An asymptotically normal out-of-sample test of equal predictive accuracy for nested models

Gray Calhoun*
Iowa State University

March 14, 2013

Abstract

This paper proposes a modification of Clark and West's (2007, *J. Econom.*) adjusted out-of-sample t-test. The alternative model is still estimated with a fixed-length rolling window, but the benchmark is estimated with a recursive window. The resulting statistic is asymptotically normal even when the models are nested. Moreover, the alternative model can be estimated using common model selection methods, such as the AIC or BIC. This paper also presents a method to compare multiple models simultaneously while controlling familywise error, and substantially improves existing block bootstrap methods for out-of-sample statistics. This procedure is then used to analyze Goyal and Welch's (2008, *Rev. Finan. Stud.*) excess returns dataset.

Keywords: Forecast Evaluation, Martingale Difference Sequence, Model Selection, Family-Wise Error Rate; Multiple Testing; Bootstrap; Reality Check

JEL Classification Numbers: C22, C53

*Economics Department; Iowa State University; Ames, IA 50011. Telephone: (515) 294-6271. Email: gcalhoun@iastate.edu. Web: http://gray.clhn.co. I'd like to thank Helle Bunzel, Todd Clark, Graham Elliott, Yu-Chin Hsu, Michael McCracken, Pablo Pincheira, Allan Timmermann, Stephane Meng-Feng Yen and participants at the 2011 Midwest Econometrics Group meeting for helpful comments and discussions. I'd also like to thank Amit Goyal for providing computer code and data for his 2008 RFS paper with Ivo Welch (Goyal and Welch, 2008).

1 Introduction

This paper proposes an out-of-sample (OOS) test statistic that is asymptotically normal with mean zero even when the models studied are nested. This paper also proves that common block bootstrap methods consistently estimate the statistic's distribution, and shows how the statistic can be used to study many models simultaneously while maintaining strong control of the familywise error (FWE), i.e. ensuring that the probability that the hypothesis that the model has equal predictive ability to a benchmark is incorrectly rejected for any of the models is no higher than a preset level.

OOS tests are commonly used in International Macroeconomics, Macroeconomics, and Finance (see, for example, Meese and Rogoff 1983; Stock and Watson 2003; and Goyal and Welch 2008) and there is a substantial literature developing the theoretical properties of these statistics, beginning primarily with Diebold and Mariano (1995) and West (1996).[1] In a pair of papers, Clark and West (2006, 2007) develop an OOS test of the null hypothesis that a small benchmark model is correctly specified. Their test compares the forecasting performance of a pair of nested models, and the null hypothesis is that the innovations in the smaller model form a martingale difference sequence. This test procedure is popular, and one assumes that this is due in part to the statistic's convenience, the statistic is approximately normal after adjusting for the estimation error of the larger model. Normality comes from a fixed-length rolling window, as in Giacomini and White (2006), and the adjustment centers the statistic appropriately. This statistic is especially convenient because other OOS tests for similar hypotheses (Chao et al. 2001; Clark and McCracken 2001, 2005a; Corradi and Swanson 2002, 2004; and McCracken 2007; among others) have a nonstandard limit distribution and place restrictions on the models under consideration, while other asymptotically normal statistics test a different null hypothesis (Giacomini and White, 2006) or place assumptions on the models and DGP that are often violated in empirical work (Diebold and Mariano 1995; West 1996; West and McCracken 1998; McCracken 2000).[2]

However, Clark and West's statistic is only "approximately normal" in an informal sense. Clark and West present Monte Carlo evidence of the statistic's distribution, but only prove that the statistic is asymptotically normal with mean zero when the benchmark model is

[1] Other papers in this literature include West and McCracken (1998), McCracken (1998, 2000), Clark and McCracken (2001, 2005a,b, 2011, 2012a,b), Chao et al. (2001), Corradi and Swanson (2002, 2004, 2007), White (2000), Inoue and Kilian (2004, 2006), Hansen (2005), Rossi (2005), Clark and West (2006, 2007), Anatolyev (2007), Giacomini and Rossi (2009, 2010), Hubrich and West (2010), Hansen et al. (2011), Inoue and Rossi (2011), Pincheira (2011), Rossi and Sekhposyan (2011a,b), and Calhoun (2011), among others. For recent reviews of this literature and additional references, see McCracken and West (2002), Corradi and Swanson (2006), West (2006), Clark and McCracken (2002), Corradi and Distaso (2011), and Giacomini (2011)

[2] Diebold and Mariano (1995) assume that the models are not estimated. West (1996), West and McCracken (1998), and McCracken (2000) implicitly assume that the models do not converge to the same limit, which rules out nesting.

a random walk (Clark and West, 2006). Estimating the parameters of the smaller model invalidates their proof.

In this paper, I show that a modified version of their statistic is asymptotically normal even when the smaller model is estimated. To achieve normality, we need a consistent estimate of the pseudo-true benchmark model, while maintaining an inconsistent estimate of the larger model so that we can ignore nesting. We can meet both needs by using different window strategies for each model: the benchmark model is estimated using a recursive window and the alternative with a fixed-length rolling window.

Mixing window strategies is uncommon but needn't be. In most applications, the null hypothesis imposes stability as well as equal accuracy between the two models. The benchmark model rarely allows for breaks, parameter drift, or other forms of instability,[3] but the researcher is typically concerned about instability. Indeed, concern about instability is often given as a reason for doing an OOS analysis, especially with a short rolling window.[4] A researcher could impose stability on both models by using a recursive window or relax stability for both by using a rolling window; either approach should not affect the test's size, but may affect power. But the researcher could instead impose stability on the benchmark and relax it for the alternative by using a recursive window for the benchmark and a rolling window for the alternative model. This approach could have a power advantage and is similar in spirit to using a Likelihood Ratio Test instead of an LM or Wald test, which depend on just the restricted or unrestricted model respectively.

This paper's statistic has a substantial advantage over existing OOS tests for nested models: the alternative can be essentially arbitrary as long as high level moment conditions hold. In particular, researchers can use model selection techniques like the AIC or BIC to determine the number of lags to include, the particular exogenous variables to include, etc. Other methods that test a similar hypothesis are unable to handle these models (except Clark and West, 2006, which does not allow the benchmark to be estimated); Giacomini and White (2006) are able to handle such models for both the alternative and the benchmark but, as mentioned earlier, they test a different aspect of forecasting performance.

This paper focuses on nested models, as they have received the most attention in the empirical and theoretical literature, but the statistic can be used with non-nested models as well. This generality is useful, since West's (1996) results do not apply to non-nested models if they both encompass the true DGP,[5] which is allowable under the null: in the limit, both models will converge to the DGP and give identical forecasts. Consequently, the

[3]Exceptions are Stock and Watson's (2007) IMA(1,1) and UC-SV models of inflation.

[4]This motivation is discussed by Stock and Watson (2003), Pesaran and Timmermann (2005, 2007), Giacomini and White (2006), Goyal and Welch (2008), Clark and McCracken (2009), and Giacomini and Rossi (2009, 2010), among others.

[5]Clark and McCracken (2011) call this scenario, "overlapping models."

naive OOS t-test is invalid, even after correcting the standard error if necessary to reflect parameter uncertainty. Clark and McCracken (2011) show that the fixed window OOS t-test remains normal for these models but the recursive and rolling windows (with the window size increasing to ∞) do not, and provide a procedure for pointwise (but not uniformly) valid tests for the recursive and rolling windows and uniformly valid tests for the fixed window. The test proposed in this paper is uniformly valid and places fewer assumptions on the models under study and the true DGP.

Since researchers often have a set of potential models and want to know which of them significantly outperform the benchmark, procedures that compare a single pair of models are of limited practical value. As White (2000) demonstrates, looking at the naive p-values of individual tests is misleading, but researchers can obtain a valid test by using the bootstrap to approximate the distribution of the largest individual test statistic under the null (White calls this procedure the *Bootstrap Reality Check* or BRC). This paper presents a test for equal predictive ability of multiple models based on Romano and Wolf's (2005) StepM, which uses a step-down procedure that iteratively rejects models to achieve higher power than the BRC and indicate which of the models improves on the benchmark (see Theorem 2 for details).

This paper also presents a new result for the validity of block bootstraps with OOS statistics, which is necessary to verify that the StepM is valid. Existing results on bootstrapping OOS statistics have some drawbacks. White (2000) and Hansen (2005) use the stationary bootstrap, but require the test sample to be much smaller than the training sample, which obviously does not hold here. Corradi and Swanson (2007) relax that requirement, but make the statistic more complicated than necessary by adjusting the objective function of the bootstrapped statistic to center it correctly.[6] A parametric bootstrap, such as that used by Mark (1995), Kilian (1999), and Clark and McCracken (2012b), is an alternative method, but requires the benchmark model to be correctly specified. Although I focus on the null hypothesis that the benchmark is correctly specified, this paper's block bootstrap methods remain valid when the benchmark is misspecified and are relatively easy to implement. So this paper's result for bootstrapping OOS statistics considerably improves on existing procedures.

The next section presents our new statistics and Section 3 our block bootstrap result. Section 4 presents simulations that compare our pairwise OOS test to Clark and West's (2006, 2007) original statistics. Section 5 demonstrates the use of our statistic by reanalyzing Goyal and Welch's (2008) study of excess return predictability. Section 6 concludes.

