and

$$\begin{bmatrix} u_1 \\ \vdots \\ u_{k_1+\cdots+k_M} \end{bmatrix} \sim N(d, \Omega).$$

The distribution of $[\chi^2_{d_1}(k_1), \chi^2_{d_2}(k_2), ..., \chi^2_{d_M}(k_M)]'$ is called a noncentral multivariate χ^2 distribution in the statistical literature (see Royen (1997) for a recent survey).

Next, consider the special case in which $w_{j,t}$ is the jth element of w_t. The alternative hypothesis of interest is

$$H_2 : \beta_j > 0 \text{ for some } j,$$

and a sequence of local alternatives is

$$H_2' : \beta_j = T^{-1/2} c_j \;\; j = 1, 2, ..., M,$$

where $c_j \geq 0$ for all j and $c_j > 0$ for at least one j. For H_2 it is natural to consider the following in-sample test

$$S_8 = \max_{j \in \{1,\ldots,M\}} \sqrt{T}\hat{\beta}_j / \hat{\sigma}_{\beta_j} \tag{16}$$

where $\hat{\sigma}^2_{\beta_j}$ is a consistent estimator of the asymptotic variance of β_j. S_8 is the data-mining-proof version of S_2. Let Σ be an $M \times M$ matrix whose (i, j)th element is given by

$$\sigma_{ij} = \sigma^2 e'_{l+1}[E(x_{i,t}x'_{i,t})]^{-1}E(x_{i,t}x_{j,t})[E(x_{j,t}x'_{j,t})]^{-1}e_{l+1}/\sigma_{\beta_i}\sigma_{\beta_j}$$

and f is an M-dimensional column vector whose jth element is given by

$$f_j \;=\; e'_{l+1}[E(x_{j,t}x'_{j,t})]^{-1}E(x_{j,t}w_{j,t})c_j/\sigma_{\beta_j}.$$

where e_{l+1} is an $(l+1)$-dimensional column vector whose last element is one and other elements are all zero.

Assumption 4. Σ is positive definite.

Proposition 5. Suppose that Assumptions 1 and 4 hold. Under the sequence of local alternatives of the form H'_2

$$S_8 \overset{d}{\to} \max_{j\in\{1,\ldots,M\}}(v_j) \tag{17}$$

where

$$\begin{bmatrix} v_1 \\ \vdots \\ v_M \end{bmatrix} \sim N(f, \Sigma).$$

For $c = \mathbf{0}$, (15) and (17) include the null distribution as a special case. Thus, we are able to derive critical values that are robust against data mining. Note that a key assumption underlying this strategy is that the regression models in question are linear. Under nonlinearity the number of models that could be constructed from a given set of predictors would be infinite and the proposed strategy for computing data-mining robust critical values would be infeasible.

Next, we consider the data-mining robust out-of-sample tests. Define the maximum version of the split-sample GM test, recursive GM test, split-sample encompassing test and recursive encompassing test as follows:

$$S_9 \;=\; \max_{j\in\{1,\ldots,M\}} \frac{\sum_{t=R+1}^{T}(\tilde{u}^2_{0t} - \tilde{u}^2_{j,t})}{\hat{\sigma}^2_j} \tag{18}$$

$$S_{10} \;=\; \max_{j\in\{1,\ldots,M\}} \frac{\sum_{t=R+1}^{T}(\bar{u}^2_{0t} - \bar{u}^2_{j,t})}{\hat{\sigma}^2_j}, \tag{19}$$

$$S_{11} \;=\; \max_{j\in\{1,\ldots,M\}} \frac{(T-R)^{-1/2}\sum_{t=R+1}^{T}(\tilde{u}^2_{0t} - \tilde{u}_{0t}\tilde{u}_{j,t})}{\hat{V}^{1/2}(\tilde{u}^2_{0t} - \tilde{u}_{0t}\tilde{u}_{j,t})}, \tag{20}$$

$$S_{12} \;=\; \max_{j\in\{1,\ldots,M\}} \frac{(T-R)^{-1/2}\sum_{t=R+1}^{T}(\bar{u}^2_{0t} - \bar{u}_{0t}\bar{u}_{j,t})}{\hat{V}^{1/2}(\bar{u}^2_{0t} - \bar{u}_{0t}\bar{u}_{j,t})}. \tag{21}$$

Proposition 6. Suppose that Assumptions 1, 2 and 3 hold. Then under a sequence of local alternatives of the form H_1'

$$S_9 \xrightarrow{d} \max_{j\in\{1,\ldots,M\}} \frac{1}{\pi}(B_j(1)-B_j(\pi)+(1-\pi)d_j)'(B_j(\pi)+\pi d_j) - \frac{1-\pi}{2\pi^2}(B_j(\pi)+\pi d_j)'(B_j(\pi)+\pi d_j) \tag{22}$$

$$S_{10} \xrightarrow{d} \max_{j\in\{1,\ldots,M\}} \int_\pi^1 \frac{1}{r}B_j'(r)dB_j(r) + d_j{}' \int_\pi^1 \frac{1}{r}B_j(r)dr + d_j'(B_j(1)-B_j(\pi)) + (1-\pi)d_j{}'d_j$$
$$-\frac{1}{2}\int_\pi^1 \frac{1}{r^2}(B(r)+rd_j)'(B_j(r)+rd_j)dr, \tag{23}$$

$$S_{11} \xrightarrow{d} \max_{j\in\{1,\ldots,M\}} \frac{(B_j(\pi)+\pi d_j)'[B_j(1)-B_j(\pi)+(1-\pi)d_j]}{\{(1-\pi)(B_j(\pi)+\pi d_j)'(B_j(\pi)+\pi d_j)\}^{1/2}}, \tag{24}$$

