An asymptotically normal out-of-sample test based on mixed estimation windows

Gray Calhoun*
Iowa State University

January 9th, 2015

Abstract

This paper develops a modification of Clark and West's (2007, *J. Econom.*) adjusted out-of-sample t-test. We propose using a recursive window to estimate the benchmark model but a fixed-length rolling window to estimate the alternative. The resulting statistic is asymptotically normal even when the models are nested. The paper also presents Monte Carlo evidence that this statistic has much higher power than existing out-of-sample statistics in a common use-case for these tests: when the DGP is subject to instability. This procedure is then used to analyze Goyal and Welch's (2008, *Rev. Finan. Stud.*) excess returns dataset and supports their finding that the equity premium is unpredictable out-of-sample.

Keywords: Forecast Evaluation, Martingale Difference Sequence, Model Selection, Forecast Encompassing

JEL Classification Numbers: C22, C53

*Economics Department; Iowa State University; Ames, IA 50011. Telephone: (515) 294-6271. Email: «gcalhoun@iastate.edu», web: «www.econ.iastate.edu/~gcalhoun». I'd like to thank Helle Bunzel, Todd Clark, Graham Elliott, Yu-Chin Hsu, Michael McCracken, Pablo Pincheira, Elie Tamer, Allan Timmermann, Ken West, Stephane Meng-Feng Yen, several referees, and participants at the 2011 Midwest Econometrics Group meeting and the 2013 NBER-NSF Time Series conference 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). An earlier version of this paper circulated under the name "An asymptotically normal out-of-sample test of equal predictive accuracy for nested models."

1 Introduction

This paper proposes an out-of-sample (OOS) test statistic that is asymptotically normal and correctly centered even when the models studied are nested. The test is based on one proposed by Clark and West (2006, 2007), but we propose estimating the benchmark model with a recursive window and the alternative model with a fixed length rolling window. The rolling window ensures asymptotic normality, as in Giacomini and White (2006), and the recursive window allows the null hypothesis to be a statement about the specification of the Data Generating Process, which is the focus of the vast majority of the OOS testing literature.[1] This combination of estimation windows also gives our test statistic high power against alternatives that cause the benchmark model to be unstable — structural breaks, time-varying coefficients, or forms of nonlinearity, for example — which is a common motivation for using these tests.

OOS tests are common 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). 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 (MDS). 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 to have mean-zero under the null. This statistic is especially convenient because other OOS tests for similar hypotheses (Chao et al. 2001; Clark and McCracken 2001, 2005; 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 not estimated (Clark and West, 2006). Estimating the parameters of the smaller model invalidates their proof.

[1] In particular, this is the focus of West (1996), Clark and McCracken (2001), McCracken (2007), Clark and West (2006), Clark and West (2007), and many others but not Giacomini and White (2006). See West (2006) and Clark and McCracken (2013) for a thorough overview of this literature.

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

This paper proposes a modified version of Clark and West's (2006, 2007) statistic and shows that it is asymptotically normal even when the smaller model is estimated. To achieve normality, the pseudotrue benchmark model must be estimated consistently, but the larger alternative model must continue to be estimated inconsistently so that the test statistic is not degenerate when the models are nested. 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. This approach has the further advantage over existing OOS tests for nested models that 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. Moreover, although we focus on nested models in this paper, the approach can be used with non-nested models as well. As Clark and McCracken (2011) have recently argued, West's (1996) results do not hold when the true DGP is nested by the benchmark and alternative models, which is allowable under the null hypothesis of interest. (Clark and McCracken, 2011, call this scenario "overlapping models.")

The next section presents the intuition and theory for our new statistic. Section 3 presents simulations that compare our pairwise OOS test to Clark and West's (2006, 2007) original statistics. Section 4 demonstrates the use of our statistic by reanalyzing Goyal and Welch's (2008) study of excess return predictability and demonstrates how our results can be used in settings with many alternative models. Section 5 concludes. Our results follow from arguments similar to West's (1996) and have been put in a separate appendix along with some supporting lemmas (Calhoun, 2015).