[6] Their recentering is required for consistency and does not imply higher order accuracy.

2 Out-of-Sample Model Comparisons

This section presents the new OOS statistic. I will present the single-comparison statistic first and then extend it to multiple comparisons. Suppose for now that a researcher is interested in predicting the target variable y_{t+1} with a vector of regressors x_t; also let v_t be another random process and suppose that (y_t, x_t, v_t) is stationary and weakly dependent (Theorem 1 lists the assumptions formally). In addition, let $\beta_0 = (\mathrm{E}\, x_t x_t')^{-1} \mathrm{E}\, x_t y_{t+1}$ be the pseudo-true coefficient of the regression of y_{t+1} on x_t and define $\varepsilon_{t+1} = y_{t+1} - x_t'\beta_0$. If the linear model is correctly specified, so ε_{t+1} is a martingale difference sequence with respect to $\mathcal{F}_t \equiv \sigma((x_t, v_t, y_t), (x_{t-1}, v_{t-1}, y_{t-1}), \dots)$, then we can see immediately that

$$\tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} \varepsilon_{t+1}(v_t - x_t'\beta_0) \tag{1}$$

obeys an MDS CLT (assuming that its variance is uniformly positive) and is asymptotically normal as $P \to \infty$, with R an arbitrary starting value[7] and $P = T - R$. Straightforward algebra (Clark and West, 2007) shows that

$$\tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} \varepsilon_{t+1}(v_t - x_t'\beta_0) = \tfrac{1}{2\sqrt{P}} \sum_{t=R+1}^{T} \Big[(y_{t+1} - x_t\beta_0)^2 - (y_{t+1} - v_t)^2 + (x_t'\beta_0 - v_t)^2\Big]. \tag{2}$$

Clark and West (2006, 2007) base their OOS statistic on the right side of (2), where R is the size of the training sample and P the size of the test sample used to evaluate the forecasting models. Clark and West (2006) use a second forecast $\hat{y}_{t+1}$ as v_t. They use a rolling window of length R (i.e., $\hat{y}_{t+1}$ is a function of $y_t, x_{t-1}, z_{t-1} \dots, y_{t-R+1}, x_{t-R}$ and z_{t-R} where z_t is another weakly dependent random process), which is kept finite as $T \to \infty$, so $\hat{y}_{t+1}$ inherits the weak dependence properties of the variables used to estimate it. Keeping R finite ensures that the conditional variance remains positive, so the sum obeys a CLT. This method of ensuring normality was introduced by Giacomini and White (2006). In Clark and West (2006), the coefficients β_0 are assumed to be zero so ε_{t+1} is observed directly (under the null hypothesis) and Clark and West propose using this statistic to test the null that ε_{t+1} is an MDS with respect to $\mathcal{F}_t$. In Clark and West (2007), β_0 is unknown but is estimated with a fixed-length rolling window as well, so

$$\tilde{\beta}_t = \Big(\sum_{s=t-R+1}^{t} x_{s-1} x_{s-1}' \Big)^{-1} \sum_{s=t-R+1}^{t} x_{s-1} y_s$$

[7] It will be clear momentarily why the summation begins at $R+1$ instead of 1.

and $\hat{\varepsilon}_{t+1} = y_{t+1} - x_t'\tilde{\beta}_t$ replaces ε_{t+1} in the test statistic for MDS. Unfortunately, $\hat{\varepsilon}_{t+1}$ is not an MDS even when ε_{t+1} is, so the statistic is no longer asymptotically mean-zero normal, even though this approximation performs well in the simulations reported by Clark and West (2007).

This paper proposes using the same basic OOS statistic, but using a recursive window to estimate β_0 and produce $\hat{\varepsilon}_{t+1}$; i.e.

$$\hat{\beta}_t = \Big(\sum_{s=2}^{t} x_{t-1}x_{t-1}'\Big)^{-1} \sum_{s=2}^{t} x_{t-1}y_t. \tag{3}$$

West's (1996) Theorem 4.1 implies that

$$\tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} \Big[(y_{t+1} - x_t\hat{\beta}_t)^2 - (y_{t+1} - v_t)^2 + (x_t'\hat{\beta}_t - v_t)^2\Big]$$

is asymptotically normal with mean zero under the null for fairly arbitrary processes v_t, as long as v_t is weakly dependent and the OOS statistic has uniformly positive variance. Just as in Clark and West (2006), these conditions are ensured if v_t is another forecast of y_{t+1} based on a fixed length rolling window. Theorem 1 presents the details of this result. The assumptions required are essentially the same in existing papers (e.g. West, 1996; West and McCracken, 1998; McCracken, 2000; Giacomini and White, 2006; and Clark and West, 2006, 2007). Please see the original papers for a discussion of these assumptions.

Theorem 1. *Suppose that we have two models $\hat{y}_{0t}$ and $\hat{y}_{1t}$ to forecast the variable y_t, and have observations for $t = 1, \ldots, T+1$. Assume the following hold:*

1. *The benchmark forecast $\hat{y}_{0,t+1}$, is estimated using OLS with a recursive window: $\hat{y}_{0,t+1} = x_t'\hat{\beta}_t$ for some vector of predictors x_t with $\hat{\beta}_t$ given by Equation (3). Also define $\beta_0 = (\mathrm{E}\, x_{t-1}x_{t-1}')^{-1}\,\mathrm{E}\, x_{t-1}y_t$ and assume that β_0 does not depend on t.*

2. *The alternative forecast, $\hat{y}_{1t}$, is estimated using a rolling window of fixed length R (which is less than T), so $\hat{y}_{1,t+1} = \psi(y_t, z_t, \ldots, y_{t-R+1}, z_{t-R+1})$ where ψ is a known function and z_t is a sequence of predictors that may include x_t.*

3. *The series y_t, $\hat{y}_{1t}$, and x_t have uniformly bounded $2r$ moments for some $r > 2$, and $\hat{y}_t y_t$ $\hat{y}_t x_{t-1}$, $y_t x_{t-1}$, and x_{t-1}, x_{t-1}' are L_2-NED of size $-\frac{1}{2}$ on a strong mixing series of size $-\frac{r}{r-2}$ for $r > 2$ or a uniform mixing series of size $-\frac{r}{2r-2}$.*

4. *Define*

$$f_t(\beta) = (y_{t+1} - x_t'\beta)^2 - (y_{t+1} - \hat{y}_{1,t+1})^2 + (x_t'\beta - \hat{y}_{1,t+1})^2,$$

$f_t = f_t(\beta_0)$, $\hat{f}_t = f_t(\hat{\beta}_t)$, $\bar{f}(\beta) = \frac{1}{P}\sum_{t=R+1}^{T} f_t(\beta)$, *and* $\bar{f} = \frac{1}{P}\sum_{t=R+1}^{T} \hat{f}_t$, *with* $P = T - R$; $\bar{f}(\beta_0)$ *has uniformly positive and finite long run variance.*

Under the null hypothesis that $y_t - \hat{y}_{0t}(\beta_0)$ *is a martingale difference sequence with respect to the filtration* $\mathcal{F}_t = \sigma((y_t, z_t), (y_{t-1}, z_{t-1}), \dots)$, $\frac{\sqrt{P}}{\hat{\sigma}}\bar{f} \to^d N(0,1)$, *where*

$$\hat{\sigma}^2 = \hat{S}_{ff} + 2\Pi(\hat{S}_{fg} + \hat{S}_{gg}), \qquad \hat{S}_{ff} = \tfrac{1}{P}\sum_{t=R+1}^{T}(\hat{f}_t - \bar{f})^2,$$

$$\hat{S}_{fg} = \tfrac{1}{P}\sum_{t=R+1}^{T}(\hat{f}_t - \bar{f})(\hat{g}_t - \bar{g})', \qquad \hat{S}_{gg} = \tfrac{1}{P}\sum_{t=R+1}^{T}(\hat{g}_t - \bar{g})(\hat{g}_t - \bar{g})',$$

$\Pi = 1 - \frac{R}{P}\log(1 + \frac{P}{R})$, $\bar{g} = \frac{1}{P}\sum_{t=R+1}^{T}\hat{g}_t$, $\mathrm{X}' = [x_1, \dots, x_T]$,

$$g_t(\beta) = \Big\{\tfrac{2}{P}\sum_{s=R}^{T} x_s(x_s'\beta - \hat{y}_{1,s+1})\Big\}'\big(\tfrac{1}{T}\mathrm{X}'\mathrm{X}\big)^{-1}x_t(y_{t+1} - x_t'\beta),$$

and $\hat{g}_t = g_t(\hat{\beta}_t)$.

The following remarks are relevant to Theorem 1:

Remark 1. Forecasters are usually interested in the one-sided alternative that $\mathrm{E}\, f_t > 0$; i.e. that the alternative model is expected to forecast better than the benchmark.

Remark 2. These results are presented for one-period-ahead forecasting for simplicity. They can be extended to forecasting at a longer horizon by appropriately modifying the variance-covariance matrix to account for the correlation structure of the forecast errors. (i.e. for τ-step-ahead forecasts the errors will be an MA($\tau - 1$) process).