$$S_{12} \xrightarrow{d} \max_{j\in\{1,\ldots,M\}} \left[\int_\pi^1 \frac{1}{r}B_j'(r)dB_j(r) + d_j' \int_\pi^1 \frac{1}{r}B_j(r)dr + d_j'(B_j(1)-B_j(\pi)) + (1-\pi)d_j'd_j\right]$$
$$/\left[\int_\pi^1 \frac{1}{r^2}(B_j(r)+rd_j)'(B_j(r)+rd_j)dr\right]^{1/2}, \tag{25}$$

where $\{B_j(\cdot)\}_{j=1}^M$ are Brownian motions that satisfy $E(B_i(r)B_j(s)') = \min(r,s)\Omega_{ij}$. By setting $c = \mathbf{0}$, (22), (23) , (24) and (25) include the null limit distribution as a special case.

4 Comparing the Power of In-Sample Tests and Out-of-Sample Tests of Predictability

The limit distributions we derived in section 3 will in general be data dependent. In practice, the process that generates the data will be unknown, but may be approximated by bootstrap methods (see e.g., White 2000; Hansen 2000, 2001). Since our main focus in this paper are the asymptotic properties of predictability tests, in this section we focus on a stylized example that is similar to processes studied in the empirical literature. We set aside for future research a detailed investigation of the small-sample properties of bootstrap versions of predictability tests. We note, however, that existing simulation evidence on the small-sample properties of bootstrap predictability tests (e.g., Kilian 1999) is fully consistent with our local asymptotic analysis.

4.1 Environments without Data Mining

We evaluate the local asymptotic power of the six predictability tests of section 3 by simulation. Let $l = 1$ and $k = 1$, corresponding to the standard specification used for example in testing the dividend-ratio model. Thus, S_1 may be interpreted as a two-sided t-test and S_2 as a one-sided t-test of the null hypothesis that w_t does not predict y_t. We postulate that $v_t = 1$ for all t, and $w_t = \rho w_{t-1} + \eta_t$. For simplicity, let $\eta_t \sim NID(0, \sigma_\eta^2)$. and $u_t \sim NID(0, \sigma^2)$. We set $\rho = 0.9$, $\sigma_\eta^2 = 0.005$ and $\sigma^2 = 0.05$. These values are close to values obtained in empirical research (see e.g., Mark 1995). Our qualitative conclusions are not sensitive to these parameter choices. Specifically, we evaluated a grid of parameter values including $\rho \in \{0.45, 0.9, 0.99\}$, $\sigma_\eta^2 \in \{0.005, 0.05\}$ and $\sigma^2 \in \{0.05, 0.5\}$. The results remained very similar.

We set $\pi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and $c \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$. The larger c, the larger the power of the test, all else equal. For $c = 5$ the power of the in-sample tests is close to 100%. We evaluate the local asymptotic power of all tests at the nominal 10% significance level. The asymptotic critical values are computed based on 20000 draws from the limit distributions in section 3 with $c = 0$. The rejection rates under the alternative are based on 20000 draws from the limit distribution with $c > 0$. All Brownian motions in turn are approximated based on discrete time approximations with T =5000.

Figure 1 shows the power results as a function of c and π. As expected, the two-sided test S_1 has lower power than the one-sided test S_2 for all c. Among the out-of-sample tests, there is no clear power ranking of the encompassing test and the GM test. Nor is there a clear ranking between the recursive test and the corresponding split-sample test. In Figure 2, the asymptotic power of the out-of-sample tests S_3, S_4, S_5, and S_6 is expressed in ratio form relative to the asymptotic power of the in-sample tests S_1 (upper panel) and S_2 (lower panel). Thus, power ratios below 1 indicate that the in-sample test is more powerful.

We find that for all values of π and c, the S_1-test has higher power than the GM tests as well as higher power than the forecast encompassing tests. This result holds for both split-sample and recursive versions of these tests. Qualitatively similar results also hold for S_2. For all values of π and c, the out-of-sample tests have lower power than the S_2-test. The relative power advantages of the S_2-test are even more pronounced than for the S_1-test. This result is not surprising because S_2 is a test of the null of no predictability against the one-sided alternative H_2. In contrast, the out-of-sample tests S_3, S_4, S_5, and S_6 test the weaker hypothesis of whether one model is a more accurate predictor than the other, which corresponds to H_1. Thus, their power is diluted. There are no out-of-sample tests, to our knowledge, that can incorporate one-sided hypotheses on β implied by economic theory. For example, standard exchange rate models imply that periods, in which the nominal exchange rate exceeds the equilibrium value of the exchange rate, should be followed by a *depreciation* of the exchange rate, as opposed to a *change* in the exchange rate. In other words, the economic model pins down the direction of change. Out-of-sample tests do not make use of that restriction.

If a researcher is interested in testing H_2 our results clearly suggest that the preferred testing strategy, at least asymptotically, would be to rely on the one-sided in-sample t-test (S_2). Even if we are interested in the two-sided alternative H_1, however, the in-sample F-test (S_1) would clearly be the preferred test. Thus, our analysis does not support the conventional wisdom that out-of-sample tests are more credible tests of the null of no predictability than in-sample tests.

Rather our local asymptotic results support the alternative explanation that the discrepancies between in-sample and out-of-sample test results tend to reflect the lower power of out-of-sample tests.