2 Theoretical results supporting the asymptotically normal OOS statistic

This section presents the new OOS statistic; first we give an informal motivation of the statistic, then present the paper's key assumptions in Section 2.1 and present our formal theoretical results in Section 2.2.

Suppose for now that a researcher is interested in predicting the target variable y_{t+1} with a vector of regressors x_t, that v_t is another random process that is believed to potentially contain information about y_{t+1}, and that (y_t, x_t, v_t) is stationary and weakly dependent. In addition, let $\beta_0 = (\mathrm{E}\, x_t x_t')^{-1} \mathrm{E}\, x_t y_{t+1}$ be the pseudotrue coefficient for the regression of y_{t+1} on x_t and define $\varepsilon_{t+1} = y_{t+1} - x_t'\beta_0$. If this linear model is correctly specified, then ε_{t+1} is an MDS with respect to $\sigma((x_t, v_t, y_t), (x_{t-1}, v_{t-1}, y_{t-1}), \dots)$ and we can see immediately that

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

obeys an MDS CLT and is asymptotically normal as $P \to \infty$,[3] with R an arbitrary starting value and $P = T - R$.

Straightforward algebra (Clark and West, 2007) shows that

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \varepsilon_{t+1}(v_t - x_t'\beta_0) = \frac{1}{2\sqrt{P}} \sum_{t=R}^{T-1} \left[(y_{t+1} - x_t\beta_0)^2 - (y_{t+1} - v_t)^2 + (x_t'\beta_0 - v_t)^2 \right] \quad (2)$$

almost surely. Clark and West (2006, 2007) base their OOS statistics on the RHS of Equation (2), but use a second forecast of y_{t+1} as v_t. (Call it $\hat{y}_{t+1}$.) They use a rolling window of length R to estimate $\hat{y}_{t+1}$,[4] and R is kept finite as $T \to \infty$ so that $\hat{y}_{t+1}$ inherits the weak dependence properties of the variables used to estimate it. Using a finite window prevents the degeneracy that can arise when comparing nested models out-of-sample (see Clark and McCracken, 2001, and McCracken, 2007), so the conditional variance of the OOS average remains positive and the average obeys a CLT.[5]

Clark and West (2006, 2007) propose using this as a test of whether the benchmark is correctly specified. In their 2006 paper, Clark and West assume that the coefficients on the benchmark model, β_0, are zero under the null, making ε_{t+1} observed directly. This restriction is relaxed in their 2007 paper, where β_0 is unknown and estimated with the same length-R rolling window as $\hat{y}_{t+1}$. Now the estimated linear model's prediction errors, $\hat{\varepsilon}_{t+1}$, replace ε_{t+1} in the OOS test statistic. 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 simulations. Since the window length is finite, the estimator of β_0 does not converge to β_0.

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

$$\hat{\beta}_t = \Big(\sum_{s=1}^{t-1} x_s x_s' \Big)^{-1} \sum_{s=1}^{t-1} x_s y_{s+1} \qquad \text{and} \qquad \hat{\varepsilon}_{t+1} = y_{t+1} - x_t'\hat{\beta}_t \quad (3)$$

for each t.[6] West's (1996) Theorem 4.1 implies that

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

is asymptotically normal with mean zero under Clark and West's MDS null for fairly arbitrary processes v_t, as long as v_t is weakly dependent and the OOS statistic has

[3] This claim assumes that the asymptotic variance of the sample average is uniformly positive, a requirement that we will address in Section 2.2.

[4] Making $\hat{y}_{t+1}$ a function of $y_t, x_{t-1}, z_{t-1} \ldots, y_{t-R+1}, x_{t-R}$ and z_{t-R}, where z_t is another weakly dependent random process

[5] This approach was first introduced by Giacomini and White (2006).