Remark 3. The requirement that the asymptotic variance of $\bar{f}(\beta_0)$ is uniformly positive is much less restrictive than in West (1996). As in Giacomini and White (2006) and Clark and West (2006, 2007), the assumption only serves to rule out pathological cases—for example, letting both the benchmark and the alternative model be perfectly correlated white noise. In West (1996), this assumption is a restriction on the DGP as well as the forecasting models, but in this paper it is a restriction only on the models.

Remark 4. As in West (1996), $\Pi \to 1$ as $T \to \infty$ since R is fixed. But using West's general formula for Π (which holds when $\lim \frac{P}{R} \in [0, \infty]$) can improve the statistic in practice, since researchers often invoke "fixed R" asymptotics when R is large. Section 4 shows via Monte Carlo that this approximation can be accurate even when R and P are equal.

Remark 5. The statistic we present tests the null hypothesis that the forecast errors from the population version of the benchmark model are a martingale difference sequence. This

hypothesis may not be appropriate, depending on the loss function or utility function of interest. If one wants to test the less restrictive null hypothesis that $y_t - \hat{y}_{0t}$ is uncorrelated with $\hat{y}_{1t}$ but not necessarily an MDS, one can replace $\hat{S}_{ff}$, $\hat{S}_{fg}$ and $\hat{S}_{gg}$ with their HAC counterparts. Our statistic can also be modified to test implications of optimal forecasts under other loss functions (see Patton and Timmermann, 2007a,b); the statistic should be expressed as a forecast encompassing test using the models' generalized forecast errors.[8] Again, Lemma B.4 can cover these other applications.

Remark 6. Giacomini and White (2006) and Clark and West (2006, 2007) emphasize model comparison using a rolling window but claim in passing that their results hold for a fixed-length fixed window as well, where the unknown coefficients are estimated only once using the first R observations. This claim is true but not obvious. The proof for the rolling window is based on weak dependence that arises because the forecast errors are a function of a finite number of consecutive weakly dependent observations. But using a fixed window introduces another source of dependence: all of the forecasts depend on the same coefficient estimates and the estimation uncertainty introduced by those estimates does not vanish asymptotically when R is bounded. One can prove that their results apply to fixed window forecasts by using a coupling argument similar to Calhoun's (2011). Interested readers should refer to Calhoun (2011) for the argument and to Merlevède and Peligrad for a review of the necessary coupling results. Theorem 1 can be shown to hold when $\hat{y}_{1t}$ is estimated with a fixed-length fixed window using the same arguments.

Since most empirical papers study more than one alternative model, Theorem 2 is of limited use on its own.[9] It can, however, be used as the basis for a procedure that allows researchers to make multiple comparisons. Before presenting a theoretical result, I will discuss some issues related to multiple hypothesis testing.

Suppose we have the family of hypotheses, $H_1, \ldots, H_J$. It is clear that testing each hypothesis individually will often have a probability greater than the tests' nominal size of rejecting at least once.[10] Econometricians, e.g. White (2000), Hansen (2005), Hubrich and West (2010), and Clark and McCracken (2012b), have emphasized statistics that test families of hypotheses that control the probability any hypothesis is rejected given that they all are true, known as *weak control of familywise error* (WFWE).[11] This paper focuses instead on controlling the probability that at least one true hypothesis is rejected given any combination

[8] See Harvey et al. (1998) and Clark and West (2007, Section 4).

[9] Moreover, we would need to account for the existence of multiple models even if each paper only considered a single model, since there are many papers studying the same data.

[10] Unless the statistics used for each test are completely interdependent, the probability will be strictly greater than the nominal size.

[11] An exception is Hsu et al. (2010) who combine Romano and Wolf's (2005) StepM with Hansen's (2005) threshold rule to control SFWE for certain families of one-sided tests.

of true and false hypotheses, known as *strong control of familywise error* (SFWE). For most empirical work, SFWE is desired; fortunately, even though these papers only prove WFWE, statistics based on nonparametric block bootstraps (Hansen, 2005, White, 2000) can be shown to control SFWE as well (this follows directly from Romano and Wolf, 2005) and those based on the parametric bootstrap (Clark and McCracken, 2012b) essentially control SFWE if the individual null hypotheses are strengthened slightly as below (see conclusion 1 of Theorem 2 as well as Remark 9). Hubrich and West (2010) propose that researchers construct critical values from the asymptotic joint distribution of the test statistics and this procedure can control SFWE as long as there is a consistent estimator of that distribution.

Only tests with SFWE can tell researchers which of the hypotheses are false. To see why, imagine that there are only two hypotheses: H_1 is true and H_2 is false. A test that rejects the true hypothesis, H_1, with arbitrarily high probability still can satisfy WFWE: since not all of the hypotheses are true, the behavior of tests with weak control is essentially unconstrained. Such a test would not satisfy SFWE, though. For SFWE in this setting, a test must reject H_1 with probability at most α (letting α be the desired level of control). In most research, we would view rejection of the true hypothesis H_1 as a mistake.

This paper uses Romano and Wolf's (2005) StepM procedure, which is an iterated extension of White's (2000) Reality Check designed to reject as many false hypotheses as possible while achieving SFWE. Their method relies on the bootstrap to estimate the joint dependence between the test statistic associated with each model. The bootstrap used in this paper is new, and is outlined in Algorithm 1. Algorithm 1 actually presents a more general version of the bootstrap that remains valid with a misspecified benchmark model. Imposing the null hypothesis of MDS leads to simpler version that will be discussed later.

Algorithm 1. *Suppose that there are m alternative forecasting models, $\hat{y}_{1t}, \ldots, \hat{y}_{mt}$, and define the data matrices $\mathrm{X_R} = (X_1, \ldots, X_R)'$ and $\mathrm{X_P} = (X_{R+1}, \ldots, X_{T-1})$ with*

$$X_t = \begin{cases} (y_{t+1}, x_t')' & t \leq R \\ (y_{t+1}, x_t', \hat{y}_{1,t+1}, \ldots, \hat{y}_{m,t+1})' & t > R. \end{cases}$$

1. *Draw B samples of P observations from $\mathrm{X_P}$ using the moving or circular blocks bootstrap with block length b or the stationary bootstrap with geometric block lengths with success probability p. Denote each sample as $\mathrm{X}^*_{\mathrm{P},l}$ and let $\mathrm{X}^*_l = [\mathrm{X}_\mathrm{R}', \mathrm{X}^{*\prime}_{\mathrm{P},l}]'$.*

2. *Estimate $\bar{f}^*_{li}$ and $\hat{\sigma}^*_{li}$ as in Theorem 1 for each bootstrap sample, X^*_l, and each alternative model, $\hat{y}_{it}$.*

This procedure exploits the fact that R is finite in Theorem 1 and that the expected block length will grow with T. Under these asymptotics, a growing proportion of the rolling-window

forecasts are the same if we bootstrap the forecasts as if we bootstrap the empirical data and reestimate the models. Bootstrapping the forecasts makes the computations faster and easier, but requires us to drop the first R observations from the bootstrap (the alternative forecasts are not defined for the first R observations). I add those observations to the beginning of each bootstrap sample so that population means under the bootstrap-induced probability distribution equal sample means.[12] The benchmark model still needs to be reestimated in each bootstrap sample.

Theorem 2 presents the final procedure.

Theorem 2. *Suppose the conditions of Theorem 1 hold, with assumptions on the alternative model, $\hat{y}_{1t}$, holding for each of the m models $\hat{y}_{1t}, \ldots, \hat{y}_{mt}$ and let the subscript l denote the different quantities associated with $\hat{y}_{lt}$ (i.e. $\mathcal{F}_{lt}$, etc.) Also assume that the long-run variance-covariance matrix of the vector $(\bar{f}_1(\beta_0), \ldots, \bar{f}_m(\beta_0))'$ is uniformly positive definite.*

Let the $*$-superscript denote a random variable with the distribution induced by the bootstrap of Algorithm 1 and consider the following procedure:

1. *Define $M_0 = \emptyset$ and, for $j = 1, \ldots, m$, let $M_j = \{l : \frac{1}{\hat{\sigma}_l} \bar{f}_l > \hat{d}_j\}$, where $\hat{d}_j$ is the $1-\alpha$ quantile of $\max_{i \notin M_{j-1}} \frac{1}{\hat{\sigma}_i^*}(\bar{f}_i^* - \bar{f}_i(\hat{\beta}_{T+1}))$.*

2. *Reject all of the models l with $l \in \bigcup_{j=1}^{m} M_j$.*

The following conclusions hold:

1. *For each $G \subset \{1, \ldots, m\}$, let $\mathcal{F}_t^G$ be the smallest σ-field containing $\mathcal{F}_{lt}$ for all $l \in G$. Under the null hypothesis that $y_t - \hat{y}_{0t}$ is an MDS with respect to $\mathcal{F}_t^I$ for some $I \subset \{1, \ldots, m\}$, the procedure outlined above, but using the moving block bootstrap with block length 1, has probability at most α of rejecting one or more models in I.*

2. *Let H_l be the null hypothesis that $y_t - \hat{y}_{0t}$ is an MDS with respect to $\mathcal{F}_{lt}$ against the one-sided alternative $\mathrm{E}\, \bar{f}_i(\beta_0) > 0$. If $p \to 0$ and $Pp \to \infty$ as $T \to \infty$ (for the stationary bootstrap) or $b \to \infty$ and $\frac{b}{P} \to 0$ (for the moving or circular block bootstraps) then the procedure outlined above controls SFWE at level α for the family of null hypotheses $H_1, \ldots, H_m$.*

3. *Let H_l be the null hypothesis that $y_t - \hat{y}_{0t}$ is uncorrelated with $\hat{y}_{lt}$ against the one-sided alternative $\mathrm{E}\, \bar{f}_i(\beta_0) > 0$. If $p \to 0$ and $Pp \to \infty$ as $T \to \infty$ (for the stationary bootstrap) or $b \to \infty$ and $\frac{b}{P} \to 0$ (for the moving or circular block bootstraps) then the procedure outlined above controls SFWE at level α for the family of null hypotheses $H_1, \ldots, H_m$.*

[12] This equality is strictly true only for the stationary and circular block bootstraps.

