Testing Predictive Ability and Power Robustification

Kyungchul Song

University of British Columbia, Department of Economics, 997-1873 East Mall, Vancouver, BC, V6T 1Z1, Canada. (*kysong@mail.ubc.ca*)

January 12, 2012

Abstract

One of the approaches to compare forecasting methods is to test whether the risk from a benchmark prediction is smaller than the others. The test can be embedded into a general problem of testing inequality constraints using a one-sided sup functional. Hansen (2005) showed that such tests suffer from asymptotic bias. This paper generalizes this observation, and proposes a hybrid method to robustify the power properties by coupling a one-sided sup test with a complementary test. The method can also be applied to testing stochastic dominance or moment inequalities. Simulation studies demonstrate that the new test performs well relative to the existing methods. For illustration, the new test was applied to analyze the forecastability of stock returns using technical indicators employed in White (2000).

Key words and Phrases: Inequality Restrictions, Testing Predictive Ability, One-sided Sup Tests, Power Robustification, Reality Check, Data Snooping

JEL Classifications: C12, C14, C52, C53.

1 Introduction

Comparing multiple forecasting methods is important in practice. Diebold and Mariano (1995) proposed tests comparing two forecasting methods, and West (1996) offered a formal analysis of inference based on out-of-sample predictions. White (2000) developed a general testing framework for multiple forecasting models, and Hansen (2005) developed a way to improve the power of the tests in White (2000). Giacomini and White (2006) introduced out-of-sample predictive ability tests that can be applied to conditional evaluation objectives. There also has been interest in evaluation of density forecasts in the literature. See for earlier contributions Diebold, Gunther, and Tay (1998), Christoffersen (1998), and Diebold, Hahn, and Tay (1999), and for more recent researches, Amisano and Giacomini (2007), and Bao, Lee, and Saltoğlu (2007), among others. For a general survey of forecast evaluations, see West (2006) and references therein.

This paper's main focus is on one-sided sup tests of predictive ability developed by White (2000) and applied by Sullivan, Timmerman, and White (1999). A one-sided sup test is based on the maximal difference between the benchmark and the candidate forecast performances. Hansen (2005) demonstrated that the one-sided sup tests are asymptotically biased, and suggested a way to improve their power. His approach is general and, in fact, related to some later literatures on testing moment inequalities such as Andrews and Soares (2007), Bugni (2010), Andrews and Shi (2009), and Linton, Song and Whang (2010).

This paper generalizes the observation by Hansen (2005), and proposes a method to robustify the local asymptotic power behavior. The main idea is to *couple* the one-sided sup test with a complementary test that shows better power properties against alternative hypotheses under which the one-sided sup test performs poorly. For a complementary test, this paper adopts a symmetrized test statistic used by Linton, Massoumi and Whang (2005) for their stochastic dominance tests. This paper calls this coupled test a *hybrid test*, as its power properties are intermediate between those of the one-sided sup test and the symmetrized test. As this paper demonstrates, one can easily apply a bootstrap procedure to obtain approximate critical values for the hybrid test by

using the existing bootstrap procedures.

Many recent researches have focused on the testing problem of inequality restrictions that hold simultaneously under the null hypothesis. See, for example, Hansen (2005), Andrews and Soares (2007), Bugni (2010), Canay (2010), Linton, Song, and Whang (2010), Andrews and Shi (2010), and references therein. This paper's approach makes contrast with the proposals mentioned above. In the context of testing predictive ability, these proposals improve the finite sample power properties of the test by eliminating forecasting methods that perform poorly beyond a threshold when computing a critical value. Since using a fixed threshold makes the test asymptotically invalid, the threshold is chosen to be less stringent as the sample size becomes larger, satisfying certain rate conditions. On the other hand, this paper's approach modifies the sup test to have a better local power against alternatives that the original test is known to have weak power, and hence using a sequence of thresholds is not required.

A test of such inequality restrictions is said to be *asymptotically similar on the boundary*, if the asymptotic rejection probability remains the same whenever any of the inequality restrictions is binding under the null hypothesis. A recent paper by Andrews (2011) showed an interesting result that a test of such inequality restrictions that is asymptotically similar on the boundary has poor power properties under general conditions. Like the researches mentioned previously, the hybrid test of this paper improves power properties by alleviating asymptotic bias of the one-sided test against certain alternatives, but does not eliminate entirely the asymptotic nonsimilarity of the one-sided test. Hence the test is not subjected to the poor power problem pointed out by Andrews (2011).