[6] The matrix inversion in $\hat{\beta}_t$ can be replaced with a pseudo-inverse if necessary for some values of t without changing the forecast.

uniformly positive variance. Just as in Clark and West (2006, 2007), these conditions are ensured if v_t is another forecast of y_{t+1} based on a fixed-length rolling window.

So far, we have presented an especially simple version of the result to make the intuition as clear as possible. The next section lists the specific assumptions for the more general case and defines additional notation.

2.1 Theoretical assumptions

Consider the following environment. There is a single linear benchmark model of the target variable, y_{t+1}:

$$y_{t+1} = x_t'\beta + \varepsilon_{t+1}, \quad t = 1, \ldots, T-1 \tag{4}$$

where β is an unknown vector of parameters and x_t is an observed vector of predictors. The parameter β is estimated with OLS using a recursive window as described by Equation (3). The alternative model is denoted $\hat{y}_{t+1}$ and is estimated with a rolling window of length R.

The main conditions on the DGP are summarized in the first assumption. The weak dependence and moment conditions are standard. The assumption of strict stationarity is stronger than necessary in practice — once the alternative forecasting method is known, it is only necessary that the OOS adjusted loss difference be weak stationary, and even that can be relaxed further — but this stronger assumption ensures that the results hold generally.

Assumption 1. *The data are generated by the relationship*

$$y_{t+1} = x_t'\beta_0 + \varepsilon_{t+1} \tag{5}$$

for $t = 1, 2, \ldots$, for some value β_0, with $\mathrm{E}\, x_t\varepsilon_{t+1} = 0$, $\mathrm{E}\,\varepsilon_{t+1}^2 > 0$, and $\mathrm{E}\, x_t x_t'$ positive definite for all t. Also assume that there is an additional sequence of random vectors z_t and the process $(\varepsilon_{t+1}, x_t, z_t)$ is stationary and strong mixing of size $-r/(r-2)$ or uniform mixing of size $-r/(2r-2)$, for $r > 2$.

The next assumption defines the forecasting models and adds additional constraints to the DGP.

Assumption 2. *The benchmark forecast is $x_t'\hat{\beta}_t$, where $\hat{\beta}_t$ is constructed with a recursive window according to* (3)*. The alternative forecast satisfies*

$$\hat{y}_{t+1} = \psi(y_t, z_t, \ldots, y_{t-R+1}, z_{t-R+1}) \tag{6}$$

where ψ is a known measurable function and the window length, R, remains finite as $T \to \infty$. Moreover, the vector $(\varepsilon_{t+1}, x_t, \hat{y}_{t+1})$ has uniformly bounded $2r$ moments where r is first defined in Assumption 1.

The requirement that the alternative forecast satisfies moment conditions, rather than the underlying predictors z_t, is somewhat unappealing but necessary. The function ψ that generates these forecasts is otherwise nearly unrestricted, so even well-behaved predictors could produce arbitrarily badly-behaved forecasts. For example, if

$$z_t \sim \textit{i.i.d. bernoulli}(1/2),$$

setting $\psi(y_t, z_t) = 1/z_t$ would prevent a CLT from holding since the forecast equals positive infinity with probability $1/2$. It is easy to construct less obvious examples of problematic functions as well. Assumption 2 implicitly rules out these functional forms by imposing moment conditions on the alternative models' forecasts.

Our next assumption ensures that the asymptotic variance of the OOS average is positive.

Assumption 3. *The asymptotic variance-covariance matrix*

$$\operatorname{var}\left(\tfrac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\begin{pmatrix} x_t \\ \hat{y}_{t+1}\end{pmatrix}\varepsilon_{t+1}\right) \tag{7}$$

is uniformly positive definite (in T).

This assumption 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 the alternative model consist of only the first regressor of the benchmark. 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.

The final assumption restricts the class of HAC variance estimators we will consider. We use the same class of estimators studied by de Jong and Davidson (2000) (their class $\mathscr{K}$); see their paper for further discussion.