The following remarks apply to Theorem 2.

Remark 7. The next section will discuss the block bootstrap procedures for OOS comparisons in detail. Note that $\bar{f}(\hat{\beta}_{T+1})$ is the bootstrap equivalent of $\mathrm{E}\, f_t(\beta_0)$, hence its role as the centering term in step 1 of the procedure.

Remark 8. It is worth emphasizing that this procedure is very easy to use under the null hypothesis of correct specification (conclusion 1). The researcher can use a standard estimate of the statistic's asymptotic variance (not a HAC estimator) and an i.i.d. bootstrap. Even better, the alternative forecasts only need to be estimated once, before the bootstrap.

Remark 9. The key difference between conclusions 1 and 2 is that, in conclusion 2, the forecast error is an MDS with respect to several individual series, but not necessarily with respect to the pooled information set generated by all of those series together.[13] Using $b = 1$ would consistently estimate the marginal distribution of each $\frac{1}{\sigma_i}\bar{f}_i$ under this weaker null hypothesis, but would not necessarily estimate the joint distribution of two or more of those terms correctly (in particular, the covariance could be wrong). The stronger null hypothesis imposed in conclusion 1 ensures that resampling with $b = 1$ estimates the joint distribution correctly as well.

The same reasoning implies that a parametric bootstrap (i.e. Clark and McCracken, 2012b) achieves control of SFWE only under the stronger hypotheses of conclusion 1. Fortunately, conclusion 1 seems to capture the goals of most empirical work, so researchers can make use of the considerable simplifications that occur in that setting.

Remark 10. If calculating $\hat{\sigma}_i$ is burdensome, one can do the same bootstrap without studentizing the statistics. This procedure is likely to have worse size and power properties, but may be more convenient. One could also estimate the variance with a second bootstrap step, but execution might take too long to be practical, or use a convenient but inconsistent approximation of the variance. The estimator of the variance in conclusion 3 should have the best performance if it is HAC (the results of Gotze and Kunsch, 1996, may be relevant), but that is not necessary for validity. See Hansen (2005) and Romano and Wolf (2005, Section 4.2) for a discussion of the benefits of studentization.

Remark 11. If the number of alternative models is large, controlling the FWE may be too strict a criterion to be useful. Romano et al. (2008) discuss procedures to control criteria other than the FWE, and one can use their generalizations of the StepM procedure here as well.

[13] It is straightforward to construct three series, u_t, v_t, and w_t, such that w_t is an MDS with respect to $\sigma((u_t, w_t), (u_{t-1}, w_{t-1}), \dots)$ and $\sigma((v_t, w_t), (v_{t-1}, w_{t-1}), \dots)$ but not $\sigma((u_t, v_t, w_t), (u_{t-1}, v_{t-1}, w_{t-1}), \dots)$. This is essentially the scenario imposed in conclusion 2.

Remark 12. Pincheira (2011) raises the issue that, since there are often several models that could be used as the benchmark, researchers may want to control for data snooping over the null hypotheses as well as the alternatives. For example, one could use either Atkeson and Ohanian's (2001) random walk or Stock and Watson's (2007) IMA(1,1) as benchmark inflation models, so it might make sense to require an alternative model to outperform them both.[14] Allowing multiple benchmark models can be expressed as an Intersection-Union test, so Theorem 2's procedure (conclusion 3) easily accommodates this extension by taking each $\bar{f}_i$ and $\hat{\sigma}_i^2$ to be a random q-vector;[15] the first element of $\bar{f}_i$, written as $\bar{f}_{i1}$, compares the first benchmark model to the ith alternative and the first element of $\hat{\sigma}_i^2$ is the corresponding estimate of asymptotic variance, the second element corresponds to the second benchmark model, etc.

Steps 1 and 2 of Theorem 2's procedure now become:

1. Define $M_0 = \emptyset$ and, for $j = 1, \ldots, m$, let $M_j = \{(i,k) : \frac{1}{\hat{\sigma}_{ik}} \bar{f}_{ik} > \hat{d}_j\}$, where $\hat{d}_j$ is the $1-\alpha$ quantile of $\max_{(i,k)\notin M_{j-1}} \frac{1}{\hat{\sigma}^*_{ik}} \left(\bar{f}^*_{ik} - \bar{f}_{ik}(\hat{\beta}_{k,T+1})\right)$.
2. Reject all of the hypotheses H_i such that $(i,k) \in \bigcup_{j=1}^m M_j$ for all k.

Also note that the asymptotic variance-covariance matrix of each $\bar{f}_i$ does not need to be positive definite as long as the variance of each of its elements is positive. A formal statement of this result is given in Appendix A.

3 The Bootstrap for Out-of-Sample Statistics

The validity of the bootstrap in Theorem 2 is a special case of a result of independent interest—the validity in general of block bootstraps for asymptotically normal OOS statistics. This section presents the general result. In this section, let $\to^{p^*}$ and $\to^{d^*}$ refer to convergence in probability or distribution conditional on the observed data. Similarly, E^*, var^*, and cov^* refer to the expectation, variance, and covariance with respect to the probability measure induced by the bootstrap, and y_t^*, etc. is the random variable y_t but under the bootstrap-induced CDF.

The notation in this section is more general than that of Section 2. The assumptions required are generalizations of the conditions of Theorems 1 and 2. See West (1996, 2006), West and McCracken (1998), and McCracken (2000) for a discussion of these conditions, as theirs are nearly identical. Theorem 3 gives the result.

[14] Pincheira (2011) studies tests that reject if the alternative model outperforms either of the models, so the details are somewhat different in our papers.

[15] If a researcher uses multiple benchmark models, it would be a mistake to assume that they are all correctly specified, so only the null hypothesis of conclusion 3 is appropriate.

Theorem 3. *Suppose the following conditions hold:*

1. *The estimator $\hat{\beta}_t$ of β_0 is estimated with a recursive, rolling, or fixed estimation window and satisfies $\hat{\beta}_t - \beta_0 = \hat{B}_t H_t$; $\hat{B}_t$ is a sequence of $k \times q$ matrices such that $\sup_t |\hat{B}_t - B| \to^p 0$; H_t is a sequence of q-vectors such that*

$$H_t = \begin{cases} \frac{1}{t}\sum_{s=1}^{t} h_s(\beta_0) & \textit{recursive window} \\ \frac{1}{R}\sum_{s=t-R+1}^{t} h_s(\beta_0) & \textit{rolling window} \\ \frac{1}{R}\sum_{s=1}^{R} h_s(\beta_0) & \textit{fixed window} \end{cases} \tag{4}$$

   and $\mathrm{E}\, h_s(\beta_0) = 0$ for all s.

2. *Define $\psi_t(\beta) = (f_t(\beta), h_t(\beta))$; $\psi_t(\beta)$ is covariance stationary and twice continuously differentiable in an open neighborhood N of β_0. Also, $\mathrm{E}\sup_{\beta\in N}|\psi_t(\beta)|^r$ and $\mathrm{E}\sup_{\beta\in N}|\frac{\partial}{\partial\beta}\psi_t(\beta)|^r$ are uniformly bounded, and there exists a sequence of random variables m_t such that $\sup_{i,\beta\in N}|\frac{\partial^2}{\partial\beta\partial\beta'}\psi_{it}(\beta)| \le m_t$ almost surely and $\mathrm{E}\, m_t^2$ is uniformly finite.*

3. *Both $\psi_t(\beta)$ and $\frac{\partial}{\partial\beta}\psi_t(\beta)$ are L_2-NED of size $-\frac{1}{2}$ on the series V_t with bounded NED-coefficients for β uniformly in N; V_t is strong mixing of size $-\frac{r}{r-2}$ or uniform mixing of size $-\frac{r}{2r-2}$ with $r > 2$.*[16]

4. *The sequence $\hat{f}_1^*, \ldots, \hat{f}_T^*$ is constructed using the same procedure as the original OOS analysis over the bootstrap sample, where the bootstrap sample is generated by moving blocks, circular blocks, or stationary bootstrap with block lengths drawn from the geometric distribution. The (expected) block length b satisfies $b \to \infty$ and $\frac{b}{P} \to 0$.*

5. *The bootstrapped estimator satisfies $\sup_t |\hat{B}_t^* - B^*| \to^p 0$, $B^* = B + o_p$ and $\frac{1}{T}\sum_{t=1}^{T} h_t^*(\hat{\beta}_{T+1}) = 0$ a.s.*

6. *The asymptotic variance matrix of $\bar{f}(\beta_0)$ is uniformly positive definite.*

7. *$R, P \to \infty$ as $T \to \infty$.*

Then

$$\Pr[\sup_x |\Pr^*[\sqrt{P}(\bar{f}^* - \bar{f}(\hat{\beta}_{T+1})) \le x] - \Pr[\sqrt{P}(\bar{f} - \mathrm{E}\,\bar{f}(\beta_0)) \le x]| > \epsilon] \to 0 \tag{5}$$

for all $\epsilon > 0$.