The performance of the hybrid test is investigated through Monte Carlo simulation studies. Overall, the new test performs as well as the tests of White (2000) and Hansen (2005), and in some cases, performs conspicuously better.

This paper applies the hybrid test to investigate the forecastability of S&P500 stock returns by technical indicators in a spirit similar to the empirical application in White (2000). Considering the

periods from March 28, 2003 through July 1, 2008, the empirical application tests the null hypothesis that no method among the 3,654 candidate forecasting methods considered by White (2000) outperforms the benchmark method based on sample means. The hybrid test has conspicuously lower p-values than the tests of Hansen (2005) and White (2000). A brief explanation behind this finding is provided in the paper.

The paper is organized as follows. The next section discusses poor power properties of one-sided sup tests. Section 3 introduces a general method of coupling the one-sided test with a complementary one. Sections 4 and 5 present and discuss results from Monte Carlo simulation studies, and an empirical application on stock returns forecastability. Section 6 concludes.

2 Testing Predictive Ability and Asymptotic Bias

In producing a forecast, one typically adopts a forecasting model, estimates the unknown parameter, and then produces a forecast using the estimated forecasting model. Since a forecasting model and an estimation method constitute eventually a single map that assigns past observations to a forecast, we follow Giacomini and White (2006) and refer to this map generically as a *forecasting method*. Given information $\mathcal{F}_T$ at time T, multiple forecasting methods are generically described by maps φ_m, $m \in \mathbf{M}$, from $\mathcal{F}_T$ to a forecast, where $\mathbf{M} \subset \mathbf{R}$ denotes the set of the indices for the forecasting methods. The set $\mathbf{M}$ can be a finite set or an infinite set that is either countable or uncountable. Let $\Lambda(m)$ denote the risk of prediction based on the m-th candidate forecasting method, and $\Lambda(0)$ the risk of prediction based on a benchmark method. Then the difference in performance between the two methods is measured by

$$d(m) \equiv \Lambda(0) - \Lambda(m).$$

We are interested in testing whether there is a candidate forecasting method that strictly dominates the benchmark method. The null and the alternative hypotheses are written as:

$$\begin{aligned} H_0 &: d(m) \leq 0, \text{ for all } m \in \mathbf{M} \text{ and} \\ H_1 &: d(m) > 0, \text{ for some } m \in \mathbf{M}. \end{aligned} \tag{1}$$

Let us consider some examples of $\Lambda(m)$.

Example 1 (Point Forecast Evaluated with the Mean Squared Prediction Error) Suppose that there is a time series $\{(Y_t, \mathbf{X}_t^\top)\}_{t=1}^\infty$, where we observe part of it, say, $\mathcal{F}_T \equiv \{(Y_t, \mathbf{X}_t^\top)\}_{t=1}^T$. The object of forecast is a τ-ahead quantity $Y_{T+\tau}$. There are M number of candidate forecasts $\hat{Y}_{T+\tau}^{(m)}$, $m = 1, \cdots, M$. Each forecast is generated by $\hat{Y}_{T+\tau}^{(m)} = f_m(\mathcal{F}_T; \hat{\beta}_{m,T})$, where $\hat{\beta}_{m,T}$ is a quantity estimated using $\mathcal{F}_T$, and $f_m(\cdot; \beta)$ the m-th forecasting model known up to β. Since $\hat{\beta}_{m,T}$ is estimated using $\mathcal{F}_T$, we can write $\hat{\beta}_{m,T} = \tau_m(\mathcal{F}_T)$ for some map τ_m. When we write $\varphi_m(\mathcal{F}_T) = f_m(\mathcal{F}_T; \tau_m(\mathcal{F}_T))$, the forecasting method is represented as a single map $\hat{Y}_{T+\tau}^{(m)} = \varphi_m(\mathcal{F}_T)$. One way to define the risk $\Lambda(m)$ is to adopt the mean squared prediction error:

$$\Lambda(m) = \mathbf{E}[\{Y_{T+\tau} - \hat{Y}_{T+\tau}^{(m)}\}^2].$$

The expectation above is with respect to the joint distribution of variables constituting information $\mathcal{F}_T$ and $Y_{T+\tau}$. ■