Assumption 4. *The kernel K is a function from $\mathbb{R}$ to $[-1,1]$ such that $K(0) = 1$, $K(x) = K(-x)$ for all x, $K(\cdot)$ is continuous at zero and all but a finite number of points, and*

$$\int_{-\infty}^{\infty} |K(x)|\,dx < \infty,$$

and

$$\int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} K(z)e^{ixz}\,dz \right| dx < \infty.$$

Last, we define some notation that will be used to derive the theoretical properties of our OOS statistics. The information set that contains the information available for forecasting y_{t+1} is

$$\mathscr{F}_t = \sigma(y_t, x_t, z_t, y_{t-1}, x_{t-1}, z_{t-1}, \dots).$$

The adjusted OOS loss difference using a hypothetical value of β to produce the benchmark forecast is denoted by

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

Define the additional terms $\hat{f}_t = f_t(\hat{\beta}_t)$, $f_t = f_t(\beta_0)$,

$$\hat{g}_t = 2\Big[\tfrac{1}{P}\sum_{s=R}^{T-1}(x_s'\hat{\beta}_s - \hat{y}_{s+1})x_s'\Big]\Big[\tfrac{1}{T-1}\sum_{s=1}^{T-1}x_s x_s'\Big]^{-1} x_t\hat{\varepsilon}_{t+1}$$

and

$$g_t = 2\,\mathrm{E}\big[(x_t'\beta_0 - \hat{y}_{t+1})x_t'\big](\mathrm{E}\,x_t x_t')^{-1}x_t\varepsilon_{t+1}$$

and the OOS averages $\bar{f} = \sum_{t=R}^{T-1}\hat{f}_t/P$, $\bar{f}^* = \sum_{t=R}^{T-1} f_t/P$, $\bar{g} = \sum_{t=R}^{T-1}\hat{g}_t/P$, and $\bar{g}^* = \sum_{t=R}^{T-1} g_t/P$.

2.2 Theoretical results

Asymptotic normality of the OOS average now follows directly from the first three assumptions without other conditions. The proof is presented in the Appendix and follows West (1996) closely.

Theorem 1. *If Assumptions 1–3 hold then*

$$\sqrt{P}(\bar{f} - \mathrm{E}\,\bar{f}^*) \to^d N(0, \sigma^2),$$

with $\sigma^2 = s_1 + 2(s_2 + s_3)$ *and*

$$s_1 = \lim \mathrm{var}(\sqrt{P}\,\bar{f}^*), \qquad s_2 = \lim \mathrm{cov}(\sqrt{P}\,\bar{f}^*, \sqrt{P}\,\bar{g}^*), \qquad s_3 = \lim \mathrm{var}(\sqrt{P}\,\bar{g}^*).$$

To use this result, we need a consistent estimator of σ^2. Define the HAC covariance estimator $\hat{\sigma}_1^2 = \hat{s}_{11} + 2(\hat{s}_{12} + \hat{s}_{13})$ and the MDS covariance estimator $\hat{\sigma}_2^2 = \hat{s}_{21} + 2(\hat{s}_{22} + \hat{s}_{23})$ with

$$\hat{s}_{11} = \tfrac{1}{P}\sum_{s,t=R}^{T-1}(\hat{f}_s - \bar{f})(\hat{f}_t - \bar{f})K(\tfrac{t-s}{P}), \qquad \hat{s}_{21} = \tfrac{1}{P}\sum_{t=R}^{T-1}(\hat{f}_t - \bar{f})^2,$$

$$\hat{s}_{12} = \tfrac{1}{P}\sum_{s,t=R}^{T-1}(\hat{f}_s - \bar{f})(\hat{g}_t - \bar{g})K(\tfrac{t-s}{P}), \qquad \hat{s}_{22} = \tfrac{1}{P}\sum_{t=R}^{T-1}(\hat{f}_t - \bar{f})(\hat{g}_t - \bar{g}),$$

and

$$\hat{s}_{13} = \tfrac{1}{P}\sum_{s,t=R}^{T-1}(\hat{g}_s - \bar{g})(\hat{g}_t - \bar{g}), \qquad \hat{s}_{23} = \tfrac{1}{P}\sum_{t=R}^{T-1}(\hat{g}_t - \bar{g})^2.$$

These estimators are consistent under similar assumptions to Theorem 1.