[16] For uniform mixing processes, $r = 2$ is also allowed as long as $\sup_{\beta\in N}\psi_t(\beta)^2$ and $\sup_{\beta\in N}(\partial\psi_t(\beta)/\partial\beta)^2$ are uniformly integrable.

Remark 13. McCracken (2000) proves asymptotic normality under weaker smoothness conditions on $f_t(\beta)$ and $h_t(\beta)$: only their expectations must be continuously differentiable. It may be possible to extend Theorem 3 to those weaker conditions, but the current proof relies on a theorem of de Jong and Davidson's (2000a) establishing the consistency of HAC estimators and their theorem requires differentiability of the observations. Extending de Jong and Davidson's (2000a) result to nondifferentiable functions should be possible, but is beyond the scope of this paper.

Remark 14. White (2000) and Hansen (2005) resample the forecasts but do not reestimate any of them which requires the additional assumption that $\frac{P}{R}\log\log R \to 0$ or that the forecasts themselves have no estimated parameters.[17]

Remark 15. Corradi and Swanson (2007) use the distribution of $\sqrt{P}(\bar{f}^* - \bar{f})$ to approximate that of $\sqrt{P}(\bar{f} - \mathrm{E}\,\bar{f}(\beta_0))$. But it is clear that $\bar{f}(\hat{\beta}_{T+1})$ is the bootstrap analogue of $\mathrm{E}\,\bar{f}(\beta_0)$, the parameter of interest. Because their bootstrap is miscentered, Corradi and Swanson (2007) must redefine $\hat{\beta}_t^*$ to achieve consistency. In this paper, though, consistency arises naturally.

Remark 16. It may be unnecessary to assume that $\bar{f}(\beta_0)$ has positive definite asymptotic variance; if so, the bootstrap would work in the setup of Clark and McCracken (2001, 2005a) and McCracken (2007). That question is left to future research.

4 Monte Carlo Results

This section presents Monte Carlo experiments demonstrating that this paper's modified version of Clark and West's (2007) statistic performs similarly to their original test in the situations they study, but can have substantially higher power when the DGP has a structural break.[18]

The DGP has three different parametrizations: one to study the tests' size, one to study power under stationarity, and one to study power if there is a single break in the relationship

[17] White (2000) lists several different sets of assumptions that give the same result, but these seem to be the most general.

[18] All of these simulations were programmed in R (R Development Core Team, 2011, version 2.14.0) and use the MASS (Venables and Ripley, 2002, 7.3-22), xtable (Dahl, 2009, version 1.6-0), and dbframe (Calhoun, 2010, version 0.2.7) packages.

between the target and predictors. The DGP is:

$$y_t = \gamma_{1t} + \gamma_{2t} z_{t-1} + e_t \qquad \gamma_t = \begin{cases} (0.5, 0) & \text{size simulations} \\ (0.5, 0.35) & \text{power (stable)} \\ (-0.5, 0) & t \leq \frac{T}{2} \quad \text{power (break)} \\ (1, 0.35) & t > \frac{T}{2} \quad \text{power (break)} \end{cases}$$

$$z_t = 0.15 + 0.95 z_{t-1} + v_t \qquad (e_t, v_t)' \sim iid\ N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 18 & -0.5 \\ -0.5 & 0.025 \end{pmatrix} \right)$$

$$R = 120, 240 \qquad P = 120, 240, 360, 720.$$

Both models are estimated by OLS. The benchmark model regresses y_t on a constant, and the alternative regresses y_t on a constant and z_{t-1}. Clark and West (2007) argue that this DGP mimics an asset pricing application similar to Goyal and Welch's (2008) which we study in Section 5.

For comparison, we study this paper's new statistic as well as Clark and West's (2006, 2007)rolling-window and recursive-window test statistics. Clark and West only prove that their rolling-window statistic is asymptotically normal, and only then if the benchmark model is not estimated, but their recursive-window statistic is popular in practice and in simulations tends to perform similarly to their rolling window test. We use all three of these statistics to test the null that the benchmark model's innovation is an MDS.[19]

Table 1 presents the simulation results. For all of the stable parameter values, the proposed new statistic has similar rejection probability to Clark and West's (2007). Both of Clark and West's tests are generally slightly undersized relative to our new test, which is itself slightly undersized: when R is 120 and P is 360 our test statistic has size 7.6% and Clark and West's rolling and recursive window tests have size 7.5% and 6.2% respectively, at a nominal size of 10%. For the stable alternative, our new statistic typically has slightly higher power than Clark and West's rolling window and lower power than their recursive window. For example, when R is 120 and P is 720, the rolling-window test rejects at 66.8%, our statistic at 73.0%, and the recrusive window statistic at 82.3%, again for a nominal size of 10%. In general, the statistics perform similarly under stability.

For the simulations with a single break, the new statistic has considerably higher power than Clark and West's (2006, 2007) original tests across all of the choices of R and P; the rejection probability is more than twice as large for most parametrizations. When R is 120 and P is 360 with a nominal size of 10%, for example, the new statistic rejects at 96.4% while

[19]Clark and West (2007) report the performance of the tests proposed by Chao et al. (2001) and Clark and McCracken (2005a) as well, and of tests based on the naive Gaussian statistic.

the rolling and recursive window statistics reject at 35.5% and 32.9% respectively. Results for other choices of nominal size and sample split give similar results. So mixing window strategies can give a large power advantage when testing for time-varying predictability, and performs similarly to the original test when testing for stable outperformance.

5 Empirical Illustration

This section demonstrates the use of our new statistic by revisiting Goyal and Welch's (2008) study of excess stock returns. Goyal and Welch argue that many variables thought to predict excess returns (measured as the difference between the yearly log return of the S&P 500 index and the T-bill interest rate) on the basis of in-sample evidence fail to do so out-of-sample. To show this, Goyal and Welch look at the forecasting performance of models using a lag of the variable of interest, and show that these models do not significantly outperform the excess return's recursive sample mean.

Here, I conduct the same analysis, but using this paper's MDS test. The benchmark model is the excess return's sample mean (as in the original) and the alternative models are of the form

$$\text{excess return}_t = \alpha_0 + \alpha_1 \text{ predictor}_{t-1} + \varepsilon_t,$$

where α_0 and α_1 are estimated by OLS using a 10-year window. The predictors used are listed in the "predictor" column of Table 2 (see Goyal and Welch, 2008, for a detailed description of the variables). We also consider Campbell and Thompson's (2008) proposed correction to the models, that the forecasts be bounded below by zero since negative forecasts are incredible, as well as two simple combination forecasts, the mean and the median (over both the original and the non-negative forecasts). The data set is annual data beginning in 1927 and ending in 2009, and the rolling window uses 10 observations.[20]

Table 2 presents the results for each model. The column "value" gives the value of the test statistic for each model, while the "naive" and "corrected" columns indicate whether the statistic is greater than the standard size-10% critical value (1.28) and the critical value estimated by the procedure of Theorem 2 (2.89).[21] Three predictors are significant at the naive critical values for both the original and bounded forecasts: the dividend yield, long term interest rate, and book to market ratio. But none are significant after accounting for data snooping, which highlights the importance of these methods. The median forecast is significant using conventional critical values as well, but not the corrected values.

[20] This statistical analysis was conducted in R (R Development Core Team, 2011) using the xtable (Dahl, 2009, version 1.6-0), and dbframe (Calhoun, 2010, version 0.2.7) packages.

[21] The bootstrap uses 599 replications with i.i.d. sampling, as proposed in conclusion 1 of Theorem 2.

6 Conclusion

This paper presents an OOS test statistic similar to Clark and West's (2006, 2007) that is asymptotically normal when comparing nested or non-nested models. Normality is achieved by estimating the alternative model using a fixed-length rolling window—as do Clark and West—but estimating the benchmark model with a recursive window. Simulations indicate that the new statistic behaves similarly to Clark and West's original test when the DGP is stable but can have much higher power when the DGP has structural breaks. The paper also presents a method for comparing the benchmark model to several alternative models simultaneously and improves block bootstrap procedures of OOS statistics.