Example 2 (Density Forecast Evaluated with the Expected Kullback-Leibler Divergence): Let $\{(Y_t, \mathbf{X}_t^\top)\}_{t=1}^\infty$ and $\mathcal{F}_T \equiv \{(Y_t, \mathbf{X}_t^\top)\}_{t=1}^T$ be as in Example 1. The object of forecast in this example is the density $f_{T+\tau}$ of a τ-ahead quantity $Y_{T+\tau}$. Suppose that $f_{m,T+\tau}(\cdot; \mathcal{F}_T)$ is the density forecast obtained using the m-th forecasting method and information $\mathcal{F}_T$. Following Bao, Lee, and Saltoğlu (2007), we may take the expected Kullback-Leibler divergence as a measure of

discrepancy between the forecast density $f_{m,T+\tau}(\cdot;\mathcal{F}_T)$ and the actual density $f_{T+\tau}(\cdot)$:

$$KL(m) = \int \log\left(f_{T+\tau}(y)\right) f_{T+\tau}(y)dy - \mathbf{E}\left[\int \log\left(f_{m,T+\tau}(y;\mathcal{F}_T)\right) f_{T+\tau}(y)dy\right].$$

where $f_{T+\tau}$ is the true density of $Y_{T+\tau}$. Since the first integral does not depend on the choice of a forecasting method, we focus on the second part only in comparing the methods. Hence we take

$$\Lambda(m) = -\mathbf{E}\left[\int \log\left(f_{m,T+\tau}(y;\mathcal{F}_T)\right) f_{T+\tau}(y)dy\right]$$

and define $e(m) = \Lambda(0) - \Lambda(m)$. ■

Example 3 (Conditional Forecast Evaluated with the Mean Squared Prediction Error): Giacomini and White (2006) proposed a general framework of testing conditional predictive abilities. Let $\{(Y_t, \mathbf{X}_t^\top)\}_{t=1}^\infty$ and $\mathcal{F}_T \equiv \{(Y_t, \mathbf{X}_t^\top)\}_{t=1}^T$ be as in Example 1. Let $\hat{Y}_{T+\tau}^{(m)} = \varphi_m(\mathcal{F}_T)$ be a point forecast of $Y_{T+\tau}$ using the m-th method φ_m, and define the conditional mean squared prediction error

$$\Lambda(m|\mathcal{G}_T) = \mathbf{E}[\{Y_{T+\tau} - \varphi_m(\mathcal{F}_T)\}^2|\mathcal{G}_T],$$

where $\mathcal{G}_T$ is part of the information $\mathcal{F}_T$. For example, the forecast is generated from $\hat{Y}_{T+\tau}^{(m)} = f_m(\mathcal{F}_T;\hat{\beta}_m)$ as in Example 1. Then the null hypothesis of interest is stated as $d(m|\mathcal{G}_T) \leq 0$ for all $m \in \mathbf{M}$, where $d(m|\mathcal{G}_T) = \Lambda(0|\mathcal{G}_T) - \Lambda(m|\mathcal{G}_T)$, with $\Lambda(0|\mathcal{G}_T)$ denoting the benchmark method's conditional mean squared prediction error. The null hypothesis states that regardless of how the information $\mathcal{G}_T$ realizes, the performance of the benchmark method dominates all the candidate methods. As in Giacomini and White (2006), we choose an appropriate test function $h(\mathcal{G}_T)$, and focus on testing (1) with $d(m) = \Lambda(0) - \Lambda(m)$, where $\Lambda(m) \equiv \mathbf{E}[h(\mathcal{G}_T)\{Y_{T+\tau} - \varphi_m(\mathcal{F}_T)\}^2]$. A remarkable feature of Giacomini and White (2006) is that their testing procedure is designed to capture the effect of estimation uncertainty when a fixed sample size is used for the estimation even as $T \to \infty$. This feature is also accommodated in this paper's framework. ■

The usual method of testing (1) involves replacing $\Lambda(m)$ by an estimator $\hat{\Lambda}(m)$ and constructing an appropriate test using $\hat{d}(m) = \hat{\Lambda}(0) - \hat{\Lambda}(m)$. For Examples 1-3 above, we can construct:

$$\begin{aligned}
\hat{\Lambda}(m) &= \frac{1}{T-R+1}\sum_{t=R}^{T}\{Y_{t+\tau} - f_m(\mathcal{F}_t;\hat{\beta}_{m,t})\}^2 \text{ (Example 1)} \\
\hat{\Lambda}(m) &= -\frac{1}{T-R+1}\sum_{t=R}^{T}\log(f_{m,t+\tau}(Y_{t+\tau};\mathcal{F}_t)) \text{ (Example 2) and} \\
\hat{\Lambda}(m) &= \frac{1}{T-R+1}\sum_{t=R}^{T}h(\mathcal{G}_t)\{Y_{t+\tau} - f_m(\mathcal{F}_t;\hat{\beta}_{m,t})\}^2 \text{ (Example 3),}
\end{aligned}$$

where the periods $R+\tau, \cdots, T+\tau$ are target periods of forecast. (For obtaining $f_{m,t+\tau}(Y_{t+\tau};\mathcal{F}_t)$ in Example 2, see Bao, Lee, and Saltoğlu (2007) and Amisano and Giacomini (2007) for details.) Let n denote the number of the time series observations used to produce $\hat{d}(m)$. The random quantity $\hat{d}(m)$ is viewed as a stochastic process indexed by $m \in \mathbf{M}$, or briefly a random function $\hat{d}(\cdot)$ on $\mathbf{M}$. The main assumption for this paper is the following:

Assumption 1: There exists a Gaussian process Z with a continuous sample path on $\mathbf{M}$ such that

$$\sqrt{n}\{\hat{d}(\cdot) - d(\cdot)\} \Longrightarrow Z(\cdot), \text{ as } n \to \infty, \tag{2}$$

where $\Longrightarrow$ denotes weak convergence of stochastic processes on $\mathbf{M}$.

When $\mathbf{M} = \{1, 2, \cdots, M\}$, Assumption 1 is satisfied if for $\hat{d}= [\hat{d}(1), \cdots, \hat{d}(M)]^\top$ and $d= [d(1), \cdots, d(M)]^\top$,

$$\sqrt{n}(\hat{\mathbf{d}} - \mathbf{d}) \xrightarrow{d} \mathbf{Z} \equiv [Z(1), \cdots, Z(M)]^\top \sim N(0, \Omega) \tag{3}$$

for a positive semidefinite matrix $\mathbf{\Omega}$ (i.e., for all $\mathbf{t} \in \mathbf{R}^M$ $\mathbf{t}^\top\mathbf{Z}$ is zero if $\mathbf{t}^\top\mathbf{\Omega}\mathbf{t} = 0$, and $\mathbf{t}^\top\mathbf{Z} \sim N(0, \mathbf{t}^\top\mathbf{\Omega}\mathbf{t})$ if $\mathbf{t}^\top\mathbf{\Omega}\mathbf{t} > 0$). Therefore, the predictive models are allowed to be nested as in White (2000) and Giacomini and White (2006). In the situation where the forecast sample is small relative to the estimation sample, the estimation error in $\hat{\beta}_{m,t}$ becomes irrelevant. On the other

hand, in the situation where parameter estimation uses a rolling window of observations with a fixed window length as in Giacomini and White (2006), the estimation error in $\hat{\beta}_{m,t}$ remains relevant in asymptotics. Assumption 1 accommodates both the situations.

Assumption 1 also admits infinite $\mathbf{M}$. The case arises, for example, when one tests the conditional mean squared prediction error as in Example 3 using *a class of* test functions instead of using a single choice of h (e.g. Stinchcombe and White (1998)). In this case, $\mathbf{M}$ also includes the indices of such a class of test functions.

However, Assumption 1 excludes the case where $\Lambda(\cdot)$ is a regression function of a continuous random variable and $\hat{\Lambda}(\cdot)$ is its nonparametric estimator. While such a case arises rarely in the context of testing predictive abilities, it does in the context of testing certain conditional moment inequalities (e.g. Lee, Song, and Whang (2011)).

We first show that tests based on the one-sided sup test statistic:

$$T^K \equiv \sqrt{n} \sup_{m \in \mathbf{M}} \hat{d}(m) \tag{4}$$

are asymptotically biased. To define a local power function, we introduce Pitman local alternatives in the direction a:

$$d(m) = a(m)/\sqrt{n}. \tag{5}$$

The direction a represents how far and in which direction the alternative hypothesis is from the null hypothesis. For example, suppose that $\mathbf{M} = \{1, 2, \cdots, M\}$, i.e., we have M candidate forecasting methods, and consider an alternative hypothesis with a such that $a(1) = c > 0$ and $a(m) = 0$ for all $m = 2, \cdots, M$. The alternative hypothesis in this case is such that the first forecasting method ($m = 1$) has a risk smaller than that of the benchmark method by $c/\sqrt{n}$.