4.2 Environments with Data Mining

We now turn to environments in which more than one candidate model was considered prior to testing. The economically most interesting alternative hypothesis is one, in which one candidate model helps to predicts y_t in population, whereas all other candidate models do not. For expository purposes we let $M = 2$. Thus, $c = [c_1\ 0]$, where $c_1 \epsilon \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$. We postulate that the two potential predictors w_{jt} follow identical AR(1) processes that are independent. The details of the design are otherwise identical to the environment without data mining. In short, the experiment differs from the previous subsection in that we allow for the possible selection of an irrelevant predictor.

Figure 3 shows the power results as a function of c and π. We find that power tends to be somewhat lower, but qualitatively the power results are very similar. In Figure 4, the asymptotic power of the out-of-sample tests S_3, S_4, S_5, and S_6 is expressed in ratio form relative to the asymptotic power of the in-sample tests S_1 (upper panel) and S_2 (lower panel). As in the case without data mining, the one-sided t test always has higher power than the out-of-sample tests. The magnitude of the power advantage of the in-sample test can be substantial. In contrast, for the F-test the earlier results have to be qualified. Although the split-sample GM and encompassing tests always have lower power than the F-test, the asymptotic power ranking of the recursive GM-test and recursive forecast encompassing test relative to the F-test is ambiguous. It depends on the the values of c and π. For some values of c and π, the out-of-sample tests have higher power, whereas in others the two-sided in-sample test has higher power. No practical recommendations can be given in favor of one or the other test. With this exception, our results suggest that even if applied users were to use appropriate critical values to protect against data mining, in-sample tests would tend to have more power than out-of-sample tests. Notably, if a one-sided hypothesis on β is of interest, the power analysis unambigously favors the in-sample test even in the presence of data mining.

5 Discussion and Conclusion

It is common for empirical researchers to find significant evidence of in-sample predictability, but no significant out-of-sample predictive relationship. The conventional wisdom is that this tendency reflects the lack of reliability of in-sample tests under the null of no predictability. As a result, there is a tendency to discount evidence in favor of predictability based on in-sample tests. We showed that this interpretation is not correct. We distinguished between environments that are subject to data mining and environments that are free from data mining.

First, we demonstrated that in-sample and out-of-sample tests of predictability are asymptotically equally reliable under the null of no predictability, provided that no data mining has taken place. Second, we analyzed environments with data mining. We showed that, contrary to conventional wisdom, out-of-sample tests of predictability are not robust against data mining.

If critical values are not adjusted to account for data mining, both in-sample and out-of-sample tests are susceptible to size distortions. We also showed that, once proper critical values are used that are robust against data mining, both tests are equally reliable under the null. Thus, with or without data mining, the conventional wisdom that in-sample t-tests and F-tests are biased in favor of detecting spurious predictability cannot be supported by theory. This reduces the choice between in-sample and out-of-sample tests of predictability to the question of which test has higher power.

We derived the local asymptotic power of these tests in environments with and without data mining. Although in-sample tests will tend to have higher power than out-of-sample tests in small samples, it is not clear whether these results extend to large samples. Our analysis suggests that if inference is based on standard critical values, both one-sided t-tests and F-tests are even asymptotically more powerful than tests of equal predictive accuracy or tests of forecast encompassing. If inference is based on data-mining robust critical values, F-tests are asymptotically more powerful than split-sample tests, but for some design parameters they may be strictly less powerful than recursive out-of-sample tests. Thus, no general recommendations about the relative merits of F-tests and out-of-sample tests are possible. In contrast, one-sided t-tests are asymptotically more powerful than tests of equal predictive accuracy or tests of forecast encompassing, even after accounting for data mining.

Our results not only dispel the conventional wisdom that out-of-sample test results are more convincing than in-sample test results, but they also provide an alternative explanation for the tendency of significant in-sample test results to break down out of sample. Rather than attributing this result to higher size distortions for in-sample tests of predictability, we attribute this result to the higher power of in-sample tests of predictability relative to out-of-sample tests of the same size. Our results are particularly clear-cut when we compare the power of various out-of-sample tests of predictability to the commonly used one-sided in-sample t-test of predictability (see e.g. Mark 1995, Kilian 1999, Faust, Rogers and Wright 2003). The reason is that out-of-sample tests are not designed to test one-sided hypotheses on regression parameters, but amount to two-sided tests on regression parameters. This fact helps to explain the stronger in-sample evidence obtained in many empirical studies using such t-tests.

An alternative rationale for the apparently contradictory results of in-sample and out-of-sample tests of predictability is the presence of structural instability (see e.g. Stock and Watson 2001). Throughout this paper we maintained the implicit assumption that the process in question is not subject to structural change. Interestingly, recent work suggests that our basic results would not necessarily be altered by the presence of structural change. For example, Clark and McCracken (2001b) study the effects of deterministic structural breaks on the power of predictability tests. Their analysis abstracts from data mining. Like our paper they consider two parametric linear models that are nested under the null. The joint null hypothesis is no predictability and no structural change. Note that by definition a structural break in the predictive relationship of interest is not possible under the null. Thus, our results on the asymptotic size of in-sample and out-of-sample tests continue to apply. Both in-sample and out-of-sample tests will by construction be reliable under the null at least asymptotically. This point is important because it implies that deterministic structural breaks in the predictive relationship could not possibly be responsible for the alleged tendency of in-sample tests to result in spurious predictability.