A Proofs of Main Theoretical Results

Define the additional notation $F_t(\beta) = \frac{\partial}{\partial\beta} f_t(\beta)$, $F_t = F_t(\beta_0)$, $\hat{f}_t^* = f_t^*(\hat{\beta}_t^*)$, $F_t^*(\beta) = \frac{\partial}{\partial\beta} f_t^*(\beta)$, $F_t^* = F_t^*(\hat{\beta}_{T+1})$, $h_t = h_t(\beta_0)$, $h_t^* = h_t^*(\hat{\beta}_{T+1})$, and

$$H_t^* = \begin{cases} \frac{1}{t}\sum_{s=1}^{t} h_s^* & \text{recursive window} \\ \frac{1}{R}\sum_{s=t-R+1}^{t} h_s^* & \text{rolling window} \\ \frac{1}{R}\sum_{s=1}^{R} h_s^* & \text{fixed window} \end{cases}$$

Proof of Theorem 1. Replace R with $\log(T)$ and P with $T - \log(T)$ in the statistic; this substitution does not affect its asymptotic distribution; the theorem is now an immediate consequence of Lemma B.4 with $h_t(\beta) = x_t(y_{t+1} - x_t'\beta)$. Note that

$$f_t(\beta_0) = 2(y_{t+1} - x_t'\beta_0)(\hat{y}_{1,t+1} - x_t'\beta_0) \quad \text{a.s.},$$

as in Clark and West (2007), which is an MDS, so we do not need to use a HAC estimator of the variance under the null. Also observe that

$$F_t(\beta) = 2x_t(x_t'\beta - \hat{y}_{1,t+1}) + 2x_t(x_t'\beta - y_{t+1})$$

and the second term is an MDS under the null, explaining the particular form of $g_t(\beta)$. □

Proof of Theorem 2. Replace R and P as in the proof of Theorem 1. Theorem 3 of this paper and Theorems 3.1 and 4.1 of Romano and Wolf (2005) complete the proof. □

Proof of Theorem 3. Expand $f_t^*(\hat{\beta}_t^*)$ around $\hat{\beta}_{T+1}$ to get

$$\begin{aligned}\sqrt{P}\big(\bar{f}^* - \bar{f}(\hat{\beta}_{T+1})\big) = \tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} \big(f_t^* - \bar{f}(\hat{\beta}_{T+1})\big) + \mathrm{E}^*\, F_t^* B^* \tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} H_t^* \\ + \tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t^* - \mathrm{E}^*\, F_t^*) B^* H_t^* + \tfrac{1}{\sqrt{P}}\, \mathrm{E}^*\, F_t^* \sum_{t=R+1}^{T} (B_t^* - B^*) H_t^* \\ + \tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t^* - \mathrm{E}^*\, F_t^*)(B_t^* - B^*) H_t^* + \tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} w_t^*\end{aligned}$$

where (as in West, 1996) the ith element of w_t^* is

$$w_{it} = \tfrac{1}{2}(\hat{\beta}_t^* - \hat{\beta}_{T+1})' \Big[\tfrac{\partial^2}{\partial\beta\partial\beta'} f_{it}^*(\tilde{\beta}_{it}^*)\Big](\hat{\beta}_t^* - \hat{\beta}_{T+1})$$

and each $\tilde{\beta}_{it}^*$ lies between $\hat{\beta}_t^*$ and $\hat{\beta}_{T+1}$. The argument that $\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} w_t^* = o_{p^*}(1)$ is identical to West's, using our Lemma B.1 in place of his Lemma A3. Then

$$\tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t^* - \mathrm{E}^*\, F_t^*) B^* H_t^* \to^{p^*} 0 \tag{6}$$

$$\tfrac{1}{\sqrt{P}}\, \mathrm{E}^*\, F_t^* \sum_{t=R+1}^{T} (B_t^* - B^*) H_t^* \to^{p^*} 0 \tag{7}$$

and

$$\tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t^* - \mathrm{E}^*\, F_t^*)(B_t^* - B^*) H_t^* \to^{p^*} 0 \tag{8}$$

from Lemma B.2 and the result follows from Lemma B.3, with the form of the asymptotic variance following directly from West (1996) and West and McCracken (1998). □

Lemma 4 (Formalization of Remark 12)**.** *Suppose that the conditions of Theorem 2 conclusion 3 hold but $f_i = (f_{i1}, \ldots, f_{iq})$ and the asymptotic variance of $(\bar{f}_{1k}(\beta_k), \ldots, \bar{f}_{mk}(\beta_k))$ is uniformly positive definite for each k. Then the procedure described in Remark 12 achieves* SFWE.

The proof is similar to that of Theorem 2. Note that we allow $\sqrt{P}(\bar{f}_{i1}(\beta_1), \ldots, \bar{f}_{iq}(\beta_q))$ to converge to a normal with singular variance-covariance matrix as long as each element has positive variance.

B Supporting Results

Lemma B.1. *Suppose $a \in [0, \frac{1}{2})$ and the conditions of Theorem 3 hold.*

1. $P^a \sup_t |H_t| \to^p 0$ *and* $P^a \sup_t |H_t^*| \to^{p^*} 0$.
2. $P^a \sup_t |\hat{\beta}_t - \beta_0| \to^p 0$ *and* $P^a \sup_t |\hat{\beta}_t^* - \hat{\beta}_T| \to^{p^*} 0$.
3. $\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t - \mathrm{E}\, F_t) = O_p(1)$ *and* $\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t^* - \mathrm{E}\, F_t^*) = O_{p^*}(1)$.

Proof of Lemma B.1. 1. The process $\frac{1}{\sqrt{T}} h_s$ satisfies de Jong and Davidson's (2000b, Theorem 3.1) functional CLT. So

$$P^a \sup_t \left| \tfrac{1}{t} \sum_{s=1}^{t} h_s \right| = P^a \sup_{\gamma \in [0,1]} \left| \tfrac{1}{\lfloor \gamma T \rfloor} \sum_{s=1}^{\lfloor \gamma T \rfloor} h_s \right| \to^p 0 \tag{9}$$

with the convergence following from the continuous mapping theorem. For the bootstrapped average, Calhoun's (2013) Theorem 2 ensures that

$$\tfrac{1}{\sqrt{\gamma T}} \sum_{s=1}^{\lfloor \gamma T \rfloor} h_s^* = O_{p^*}(\Omega^*(\hat{\beta}_{T+1})^{1/2}) \tag{10}$$

where $\Omega^*(\beta) = \mathrm{var}^*\, h_t^*(\beta)$. Since $\hat{\beta}_{T+1} \to^p \beta_0$, it suffices to show that $\sup_{\beta \in N} |\Omega^*(\beta) - \mathrm{var}\, h_t(\beta)| \to^p 0$. But this convergence holds for any fixed β and our assumptions guarantee stochastic equicontinuity as in de Jong and Davidson (2000a) (also see Davidson, 1994), completing the proof.

2. We have

$$\begin{aligned} P^a \sup_t |\hat{\beta}_t - \beta_0| &= P^a \sup_t |\hat{B}_t H_t| \\ &\leq \sup_{t,u} \left| [\hat{B}_u - B] \tfrac{P^a}{t} \sum_{s=1}^{t} h_s \right| + \sup_t \left| B \tfrac{P^a}{t} \sum_{s=1}^{t} h_s \right| \end{aligned} \tag{11}$$

and both terms converge to zero in (conditional) probability by the previous argument and by assumption. The same argument holds for $\hat{\beta}_t^* - \hat{\beta}_{T+1}$ as well.

3. $\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t - \mathrm{E}\, F_t)$ obeys de Jong's (1997) CLT, so the result is trivial. Also,

$$\left| \tfrac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t^* - \mathrm{E}^*\, F^*) \right| = O_{p^*}(\Omega^*(\hat{\beta}_{T+1})^{1/2}) = O_p(1) \tag{12}$$

as in the proof of part 1, where now $\Omega^*(\beta) = \mathrm{var}^*\, F_t^*(\beta)$.

□

Lemma B.2. *Under the conditions of Theorem 3, Equations* (6)–(8) *and* (20)–(22) *hold.*

Proof of Lemma B.2. We can write

$$\left|\tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(F_t^*-\mathrm{E}^*\,F_t^*)B^*H_t^*\right| \leq \left|\tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(F_t^*-\mathrm{E}^*\,F_t^*)B^*\right|\sup_t|H_t^*|.$$

From Lemma B.1, $\sup_t|H_t^*|\to^p 0$ and $\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(F_t^*-\mathrm{E}^*\,F_t^*)=O_p(1)$, establishing (6). The proofs of (7), (8), and (20)–(22) are similar. □

Lemma B.3. *Suppose the conditions of Theorem 3 hold, let*

$$\psi_t^*(\beta)=\begin{pmatrix}f_t^*(\beta)-\bar{f}(\beta)\\ h_t^*(\beta)-\bar{h}(\beta)\end{pmatrix},$$

and define $\psi_t=\psi_t(\beta_0)$, $\hat{\psi}_t=\psi_t(\hat{\beta}_{T+1})$, *and* $\psi_t^*=\psi_t^*(\hat{\beta}_{T+1})$. *Then*

$$\Pr\left[\sup_x\left|\Pr^*\left[\tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\psi_t^*\leq x\right]-\Pr\left[\tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\psi_t\leq x\right]\right|>\epsilon\right]\to 0 \tag{13}$$

for all positive ϵ, *where the inequalities hold element-by-element.*

Proof of Lemma B.3. Note that both ψ_t and ψ_t^* satisfy CLTs, (de Jong, 1997, Theorem 2 and Calhoun, 2013, Theorem 1) so it suffices to prove that

$$\tfrac{1}{P}\operatorname{var}^*\sum_{t=R+1}^{T}\psi_t^*-\tfrac{1}{P}\operatorname{var}\sum_{t=R+1}^{T}\psi_t\to^p 0.$$

I will prove this result for the Circular Block Bootstrap (Politis and Romano, 1992)—the proof for other block bootstraps is similar and follows as in Calhoun (2013).