Proposition 1: *Suppose that Assumption 1 holds. Then for any* $c_\alpha > 0$ *with* $\lim_{n\to\infty} P\left\{T^K > c_\alpha\right\} \leq$

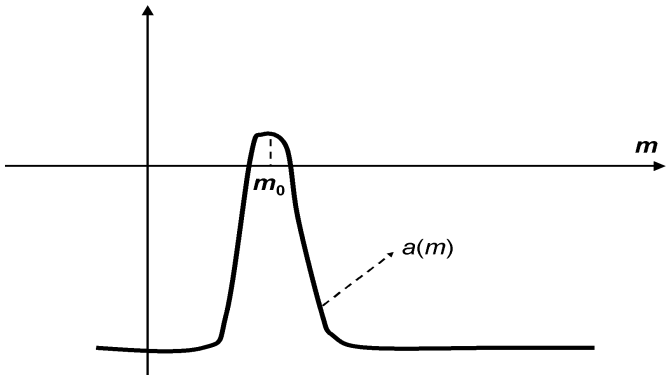

Figure 1: Illustration of An Alternative Hypothesis Associated with Low Power: Recall that when $a(m)$ takes a positive value for some m, the situation corresponds to an alternative hypothesis. When $a(m)$ is highly negative at m's away from m_0 and $a(m_0)$ is positive but close to zero, $\sup_{m\in\mathbf{M}} Z(m) + a(m)$ is close to $Z(m_0)$ with high probability, making it likely that the rejection probability is below α.

α under H_0, there exists a map $a : \mathbf{M} \to \mathbf{R}$ such that under the local alternatives of type (5),

$$\lim_{n\to\infty} P_a \left\{ T^K > c_\alpha \right\} \leq \alpha - \Delta + \varepsilon,$$

where P_a denotes the sequence of probabilities under (5) *and* $\Delta = P\{\sup_{m\in\mathbf{M}} Z(m) > c_\alpha\} - \inf_{m\in\mathbf{M}} P\{Z(m) > c_\alpha\}$.

Proposition 1 shows that the sup test of (1) has a severe bias when Δ is large. Proposition 1 relies only on generic features of the testing set-up such as (1) and (2), and hence also applies to many inequality tests in contexts beyond those of testing predictive abilities. A general version of Proposition 1 and its proof is found in the supplemental note.

The intuition behind Proposition 1 is simple. Suppose that $P\{\sup_{m\in\mathbf{M}} Z(m) > c_\alpha\} = \alpha$. Then, the asymptotic power under the local alternatives with a is given by $P\{\sup_{m\in\mathbf{M}} Z(m)+a(m) > c_\alpha\}$. Suppose we take $a(m)$ of the form in Figure 1 with $a(m_0)$ positive but close to zero, whereas for other m's, $a(m)$ is very negative. Then $\sup_{m\in\mathbf{M}} Z(m) + a(m)$ is close to $Z(m_0) + a(m_0)$ with high

probability. Since $a(m_0)$ is close to zero,

$$P\left\{\sup_{m\in\mathbf{M}} Z(m) + a(m) > c_\alpha\right\} \leq P\left\{Z(m_0) > c_\alpha\right\} + \varepsilon$$

for small $\varepsilon > 0$. Since typically $P\left\{Z(m_0) > c_\alpha\right\} < P\left\{\sup_{m\in\mathbf{M}} Z(m) > c_\alpha\right\} = \alpha$, we obtain the asymptotic bias result.

3 Power Robustification via Coupling

3.1 A Complementary Test

The previous section showed that the sup test has very poor power against certain local alternatives. This section proposes a hybrid test that improves power against such local alternatives. Given $\hat{d}$ as before, we construct another test statistic:

$$T^S = \min\left\{\max_{m\in\mathbf{M}} \hat{d}(m), \max_{m\in\mathbf{M}}(-\hat{d}(m))\right\}. \tag{6}$$

This type of test statistic was introduced by Linton, Massoumi, and Whang (2005) for testing stochastic dominance. For testing (1), T^S is *complementary* to T^K in the sense that using T^S results in a greater power against such local alternatives that the test T^K performs very poorly.