In practice, the possible presence of structural breaks is important not because it affects the size of the test, but because it will tend to affect the power of predictability tests. For example, Clark and McCracken allow for a one-time structural shift in the predictive relationship under the alternative hypothesis. They postulate that for some part of the sample $\beta = 0$, whereas for the remainder of the sample $\beta \neq 0$. They show that the usual in-sample F-test will detect this form of predictability with probability 1 asymptotically. In this sense, in-sample tests of predictability are robust at least to simple forms of structural change. In contrast, out-of-sample tests such as tests of equal forecast accuracy or forecast encompassing may lack power against some structural break alternatives. Put differently, if there is predictability at least for some part of the sample, but is subject to structural change, out-of-sample tests may fail to detect it, even as in-sample tests correctly reject the no predictability null. Thus, out-of-sample tests of predictability may actually be less robust to structural change than in-sample tests.[5] Moreover, the power of out-of-sample tests will be highly dependent on the nature and timing of the structural change and on the choice of forecast window. This evidence is important because it is sometimes suggested that recursive or rolling estimates may offer some protection against model instability (see, e.g., Stock and Watson 1999, p. 298, 326). The example of Clark and McCracken (2001b) shows that this need not be the case.[6]

We conclude that the empirical evidence on in-sample and out-of-sample predictability tests needs to be reconsidered. Especially in the exchange rate literature there recently has been a shift in empirical work in favor of out-of-sample predictive inference (see, e.g., Frankel and Rose 1995, pp. 1702-1705). Our analysis suggests that there is no theoretical basis for this shift and that many economic forecast models may have been rejected incorrectly based on weak out-of-sample evidence.

There are two obvious caveats to our conclusions. First, the limit distributions of in-sample and of out-of-sample tests of predictability are in general data-dependent. This means that applied researchers will need to give careful attention to the size and power properties of these tests on a case-by-case basis. Second, our analysis has been asymptotic in nature. This is an advantage in that were able to derive rigorous results. The asymptotic nature of our results also is a disadvantage in that many applications involve fairly small samples and asymptotic approximations may not be accurate enough. Given the favorable experience with bootstrap methods in Kilian (1999) and Kilian and Taylor (2003), however, we conjecture that in practice bootstrap methods may be used to address both of these concerns.

[5] There is one counterexample to this tendency, in which out-of-sample tests will tend to have higher power than in-sample tests: Suppose that the break in β occurs at exactly $[\lambda T]$ where $\lambda = 0.5$. Further suppose that in the first half of the sample $\beta = -c$ and in the second half $\beta = c$ where c is some constant. In that case, the in-sample test will have zero power asymptotically, whereas the out-of-sample test may have some power. This counterexample, however, seems more of an intellectual curiosity because it requires three unrealistic conditions. First, a switch in sign seems unlikely in situations that would suggest the use of a one-sided t-test, as is typically the case in applied work. Second, it is unlikely that the deviations from $\beta = 0$ exactly offset one another. Third, it is unlikely that the break occurs exactly at $[0.5T]$. Even for small deviations from these assumptions the counterexample breaks down.

[6] In related work, Rossi (2001b) develops an optimal test of the joint null hypothesis of no predictability and no parameter instability. This test differs from both the in-sample and the out-of-sample tests analyzed in this paper. Rossi shows that her test is locally asymptotically more powerful than either rolling out-of-sample tests of predictability or sequential in-sample tests first for parameter stability and then for predictability.

Our results should not be interpreted as evidence against the use of out-of-sample tests of predictability in general, but against their uncritical use in applied work. We conclude that it is important to be clear about the objective of predictability testing. It will be difficult to find convincing applications of out-of-sample tests in standard environments, such as the setting considered in this paper. There are, however, real-time forecasting problems for which out-of-sample tests seem well suited. For example, Amato and Swanson (2001) and Chao, Corradi and Swanson (2001) draw a distinction between predictability that can be exploited in real time and predictability that exists in population. An interesting topic for future research would be the potential advantages of out-of-sample tests of predictability in models that are misspecified under the null hypothesis or in models that evolve smoothly over time.

Proofs

The proofs are similar to those in McCracken (1999) and Clark and McCracken (2000), so we will only sketch them.

Proofs for the split-sample tests: (8) and (12)

By Lemma A.4 of Clark and McCracken (2001a), Q is idempotent. By Schur's decomposition theorem (Theorem 13 of Magnus and Neudecker, 1999, p.16), it follows that there is a $l \times k$ matrix L such that $LL' = Q$ and $L'L = I_k$. Let $\tilde{h}_t = L'E(x_t x_t')^{-1/2} x_t (u_t + T^{-1/2} c' w_t)$. Then it follows from the functional central limit theorem that

$$T^{-1/2} \sum_{t=1}^{[rT]} h_t \Rightarrow \sigma\delta r + \sigma W(r), \tag{26}$$

where $[x]$ denotes the integer part of x and $\Rightarrow$ denotes the weak convergence in the space of cadlag functions on $[0, 1]$. Following the arguments used in the proofs of McCracken (1999, Lemma 3.2) and Clark and McCracken (2000, Lemmas A10 and A12), one can show that

$$\begin{aligned} S_3 &= \sum_{s=R+1}^{T} \tilde{h}_s' \frac{1}{R} \sum_{t=1}^{R} h_t + \frac{T-R}{2R^2} \sum_{s=1}^{R} h_s' \sum_{t=1}^{R} h_t + o_p(1), \\ S_5 &= (T-R)^{-1/2} (\sum_{s=1}^{R} h_s' \sum_{t=R+1}^{T} h_t) / (\sum_{s=1}^{R} h_s' \sum_{u=1}^{R} h_u)^{1/2} + o_p(1). \end{aligned} \tag{27}$$

Combining (26), (27) and (27) with the continuous mapping theorem completes the proof of (8) and (12).