Since $\mathrm{E}^*\,\psi_t^*=0$ by construction, we have

$$\tfrac{1}{P}\operatorname{var}^*\sum_{t=R+1}^{T}\psi_t^*=\tfrac{1}{bT}\sum_{\tau=1}^{T}\sum_{s,t\in I(\tau)}\hat{\psi}_s\hat{\psi}_t' \tag{14}$$

almost surely, with

$$I(\tau)=\begin{cases}\{\tau,\ldots,\tau+b-1\} & \text{if } \tau\leq T-b+1\\ \{\tau,\ldots,T\}\cup\{1,\ldots,\tau-(T-b+1)\} & \text{if } \tau>T-b+1.\end{cases}$$

We can rewrite the right side of (14) as

$$\tfrac{1}{bT}\sum_{\tau=1}^{T}\sum_{s,t\in I(\tau)}\hat{\psi}_s\hat{\psi}_t' = \tfrac{1}{T}\sum_{s,t=1}^{T}\hat{\psi}_s\hat{\psi}_t'\left(1-\tfrac{|s-t|}{b}\right)^{+} + \tfrac{1}{bT}\sum_{s=1}^{b-1}\sum_{t=T-b+s}^{T}\left(\hat{\psi}_s\hat{\psi}_t'+\hat{\psi}_t\hat{\psi}_s'\right)\left(1-\tfrac{|s-t+T|}{b}\right)^{+}. \quad (15)$$

The first term of (15) can be shown to converge in probability to $\frac{1}{P}\operatorname{var}\sum_{t=R+1}^{T}\psi_t$ by Theorem 2.2 of de Jong and Davidson (2000a). Most of the necessary conditions of their theorem hold by assumption and it remains to prove their condition (2.9) holds, namely that

$$\sup_{\beta\in N}\left\|\tfrac{1}{T}\sum_{t=1}^{T}e^{i\xi t/b}\left(\frac{\partial\psi_{jt}(\beta)}{\partial\beta_k}-\mathrm{E}\,\frac{\partial\psi_{jt}(\beta)}{\partial\beta_k}\right)\right\|_2\to 0 \quad (16)$$

for all j, k, and ξ. For any fixed $\beta\in N$,

$$\tfrac{1}{T}\sum_{t=1}^{T}\cos(\xi t/b)\left(\frac{\partial\psi_{jt}(\beta)}{\partial\beta_k}-\mathrm{E}\,\frac{\partial\psi_{jt}(\beta)}{\partial\beta_k}\right)\to^{p}0$$

and

$$\tfrac{1}{T}\sum_{t=1}^{T}\sin(\xi t/b)\left(\frac{\partial\psi_{jt}(\beta)}{\partial\beta_k}-\mathrm{E}\,\frac{\partial\psi_{jt}(\beta)}{\partial\beta_k}\right)\to^{p}0$$

by a mixingale LLN (Davidson, 1993).[22] Convergence in L_2 for each $\beta\in N$ then follows from the existence of the rth moment ($r>2$) and uniform convergence is a consequence of our bound on the second derivative of ψ_t.

The conclusion then holds if the second term of (15) is $o_p(1)$. We have, for each element of that matrix,

$$\left|\tfrac{1}{bT}\sum_{s=1}^{b-1}\sum_{t=T-b+s}^{T}\hat{\psi}_{ks}\hat{\psi}_{jt}\left(1-\tfrac{|s-t+T|}{b}\right)^{+}\right|\le\tfrac{1}{bT}\sum_{s=1}^{b-1}|\hat{\psi}_{ks}|\sum_{t=T-b+1}^{T}|\hat{\psi}_{jt}|$$

almost surely. For any positive δ,

$$\Pr\left[\tfrac{1}{T}\sum_{s=1}^{b-1}|\hat{\psi}_{is}|>\delta\right]\le\Pr\left[\tfrac{1}{T}\sum_{s=1}^{b-1}\sup_{\beta\in N}|\psi_{is}(\beta)|>\delta\right]+\Pr\left[\hat{\beta}_{T+1}\notin N\right].$$

[22] Remember that NED processes are also mixingales. See, for example, Davidson (1994, Section 17.2).

Both terms converge to zero by assumption, so $\frac{1}{T}\sum_{s=1}^{b-1}|\hat{\psi}_{is}| = o_p(1)$. A similar argument shows that $\frac{1}{b}\sum_{t=T-b+1}^{T}|\hat{\psi}_{jt}| = O_p(1)$, completing the proof. □

Lemma B.4. *If the conditions of Theorem 3, except for those governing the bootstrap process, hold then*

$$\tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\hat{f}_t - \mathrm{E}\, f_t) \to^d N(0,\sigma^2), \tag{17}$$

with

$$\sigma^2 = S_{ff} + \lambda_{fh}(FBS'_{fh} + S_{fh}B'F') + 2\lambda_{hh}FV_\beta F' \tag{18}$$

$$(\lambda_{fh},\lambda_{hh}) = \begin{cases} (1,2)\times\left(1-\frac{1}{\pi}\ln(1+\pi)\right) & \textit{recursive window} \\ \left(\frac{\pi}{2},\pi-\frac{\pi^2}{3}\right) & \textit{rolling window with } \pi\le 1 \\ \left(1-\frac{1}{2\pi},1-\frac{1}{3\pi}\right) & \textit{rolling window with } \pi>1 \\ (0,\pi) & \textit{fixed window,} \end{cases} \tag{19}$$

$\begin{pmatrix} S_{ff} & S_{fh} \\ S'_{fh} & S_{hh} \end{pmatrix}$ *the asymptotic variance of* $(\bar{f}(\beta_0)', \bar{h}(\beta_0)')'$, *and* V_β *the asymptotic variance of* $\hat{\beta}_{T+1}$.

Proof of Theorem 3. As in West (1996) and West and McCracken (1998) (and as in the proof of Theorem 3), expand $\hat{f}_t$ around β_0 to get

$$\begin{aligned}\sqrt{P}\big(\bar{f}-\bar{f}(\beta_0)\big) = {} & \tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}\big(f_t-\mathrm{E}\, f_t\big) + \mathrm{E}\, F_t B \tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T} H_t \\ & + \tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(F_t-\mathrm{E}\, F_t)BH_t + \tfrac{1}{\sqrt{P}}\mathrm{E}\, F_t\sum_{t=R+1}^{T}(B_t-B)H_t \\ & + \tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(F_t-\mathrm{E}\, F_t)(B_t-B)H_t + \tfrac{1}{\sqrt{P}}\sum_{t=R+1}^{T} w_t\end{aligned}$$

where (again, as in West, 1996) the ith element of w_t is

$$w_{it} = \tfrac{1}{2}(\hat{\beta}_t-\beta_0)'\Big[\tfrac{\partial^2}{\partial\beta\partial\beta'}f_{it}(\tilde{\beta}_{it})\Big](\hat{\beta}_t-\beta_0)$$

and each $\tilde{\beta}_{it}$ lies between $\hat{\beta}_t$ and β_0; $\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T} w_t = o_p(1)$ as in Theorem 3. Then, as

before,

$$\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t - \mathrm{E}\, F_t) B H_t \to^p 0 \tag{20}$$

$$\frac{1}{\sqrt{P}} \mathrm{E}\, F_t \sum_{t=R+1}^{T} (B_t - B) H_t \to^p 0 \tag{21}$$

and

$$\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (F_t - \mathrm{E}\, F_t)(B_t - B) H_t \to^p 0 \tag{22}$$

from Lemma B.2. The formula of the asymptotic variance can be derived exactly as in West (1996) and West and McCracken (1998). Since f_t and H_t both satisfy CLTs, this completes the proof. □

References

S. Anatolyev. Inference about predictive ability when there are many predictors. Working Paper, 2007.

A. Atkeson and L. E. Ohanian. Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1):2, 2001.

G. Calhoun. *dbframe: An R to SQL interface*, 2010.

G. Calhoun. Out-of-sample comparisons of overfit models. Working Paper 11002, Iowa State University, 2011.

G. Calhoun. Block bootstrap consistency under weak assumptions. Working Paper 11017, Iowa State University, 2013.

J. Y. Campbell and S. B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008.

J. C. Chao, V. Corradi, and N. R. Swanson. An out of sample test for Granger causality. *Macroeconomic Dynamics*, 5(4):598–620, 2001.

T. E. Clark and M. W. McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110, Nov. 2001.

T. E. Clark and M. W. McCracken. Testing for unconditional predictive ability. In M. P. Clements and D. F. Hendry, editors, *A Companion to Economic Forecasting*, chapter 14, pages 415–440. Blackwell, 2002.

T. E. Clark and M. W. McCracken. Evaluating direct multistep forecasts. *Econometric Reviews*, 24(4):369, 2005a.

T. E. Clark and M. W. McCracken. The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics*, 124(1):1–31, 2005b.

T. E. Clark and M. W. McCracken. Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, 50(2):363–395, 2009.

T. E. Clark and M. W. McCracken. Tests of equal forecast accuracy for overlapping models. Working Paper 11-21, Federal Reserve Bank of Cleveland, 2011.

T. E. Clark and M. W. McCracken. In-sample tests of predictive ability: a new approach. *Journal of Econometrics*, 170(1):1–14, 2012a.

T. E. Clark and M. W. McCracken. Reality checks and nested forecast model comparisons. *Journal of Business and Economic Statistics*, 30(1):53–66, 2012b.