To illustrate this point, let $\mathbf{M} = \{1, 2\}$ and X_1 and X_2 be given observations which are positively correlated and jointly normal with a mean vector $d = [d(1), d(2)]^\top$. We are interested in testing

$$\begin{aligned} H_0 &: d(1) \leq 0 \text{ and } d(2) \leq 0, \text{ against} \\ H_1 &: d(1) > 0 \text{ or } d(2) > 0. \end{aligned}$$

Consider $T^K = \max\{X_1, X_2\}$ and $T^S = \min\{\max\{X_1, X_2\}, \max\{-X_1, -X_2\}\}$. Complementarity between T^K and T^S is illustrated in Figure 2 in a form borrowed from Hansen (2005). The ellipses

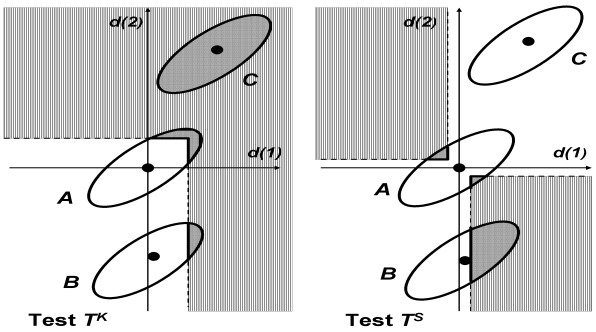

Figure 2: Complementarity of T^K and T^S : Both panels depict three elipses that represent contours from the joint density of (X_1, X_2) under different probabilities. Contour A corresponds to the least favorable configuration under the null hypothesis, and contours B and C different probabilities under the alternative hypothesis. The lighter gray areas on both panels represent the rejection regions of the tests. Under the alternative hypothesis of type C, the test T^K has a better power than T^S, while under the alternative hypothesis of type B, the test T^S has a better power than the test T^K (as shown by a larger dark area within the contour B for the test T^S than for the test T^K.)

in Figure 2 indicate representative contours of the joint density of X_1 and X_2, each corresponding to different distributions denoted by A, B, and C. While A represents the null hypothesis under a least favorable configuration (LFC), i.e., $d(1) = d(2) = 0$, B and C represent alternative hypotheses. Under B, the rejection probability of the test T^K is lower than that under A, implying the biasedness of the test. (This is illustrated by the dark area in ellipsis B in the left panel which is smaller than the dark area in ellipsis A in the same panel.) However, the rejection probability of the test T^S against B is better than the test T^K as indicated by a larger dark area in the ellipsis B on the right panel than that on the left panel. (This contrast may be less stark when X_1 and X_2 are negatively correlated.) Hence against B, test T^S has a better power than test T^K. This order of performance is reversed in the case of an alternative C where the test T^S has a power close to zero while the test T^K has a power close to 1.

3.2 Coupling

We construct a hybrid test by coupling T^K and T^S. For a given level $\alpha \in [0,1]$ and $\gamma \in [0,1]$, we define a hybrid test of (1) as follows:

$$\begin{aligned} \text{Reject } H_0 \text{ if } T^S &> c_\alpha^S(\gamma) \\ &\text{or} \\ \text{if } T^S &\leq c_\alpha^S(\gamma) \text{ and } T^K > c_\alpha^K(\gamma), \end{aligned} \tag{7}$$

where $c_\alpha^S(\gamma)$ and $c_\alpha^K(\gamma)$ are threshold values such that

$$\begin{aligned} \lim_{n\to\infty} P\{T^S &> c_\alpha^S(\gamma)\} = \alpha\gamma \text{ and} \\ \lim_{n\to\infty} P\{T^S &\leq c_\alpha^S(\gamma) \text{ and } T^K > c_\alpha^K(\gamma)\} = \alpha(1-\gamma). \end{aligned}$$

The hybrid test runs along a locus between T^K and T^S as we move γ between 0 and 1. When γ is close to 1, the hybrid test becomes close to T^S, and when γ is close to 0, it becomes close to T^K. The power-reducing effect of the negativity of $d(m)$ for most m's on the test T^K is counteracted by the power-enhancing effect of the positivity of $-d(m)$ for most m's on the test T^S. By coupling with T^S, the hybrid test shares this counteracting effect. Without reasons to do otherwise, this paper proposes using $\gamma = 1/2$.