Proofs for the recursive tests: (9) and (13))

It follows from applications of Theorem 2.1 of Hansen (1992), (26) and the continuous mapping theorem that

$$\sum_{t=[\pi T]+1}^{T} h_t' \frac{1}{t} \sum_{s=1}^{t} h_t \xrightarrow{d} \int_{\pi}^{1} \frac{1}{r} W'(r) dW(r) + \delta' \int_{\pi}^{1} \frac{1}{r} W(r) dr + \delta'(W(1) - W(\pi)) + (1-\pi)\delta'\delta. \tag{28}$$

Following the arguments used in the proofs of McCracken (1999, Lemma 3.2) and Clark and McCracken (2000, Lemmas A10 and A12), one can show that

$$\begin{aligned} S_4 &= \sum_{t=R+1}^{T} h_t' \frac{1}{t} \sum_{s=1}^{t} h_s + \frac{1}{2} \sum_{t=R+1}^{T} \frac{1}{t^2} \sum_{s=1}^{t} h_s' \sum_{u=1}^{t} h_u + o_p(1), \\ S_6 &= (\sum_{t=R+1}^{T} h_t' \frac{1}{t} \sum_{s=1}^{t} h_s) / (\sum_{t=R+1}^{T} \frac{1}{t^2} \sum_{s=1}^{t} h_s' \sum_{u=1}^{t} h_u)^{1/2} + o_p(1). \end{aligned} \tag{29}$$

The desired results (9) and (13)) follow from (26), (28), (29) and (29).

Proof of Propositions 4 and 5. Under Assumptions 1 and 3 and the sequence of local alternatives, we have

$$\left[\frac{\sum_{t=1}^{T}\hat{u}_{1,t}^2-\hat{u}_{0,t}^2}{\hat{\sigma}_1^2},...,\frac{\sum_{t=1}^{T}\hat{u}_{M,t}^2-\hat{u}_{0,t}^2}{\hat{\sigma}_M^2}\right]' \overset{d}{\to} [\chi_{d_1}^2(k_1),...,\chi_{d_M}(k_M)]' \tag{30}$$

Under Assumptions 1 and 4 and the sequence of local alternatives,

$$\left[\sqrt{T}\hat{\beta}_{1,T}/\hat{\sigma}_{\beta_1},...,\sqrt{T}\hat{\beta}_{M,T}/\hat{\sigma}_{\beta_M}\right]' \overset{d}{\to} [v_1,...,v_M]' \tag{31}$$

Applications of the continuous mapping theorem to (30) and (31) complete the proof of Propositions 4 and 5, respectively.

Proof of Proposition 6:

Let

$$h_t = L_j'[E(x_{j,t}x_{j,t}')]^{-1/2}x_{j,t}(u_t+T^{-1/2}c(\zeta_j)'w_{j,t})$$

The proof of Proposition 6 is analogous to those of Propositions 2 and 3 except that (26) is replaced with

$$T^{-1/2}\sum_{t=1}^{rT}[\begin{array}{cccc} h_{1,t} & h_{1,t} & \cdots & h_{M,t}\end{array}]' \Rightarrow \sigma dr+\Omega^{1/2}B(r) \tag{32}$$

and that the continuous mapping theorem is applied to the max functional.

References

1. Amatao, J.D., and N.R. Swanson (2001), "The Real-Time Predictive Content of Money for Output," *Journal of Monetary Economics,* 48, 3–24.
2. Anderson, T.W. (1994), *The Statistical Analysis of Time Series*, Wiley: New York.
3. Ashley, R., C.W.J. Granger and R. Schmalensee (1980), "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica*, 48, 1149–1167.
4. Berben, R.-P., and D. van Dijk (1998), "Does the Absence of Cointegration Explain the Typical Findings in Long-Horizon Regressions?," manuscript, Tinbergen Institute, Erasmus University Rotterdam.
5. Berkowitz, J., and L. Giorgianni (2001), "Long-Horizon Exchange Rate Predictability?," *Review of Economics and Statistics,* 83, 81–91.
6. Campbell, J.Y., and R.J. Shiller (1988a), "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies,* 1, 195–228.
7. Campbell, J.Y., and R.J. Shiller (1988b), "Stock Prices, Earnings, and Expected Dividends," *Journal of Finance,* 43, 661–676.
8. Chao J.C., V. Corradi and N.R. Swanson (2001), "An Out-of-Sample test for Granger Causality", *Macroeconomic Dynamics*, 5, 598-620.
9. Chinn, M.D., and R.A. Meese (1995), "Banking on Currency Forecasts: How Predictable is Change in Money?," *Journal of International Economics*, 38, 161–178.
10. Clarida, R.H., and M.P. Taylor (1997), "The Term Structure of Forward Exchange Premiums and the Forecastability of Spot Exchange Rates: Correcting the Errors," *Review of Economics and Statistics,* 89, 353–361.
11. Clarida, R.H., L. Sarno, M.P. Taylor and G. Valente (2003), "The Out-of_Sample Success of Term Structure Models as Exchange Rate Predictors: A Step Beyond," forthcoming: *Journal of International Economics.*
12. Clark, T.E. (2000), "Can Out-of-Sample Forecast Comparisons Help Prevent Overfitting?" manuscript, Research Division, Federal Reserve Bank of Kansas City.
13. Clark, T.E., and M.W. McCracken (2000), "Not-for-Publication Appendix to "Tests of Equal Forecast Accuracy and Encompassing for Nested Models"," manuscript, Federal Reserve Bank of Kansas City and Louisiana State University.
14. Clark, T.E., and M.W. McCracken (2001a), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics,* 105, 85–110.
15. Clark, T.E., and M.W. McCracken (2001b), "Forecast-Based Model Selection in the Presence of Structural Breaks," manuscript, Department of Economics, University of Missouri at Columbia.
16. Corradi V., N.R. Swanson and C. Olivetti (2001), "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 104, 315-358.