T. E. Clark and K. D. West. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1):155–186, 2006.

T. E. Clark and K. D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311, May 2007.

V. Corradi and W. Distaso. Multiple forecast model evaluation. In M. P. Clements and D. F. Hendry, editors, *Oxford Handbook of Economic Forecasting*, chapter 13, pages 391–414. Oxford University Press, 2011.

V. Corradi and N. R. Swanson. A consistent test for nonlinear out of sample predictive accuracy. *Journal of Econometrics*, 110(2):353–381, Oct. 2002.

V. Corradi and N. R. Swanson. Some recent developments in predictive accuracy testing with nested models and (generic) nonlinear alternatives. *International Journal of Forecasting*, 20(2):185–199, 2004.

V. Corradi and N. R. Swanson. Predictive density evaluation. In G. Elliott, C. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, chapter 5, pages 197–284. Elsevier, 2006.

V. Corradi and N. R. Swanson. Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *International Economic Review*, 48:67–109, 2007.

D. B. Dahl. *xtable: Export tables to LaTeX or HTML*, 2009. URL `http://CRAN.R-project.org/package=xtable`. R package version 1.5-6.

J. Davidson. An L_1-Convergence theorem for heterogeneous mixingale arrays with trending moments. *Statistics & Probability Letters*, 16(4):301–304, Mar. 1993.

J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Advanced Texts in Econometrics. Oxford University Press, 1994.

R. M. de Jong. Central limit theorems for dependent heterogeneous random variables. *Econometric Theory*, 13(3):353–367, 1997.

R. M. de Jong and J. Davidson. Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, 68(2):407–423, Mar. 2000a.

R. M. de Jong and J. Davidson. The functional central limit theorem and weak convergence to stochastic integrals I: Weakly dependent processes. *Econometric Theory*, 16(5):621–642, 2000b.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263, 1995.

R. Giacomini. Testing conditional predictive ability. In M. P. Clements and D. F. Hendry, editors, *Oxford Handbook of Economic Forecasting*, chapter 15, pages 441–456. Oxford University Press, 2011.

R. Giacomini and B. Rossi. Detecting and predicting forecast breakdowns. *Review of Economic Studies*, 76(2):669–705, 2009.

R. Giacomini and B. Rossi. Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620, 2010.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74(6): 1545–1578, 2006.

F. Gotze and H. Kunsch. Second-order correctness of the blockwise bootstrap for stationary observations. *The Annals of Statistics*, 24(5):1914–1933, 1996.

A. Goyal and I. Welch. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008.

P. R. Hansen. A test for superior predictive ability. *Journal of Business and Economic Statistics*, 23(4):365–380, 2005.

P. R. Hansen, A. Lunde, and J. M. Nason. The model confidence set. *Econometrica*, 79(2): 453–497, 2011.

D. I. Harvey, S. J. Leybourne, and P. Newbold. Tests for forecast encompassing. *Journal of Business & Economic Statistics*, 16(2):254–259, Apr. 1998.

P.-H. Hsu, Y.-C. Hsu, and C.-M. Kuan. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17 (3):471–484, 2010.

K. Hubrich and K. D. West. Forecast evaluation of small nested model sets. *Journal of Applied Econometrics*, 25(4):574–594, 2010.

A. Inoue and L. Kilian. In-sample or out-of-sample tests of predictability: which one should we use? *Econometric Reviews*, 23(4):371–402, 2004.

A. Inoue and L. Kilian. On the selection of forecasting models. *Journal of Econometrics*, 130(2):273–306, Feb. 2006.

A. Inoue and B. Rossi. Out-of-sample forecast tests robust to the window size choice. ERID Working Paper 94, Duke University, 2011.

L. Kilian. Exchange rates and monetary fundamentals: What do we learn from long-horizon regressions? *Journal of Applied Econometrics*, 14(5):491–510, 1999.

N. C. Mark. Exchange rates and fundamentals: Evidence on long-horizon predictability. *American Economic Review*, 85(1):201–218, 1995.

M. W. McCracken. Data mining and out-of-sample inference. Manuscript, Louisiana State University, 1998.

M. W. McCracken. Robust out-of-sample inference. *Journal of Econometrics*, 99(2):195–223, 2000.

M. W. McCracken. Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, 140(2):719–752, Oct. 2007.

M. W. McCracken and K. D. West. Inference about predictive ability. In M. P. Clements and D. F. Hendry, editors, *A Companion to Economic Forecasting*, chapter 21, pages 299–321. Blackwell, 2002.

R. A. Meese and K. Rogoff. Empirical exchange rate models of the seventies: do they fit out of sample? *Journal of International Economics*, 14(1-2):3–24, Feb. 1983.

F. Merlevède and M. Peligrad. On the coupling of dependent random variables and applications. pages 171–193.

A. J. Patton and A. Timmermann. Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics*, 140(2):884–918, 2007a.

A. J. Patton and A. Timmermann. Testing forecast optimality under unknown loss. *Journal of the American Statistical Association*, 102(480):1172–1184, 2007b.

M. H. Pesaran and A. Timmermann. Real-time econometrics. *Econometric Theory*, 21(01): 212–231, 2005.

M. H. Pesaran and A. Timmermann. Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1):134–161, 2007.

P. Pincheira. A bunch of models, a bunch of nulls and inference about predictive ability. Working Paper 607, Central Bank of Chile, 2011.

D. N. Politis and J. P. Romano. A circular block resampling procedure for stationary data. In R. Page and R. LePage, editors, *Exploring the Limits of Bootstrap*, pages 263–270. John Wiley, New York, 1992.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`.

J. P. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

J. P. Romano, A. M. Shaikh, and M. Wolf. Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447, 2008.

B. Rossi. Testing long-horizon predictive ability with high persistence, and the Meese-Rogoff puzzle. *International Economic Review*, 46(1):61–92, 2005.

B. Rossi and T. Sekhposyan. Understanding models' forecasting performance. *Journal of Econometrics*, 164(1):158–172, 2011a.

B. Rossi and T. Sekhposyan. Forecast optimality tests in the presence of instabilities. ERID Working Paper 109, Duke University, 2011b.

J. H. Stock and M. W. Watson. Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, 41(3):788–829, 2003.

J. H. Stock and M. W. Watson. Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39:3–33, 2007.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4th edition, 2002.

K. D. West. Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–1084, Sept. 1996.

K. D. West. Forecast evaluation. In G. Elliott, C. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 99–134. Elsevier, 2006.

K. D. West and M. W. McCracken. Regression-based tests of predictive ability. *International Economic Review*, 39(4):817–840, 1998.

H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

Sim. type	R	P	Pr[CW roll.]	Pr[CW rec.]	Pr[new]
size	120	120	7.3	7.8	7.8
		240	5.5	5.5	6.2
		360	7.5	6.2	7.6
		720	8.5	5.4	7.2
	240	120	7.2	7.2	7.8
		240	6.3	6.5	7.2
		360	6.8	5.9	6.9
		720	7.0	5.9	7.4
power (stable)	120	120	26.2	30.0	29.3
		240	39.2	47.2	42.4
		360	47.3	59.8	50.9
		720	66.8	82.3	73.0
	240	120	34.5	36.1	34.3
		240	45.9	50.1	46.7
		360	56.7	63.8	56.9
		720	78.2	87.0	78.5
power (breaks)	120	120	25.9	29.9	62.9
		240	30.1	31.0	87.2
		360	35.5	32.9	96.4
		720	46.1	38.2	99.8
	240	120	28.1	30.6	58.2
		240	37.6	36.1	87.8
		360	43.1	39.0	97.4
		720	56.9	42.5	100.0

Table 1: Size and power of the OOS tests in the simulations described by Section 4, at 10% confidence. These percentages are calculated from 2000 samples. Pr[CW roll.] shows the fraction of simulations for which Clark and West's (2007) rolling-window statistic rejects; Pr[CW rec.] shows the fraction of simulations for which their recursive-window statistic rejects; and Pr[new] shows the fraction of simulations for which this paper's test rejects.

	value	naive	corrected
book to market CT	2.06	sig.	
long term rate CT	1.64	sig.	
median	1.59	sig.	
long term rate	1.56	sig.	
book to market	1.41	sig.	
dividend yield CT	1.32	sig.	
dividend yield	1.27		
stock variance CT	1.23		
dividend payout ratio CT	1.20		
average	1.04		
dividend price ratio	0.95		
treasury bill CT	0.89		
dividend price ratio CT	0.81		
default yield spread CT	0.71		
net equity	0.70		
net equity CT	0.68		
earnings price ratio CT	0.65		
dividend payout ratio	0.64		
treasury bill	0.53		
stock variance	0.50		
inflation CT	0.20		
default return spread	0.16		
default return spread CT	0.12		
default yield spread	0.09		
inflation	−0.09		
term spread CT	−0.29		
term spread	−0.43		
earnings price ratio	−0.56		
long term yield	−0.74		
long term yield CT	−0.89		

Table 2: Results from OOS comparison of equity premium prediction models; the benchmark is the recursive sample mean of the equity premium and each alternative model is a constant and single lag of the variable listed in the "predictor" column. The dataset begins in 1927 and ends in 2009 and is annual data. The "value" column lists the value of this paper's OOS statistic, the "naive" column indicates whether the statistic is significant at standard critical values, and the "corrected" column indicates significance using the critical values proposed in Theorem 2 that account for the number of models. See Section 5 for details.