Lemma 2. *If Assumptions 1–4 hold then*

$$\hat{\sigma}_1^2 \to^p \sigma^2.$$

If Assumptions 1–3 hold and $\{\varepsilon_t, \mathscr{F}_t\}$ *is an MDS then*

$$\hat{\sigma}_2^2 \to^p \sigma^2.$$

Note that these results allow misspecification; asymptotic normality follows from the weak dependence of the underlying series and from the design of the test statistic. These statistics have typically been used to test the null hypothesis that the benchmark model is correctly specified — that $\{\varepsilon_t, \mathscr{F}_t\}$ is an MDS — which implies that f_t is an MDS as discussed at the beginning of this section. This is especially appealing in our framework, since the benchmark can be theoretically motivated so the MDS null would be a test of rationality. For example, Goyal and Welch (2008) test whether excess returns for the S&P 500 are predictable out-of-sample, and any deviation of ε_{t+1} from an MDS is potentially interesting. But the MDS null hypothesis only affects the estimator of σ^2 (see Lemma 2); Theorem 1 continues to hold under any DGP that satisfies Assumptions 1 – 3.

In other settings, a researcher may want to test the weaker hypothesis that $\mathrm{E}\,\bar{f}^* = 0$ but the benchmark may be misspecified. Our statistic can then be interpreted as an encompassing test as in Harvey et al. (1998), and would test whether the alternative model contains additional information that could make the benchmark model more accurate. This interpretation can be motivated by the combination forecasting model

$$\hat{y}_{avg,t+1} = (1-w)x_t'\beta_0 + w\hat{y}_{t+1}$$

which can be rewritten in terms of forecast errors as

$$y_{t+1} - \hat{y}_{avg,t+1} = \varepsilon_{t+1} + w(x_t'\beta_0 - \hat{y}_{t+1}).$$

The value

$$w = \frac{\mathrm{E}\,\varepsilon_{t+1}(\hat{y}_{t+1} - x_t'\beta_0)}{\mathrm{E}(x_t'\beta_0 - \hat{y}_{t+1})^2}$$

minimizes the MSE of the combination forecast, so the combination model will have smaller MSE than the benchmark model, implying that the alternative uses information not in the benchmark, unless ε_{t+1} and $\hat{y}_{t+1} - x_t'\beta_0$ are uncorrelated. This correlation is exactly the quantity measured by our statistic.

The final result puts together Theorem 1 and Lemma 2 to produce our test statistics. The null hypothesis under misspecification is written in terms of $\mathrm{E}\,\varepsilon_{t+1}\hat{y}_{t+1}$ and not $\mathrm{E}\,\varepsilon_{t+1}(\hat{y}_{t+1} - x_t'\beta_0)$, since $\mathrm{E}\,\varepsilon_{t+1}x_t = 0$ by construction. This result is an immediate consequence of the previous two results and its proof is omitted.