17. Dasgupta, N., and J.D. Spurrier (1997), "A Class of Multivariate Chi-Square Distributions with Applications to Comparisons with a Control," *Communications in Statistics: Theory and Method*, 26, 1559-1573.

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

19. Fama, E.F., and K.R. French (1988), "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics,* 22, 3-25.

20. Faust, J., J. Rogers and J. Wright (2003), "Exchange Rate Forecasting: The Errors We've Really Made," forthcoming: *Journal of International Economics.*

21. Frankel, J.A., and A.K. Rose (1995), "Empirical Research on Nominal Exchange Rates," in G. Grossman and K. Rogoff (eds.), *Handbook of International Economics, vol. III,* 1689-1729.

22. Gilbert, S. (2001), "Sampling Schemes and Tests of Regression Models," manuscript, Department of Economics, Southern Illinois University at Carbondale.

23. Ghysels, E., and A.R. Hall (1990), "A Test for Structural Stability of Euler Conditions Parameters Estimated via the Generalized Method of Moments Estimator," *International Economic Review,* 31, 355–364.

24. Granger, C.W.J. (1990), *Modeling Economic Time Series: Readings in Econometric Methodology*, Oxford University Press: Oxford, UK.

25. Goetzmann, W.N., and P. Jorion (1993), "Testing the Predictive Power of Dividend Yields," *Journal of Finance,* 48, 663–679.

26. Goetzmann, W.N., and P. Jorion (1995), "A Longer Look at Dividend Yields," *Journal of Business,* 68, 483–508.

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

28. Hansen, P.R. (2000), "The Distribution of the Maximal R^2," manuscript, Department of Economics, University of California at San Diego.

29. Hansen, P.R. (2001), "The Reality Check for Data Snooping: A Comment on White," manuscript, Department of Economics, Brown University.

30. Harvey, D.I., S.J. Leybourne and P. Newbold (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting,* 13, 281–291.

31. Harvey, D.I., S.J. Leybourne and P. Newbold (1998), "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics,* 16, 254–259.

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

33. Kilian, L. (1999), "Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?," *Journal of Applied Econometrics,* 14, 491–510.

34. Kilian, L., and M.P. Taylor (2003), "Why is it so Difficult to Beat the Random Walk Forecast of Exchange Rates?", forthcoming: *Journal of International Economics.*

35. McCracken, M.W. (1999), "Asymptotics for Out of Sample Tests of Causality," manuscript, Department of Economics, Louisiana State University.

36. McCracken, M.W. (2001), "Data Mining and Out-of-Sample Inference," manuscript, Department of Economics, Louisiana State University.

37. Mark, N.C. (1995), "Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability," *American Economic Review*, 85, 201–218.

38. 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.

39. 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.

40. Meese, R.A., and A.K. Rose (1991), "An Empirical Assessment of Non-Linearities in Models of Exchange Rate Determination," *Review of Economic Studies*, 58, 603–619.

41. Rossi, B. (2001a), "Testing Out-of-Sample Predictive Ability with High Persistence. An Application to Models of Exchange Rate Determination," manuscript, Department of Economics, Princeton University.

42. Rossi, B. (2001b), "Optimal Tests for Nested Model Selection with Underlying Parameter Instability," manuscript, Department of Economics, Duke University.

43. Royen, T. (1984), "Multivariate Comparisons of Polynomial Distributions," *Biometrical Journal*, 26, 319-332.

44. Royen, T (1997), "Multivariate Gamma Distributions (Update)" in S. Kotz (ed.), *Encyclopedia of Statistical Sciences,* Update Vol. 1, Wiley: New York.

45. Stock, J.H. (1994), "Unit Roots, Structural Breaks and Trends," in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics,* Vol. 4, Elsevier, Amsterdam, 2739–2841.

46. Stock, J.H., and M.W. Watson (1999), "Forecasting Inflation," *Journal of Monetary Economics,* 44, 293–335.

47. Stock, J.H., and M.W. Watson (2001), "Forecasting Output and Inflation: The Role of Asset Prices," manuscript, Kennedy School of Government, Harvard University.

48. Sullivan, R., A. Timmermann and H. White (2001), "Dangers of Data Mining: The Case of Calendar Effects in Stock Returns," *Journal of Econometrics,* 105, 249–286.