Critical values can be computed using bootstrap. First, we simulate the bootstrap distribution P^* of (T^S, T^K) by generating $(T_b^{S*}, T_b^{K*})_{b=1}^B$, where B denotes the bootstrap number. (When observations are stationary series, this can be done using the stationary bootstrap method of Politis and Romano (1994). See also for details White (2000) and Hansen (2005).) Using the empirical distribution of $\{T_b^{S*}\}_{b=1}^B$, we first compute $c_\alpha^{S*}(\gamma)$ such that

$$\frac{1}{B}\sum_{b=1}^{B} 1\{T_b^{S*} > c_\alpha^{S*}(\gamma)\} = \alpha\gamma. \tag{8}$$

Given $c_\alpha^{S*}(\gamma)$, we can take $c_\alpha^{K*}(\gamma)$ to be the $(1-\alpha(1-\gamma))$-percentile of the bootstrap series, $T_b^{K*} \cdot 1\{T_b^{S*} \leq c_\alpha^{S*}(\gamma)\},\ b = 1, \cdots, B$.

The method of coupling hardly entails additional computational cost. The computational cost in most cases arises when one computes $\hat{d}^*(m)$ using the bootstrap samples, which is a step common in other bootstrap-based tests. Once $\hat{d}^*(m)$ is computed, finding T_b^{K*} and T_b^{S*} and obtaining bootstrap critical values are straightforward.

We define the p-values for the test as follows:

$$\hat{p} = \sup\left\{\alpha \in [0,1] : T^S \leq c_\alpha^{S*}(\gamma) \text{ and } T^K \leq c_\alpha^{K*}(\gamma)\right\}, \tag{9}$$

where $c_\alpha^{S*}(\gamma)$ and $c_\alpha^{K*}(\gamma)$ are critical values defined in (8). The event that $T^S \leq c_\alpha^{S*}(\gamma)$ and $T^K \leq c_\alpha^{K*}(\gamma)$ arises if and only if the hybrid test does not reject the null hypothesis. In practice, one starts from $\alpha = 0$ and increases along a grid point until $T^S > c_\alpha^{S*}(\gamma)$ or $T^K > c_\alpha^{K*}(\gamma)$. Since bootstrap statistics T_b^{K*} and T_b^{S*} have already been computed, the grid search can be done very fast.

3.3 A Recursive Search for a Better Forecasting Method

When the search for a better forecasting method is an ongoing process with candidate models continuing to expand at each search, it is convenient to have a search algorithm that properly takes account of the past searches. White (2000) proposed such an algorithm for practitioners. In this section, we similarly offer the method of recursive search based on the hybrid test.

Given bootstrap versions $\{\hat{d}_b^*(m)\}_{b=1}^B$, $m = 1, \cdots, M$, and a consistent asymptotic variance estimator $\hat{\omega}^2(m)$ such that $\sqrt{n}(\hat{d}(m) - d(m))/\hat{\omega}(m) \xrightarrow{d} N(0,1)$, we define

$$\tilde{d}_b^*(m) = \hat{d}_b^*(m) - \hat{d}(m). \tag{10}$$

(One may construct $\hat{\omega}^2(m)$ using an HAC (heteroskedasticity-autocorrelation consistent) type es-

timator as done in Hansen (2005), p. 372.) Now, the recursive search that this paper suggests proceeds as follows:

Step 1: For model 1, compute $\hat{d}(1) = \hat{\Lambda}(0) - \hat{\Lambda}(1)$, its asymptotic variance estimator $\hat{\omega}^2(1)$, and the bootstrap version $\{\hat{d}_b^*(1)\}_{b=1}^B$. Set $T_{1,+}^K = \sqrt{n}\hat{d}(1)/\hat{\omega}(1)$, $T_{1,-}^K = -\sqrt{n}\hat{d}(1)/\hat{\omega}(1)$, $T_1^S = \min\{T_{1,+}^K, T_{1,-}^K\}$, and bootstrap versions, $T_{1,+,b}^{K*} = \sqrt{n}\tilde{d}_b^*(1)/\hat{\omega}(1)$, $T_{1,-,b}^{K*} = -\sqrt{n}\tilde{d}_b^*(1)/\hat{\omega}(1)$, and $T_{1,b}^{S*} = \min\{T_{1,+,b}^{K*}, T_{1,-,b}^{K*}\}$.