Theorem 3. *If Assumptions 1–4 hold, then*

$$\sqrt{P}\bar{f}/\hat{\sigma}_1 \to^d N(0,1)$$

under the null hypothesis $\mathrm{E}(\varepsilon_{t+1}\hat{y}_{t+1}) = 0$ *for all* $t = R, \ldots, T-1$. *If, instead, Assumptions 1–3 hold, then*

$$\sqrt{P}\bar{f}/\hat{\sigma}_2 \to^d N(0,1)$$

under the null hypothesis that $\{\varepsilon_t, \mathscr{F}_t\}$ *is an MDS.*

The test statistic proposed in Theorem 3 can be easily extended in several ways. For longer-horizon forecasts (two or more periods ahead), $\hat{\sigma}_1$ will remain consistent but $\hat{\sigma}_2$ will not — the forecast errors for a correctly specified h-step-ahead forecast have an MA$(h-1)$ dependence structure — but using a generalized $\hat{\sigma}_2$ that reflects this covariance structure restores consistency. To test optimality under loss functions other than squared-error, one can replace the forecast error with the generalized forecast error (see, for example Patton and Timmermann, 2007a,b) and replace the OLS estimator of β with the corresponding M-estimator. And the benchmark model can be replaced in general with a nonlinear model that satisfies the assumptions of West (1996) or McCracken (2000) by making the appropriate changes to f_t and g_t. (See West, 1996, and McCracken, 2000, for details.) The general approach of using a recursive window to estimate the benchmark and a fixed-length rolling window to estimate the alternative applies quite broadly.

3 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.[7]

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

[7] 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) package.

relationship between the target and predictors. The DGP is:

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

$$x_{t+1} = 0.15 + 0.95 x_t + u_{t+1} \quad (\varepsilon_t, u_t)' \sim iid\ N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 18 & -0.5 \\ -0.5 & 0.025 \end{pmatrix}\right)$$

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

Both models are estimated by OLS. The benchmark model regresses y_{t+1} on a constant, and the alternative regresses y_{t+1} on a constant and x_t. 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 4.

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.[8]

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 recursive 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 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.

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

Sim. type	R	P	Pr[CW roll.]	Pr[CW rec.]	Pr[new]
size	120	120	7.2	8.0	7.5
		240	5.6	5.6	6.2
		360	7.2	6.1	7.2
		720	8.5	5.4	7.2
	240	120	7.2	7.2	7.7
		240	6.3	6.5	7.1
		360	6.8	5.9	6.8
		720	7.0	5.9	7.3
power (stable)	120	120	26.2	30.0	29.2
		240	39.2	47.2	42.4
		360	47.3	59.8	51.1
		720	66.8	82.3	73.1
	240	120	34.5	36.1	34.1
		240	45.9	50.1	46.9
		360	56.7	63.8	56.9
		720	78.2	87.0	78.7
power (breaks)	120	120	25.9	29.9	62.2
		240	30.1	31.0	87.4
		360	35.5	32.9	96.5
		720	46.1	38.2	99.8
	240	120	28.1	30.6	58.2
		240	37.6	36.1	87.7
		360	43.1	39.0	97.2
		720	56.9	42.5	100.0

Table 1: Size and power of the OOS tests in the simulations described by Section 3, 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.

4 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, we 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

$$\textit{excess return}_{t+1} = \beta_0 + \beta_1 \textit{ predictor}_t + \varepsilon_{t+1},$$

where β_0 and β_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.[9]

Table 2 presents the results for each model. The column *value* gives the value of the test statistic for each model, while *naive* indicates whether the statistic is greater than the standard 10% critical value (1.28). 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, and the median forecast is significant as well. This could suggest that excess returns are not an MDS and that information in these three variables is useful for predicting returns.

However, we know that this is an extremely optimistic assessment of the models' performance. We are conducting 30 simultaneous hypothesis tests, so it is likely that some will reject by chance. There are several approaches that could accommodate this multiplicity and a full treatment is beyond the scope of this paper, however, it is straightforward to use our results to derive a valid critical value similar to White (2000).