49. West, K.D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1084.

50. West, K.D. (2001), "Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics*, 19, 29–33.

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

Figure 1: **Power of Predictability Tests in Environments without Data Mining**

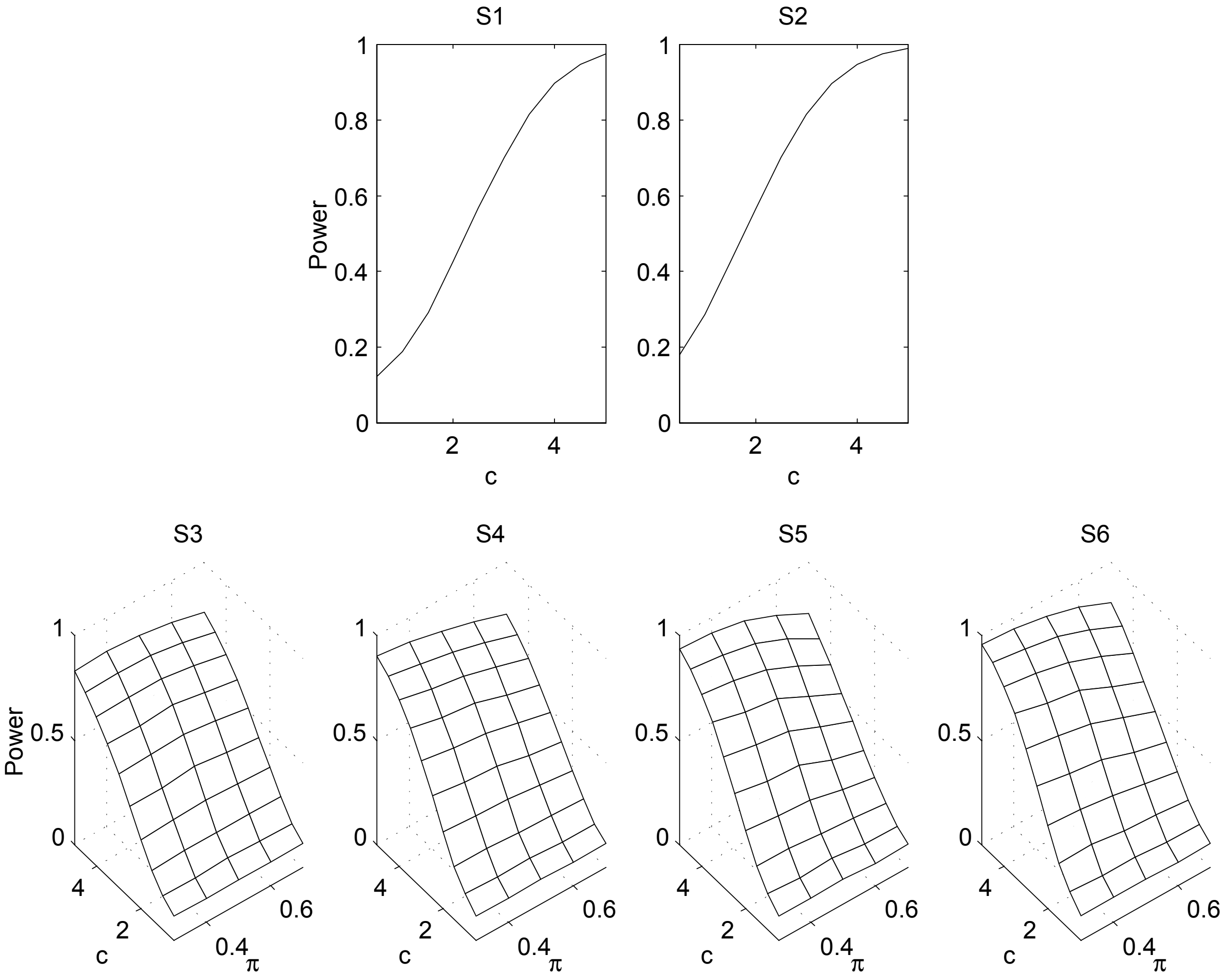

Figure 2: **Relative Power of Predictability Tests in Environments without Data Mining**