Let $\bar{f}_i$ be the OOS statistic associated with the ith alternative forecast, $\hat{y}_{i,t+1}$. The arguments underlying our results apply essentially unchanged to multivariate f_t, so the continuous mapping theorem implies that

$$\max_{i=1,\ldots,30} \sqrt{P}\bar{f}_i/\hat{\sigma}_{2i} \to^d \max_{i=1,\ldots,30} W_i,$$

where $W \sim N(0, V)$ and V is the 30×30 correlation matrix with elements

$$V_{ij} = \lim \frac{\operatorname{cov}(\bar{f}_i, \bar{f}_j)}{\operatorname{var}(\bar{f}_i)^{1/2} \operatorname{var}(\bar{f}_j)^{1/2}}.$$

To estimate V, we use the correlation matrix associated with the multivariate analogue of $\hat{\sigma}_2$,

$$\tfrac{1}{P}\sum_{t=R}^{T-1}\left[(\hat{f}_t - \bar{f})(\hat{f}_t - \bar{f})' + (\hat{f}_t - \bar{f})(\hat{g}_t - \bar{g})' + (\hat{g}_t - \bar{g})(\hat{f}_t - \bar{f})' + 2(\hat{g}_t - \bar{g})(\hat{g}_t - \bar{g})'\right]$$

[9] This statistical analysis was conducted in R (R Development Core Team, 2011) and uses the MASS (Venables and Ripley, 2002, 7.3-22) package.

	value	naive	corrected
book to market CT	2.04	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	1.30	sig.	
dividend yield	1.26		
stock variance CT	1.22		
dividend payout ratio CT	1.18		
average	1.04		
dividend price ratio	0.95		
treasury bill CT	0.89		
dividend price ratio CT	0.82		
default yield spread CT	0.70		
net equity	0.70		
net equity CT	0.69		
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	−0.73		
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 critical values that account for the number of models. See Section 4 for details.

where $\hat{f}_t$ and $\hat{g}_t$ are vectors with ith elements

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

and

$$\hat{g}_{it} = 2\Big[\tfrac{1}{P}\sum_{s=R}^{T-1}(x_s'\hat{\beta}_s - \hat{y}_{i,s+1})x_s'\Big]\Big[\tfrac{1}{T-1}\sum_{s=1}^{T-1}x_s x_s'\Big]^{-1}x_t\hat{\varepsilon}_{t+1}$$

respectively. Call this estimate $\hat{V}$ and let $\hat{c}$ denote the 0.90 quantile of the distribution of $\max_i \hat{W}_i$, with $\hat{W} \sim N(0,\hat{V})$. Then

$$\limsup_{T\to\infty} \Pr\Big[\max_{i=1,\ldots,30} \sqrt{P}\bar{f}_i/\hat{\sigma}_{2i} > \hat{c}\Big] \leq 0.10,$$

under the null hypothesis that excess returns are an MDS with respect to all of the information contained in the variables listed in Table 2, making $\hat{c}$ an asymptotically valid critical value.[10]

We calculate $\hat{c}$ by generating 1999 draws from $N(0,\hat{V})$, giving a value of 2.49, and the *corrected* column of Table 2 denotes the models that remain significant at 10% with this critical value. Using this critical value, none of the predictors are significant, which gives additional support to Goyal and Welch's (2008) conclusion that excess returns are unpredictable and also demonstrates the importance of correcting for multiplicity in these studies.

5 Discussion

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. We also have presented an empirical study of the equity premium that demonstrates how to use these results with several alternative models.

References

G. Calhoun. Supplemental appendix for "An asymptotically normal out-of-sample test based on mixed estimation windows". Available at http://www.econ.iastate.edu/~gcalhoun, 2015.

[10] Hansen (2005) makes the point that multiple one-sided comparisons can have poor power if irrelevant predictors are included in these tests and proposes a threshold for discarding very poor forecasts. His threshold is well below our worst performing model, so this issue is not a concern here.

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. Evaluating direct multistep forecasts. *Econometric Reviews*, 24(4):369, 2005.

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. Advances in forecast evaluation. In *Handbook of Economic Forecasting*, volume 2, pages 1107–1201. North Holland, 2013.

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 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.

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

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

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

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.

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

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.

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.

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

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.