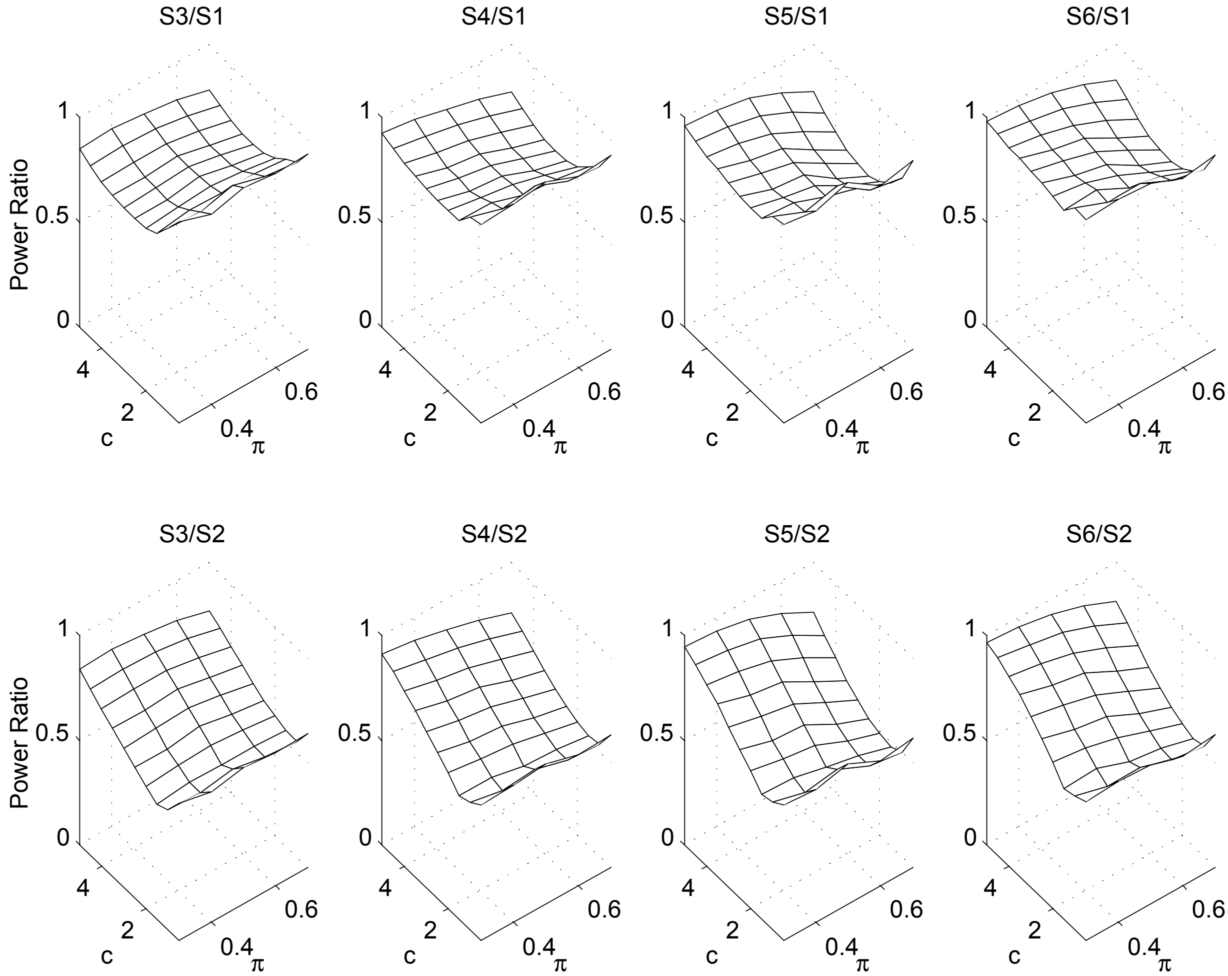

Figure 3: **Power of Predictability Tests in Environments with Data Mining**

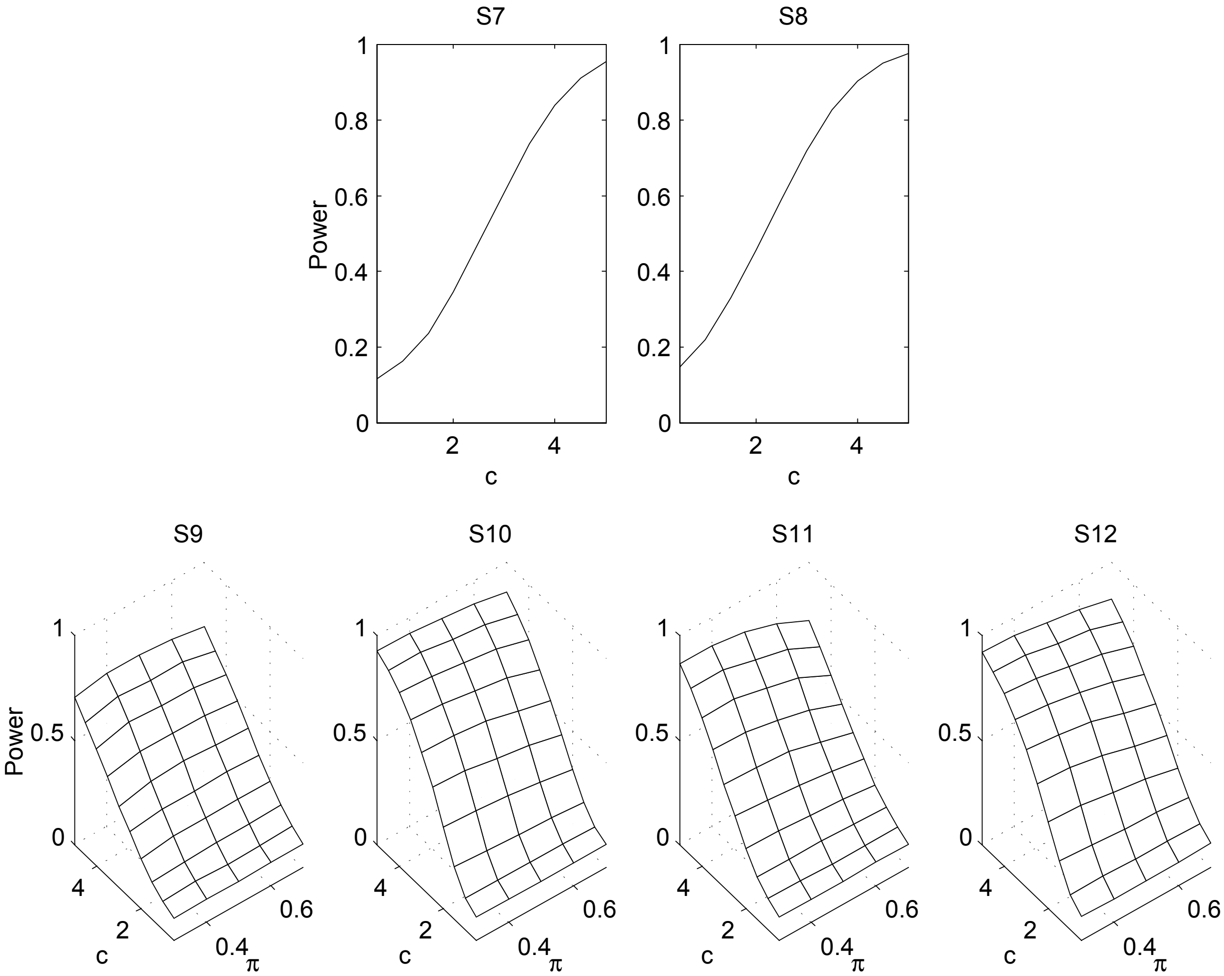

Figure 4: **Relative Power of Predictability Tests in Environments with Data Mining**