Step m: For model m, we compute $\hat{d}(m) = \hat{\Lambda}(0) - \hat{\Lambda}(m)$, its asymptotic variance estimator $\hat{\omega}^2(m)$, and the bootstrap version $\{\hat{d}_b^*(m)\}_{b=1}^B$. Set $T_{m,+}^K = \max\{\sqrt{n}\hat{d}(m)/\hat{\omega}(m), T_{m-1,+}^K\}$, $T_{m,-}^K = \max\{-\sqrt{n}\hat{d}(m)/\hat{\omega}(m), T_{m-1,-}^K\}$, $T_m^S = \min\{T_{m,+}^K, T_{m,-}^K\}$, and bootstrap versions

$$\begin{aligned} T_{m,+,b}^{K*} &= \max\left\{\sqrt{n}\tilde{d}_b^*(m)/\hat{\omega}(m), T_{m-1,+,b}^{K*}\right\}, \\ T_{m,-,b}^{K*} &= \max\left\{-\sqrt{n}\tilde{d}_b^*(m)/\hat{\omega}(m), T_{m-1,-,b}^{K*}\right\}, \text{ and} \\ T_{m,b}^{S*} &= \min\{T_{m,+,b}^{K*}, T_{m,-,b}^{K*}\}. \end{aligned}$$

At each step m, the bootstrap p-value can be computed as in (9), where we replace T_b^{S*}, T_b^{K*} and T^S, T^K by $T_{m,b}^{S*}$, $T_{m,+,b}^{K*}$ and T_m^S, $T_{m,+}^K$.

The recursive search at Step m carries along the history of previous searches done using the same data. Similarly as in the spirit of White (2000), we emphasize that as for the previous searches, the recursive search at any Step m requires only knowledge of $T_{m-1,+}^K$, $T_{m-1,-}^K$, $\{T_{m-1,+,b}^{K*}\}_{b=1}^B$, and $\{T_{m-1,-,b}^{K*}\}_{b=1}^B$. In other words, one does not need to know the entire history of the searches and performances, before one turns to the next search.

4 Monte Carlo Simulations

4.1 Data Generating Processes and Three Tests

The first part of simulations focuses on the simulation design considered by Hansen (2005) and compare three types of tests, a test of White (2000) (Reality Check: RC), a test of Hansen (2005) (Superior Predictive Ability: SPA) and this paper's proposal (Hybrid Test: Hyb). The second part considers local alternatives that are different from those of Hansen (2005).

Suppose that $\hat{Y}_{T+\tau}^{(m)}$ is a τ-step ahead forecast of $Y_{T+\tau}$ using the m-th method. The relative performance is represented by $L(Y_{T+\tau}, \hat{Y}_{T+\tau}^{(m)})$ for some loss function L, and we simply write $L_{m,T} = L(Y_{T+\tau}, \hat{Y}_{T+\tau}^{(m)})$. Suppose that $\hat{Y}_{T+\tau}^{(0)}$ is a forecast from a benchmark method. The risk difference is given by $d(m) = \mathbf{E}\left[L_{0,T} - L_{m,T}\right]$.

In simulation studies, we drew for $m = 1, 2, \cdots, M$ and $t = 1, 2, \cdots, n$,

$$L_{m,t} \sim \text{i.i.d. } N(\lambda(m)/\sqrt{n}, \sigma_m^2)$$

for constants $\lambda(m)$ and $\sigma_m^2 = \frac{1}{2}\exp\left(\arctan(\lambda(m))\right)$. We set $\lambda(0) = 0$. As for $\lambda(m)$, we considered two different schemes: alternatives with local positivity and alternatives with both local positivity and local negativity. These two schemes are to be specified in Sections 4.2.1 and 4.2.2 later.

First, consider two test statistics, one according to White (2000) and the other according to Hansen (2005):

$$T^{RC} = \sqrt{n} \max_{m \in \mathbf{M}} \hat{d}(m) \text{ and } T^{SPA} = \sqrt{n} \max_{m \in \mathbf{M}} \frac{\hat{d}(m)}{\hat{\omega}(m)}, \tag{11}$$

where $\hat{\omega}^2(m)$ is taken to be the sample variance of $\{L_{0,t} - L_{m,t}\}_{t=1}^n$.

To construct critical values, we generated the bootstrap version $\{\hat{d}_b^*(m)\}_{b=1}^B$, $b = 1, 2, \cdots, B$, of $\hat{d}(m)$ by resampling from observations $\{L_{0,t} - L_{m,t}\}_{t=1}^n$ with replacement, and let $\tilde{d}_b^*(m)$ be as

defined in (10). We constructed

$$\begin{aligned} \text{Reality Check Test (RC):} \quad & \text{Reject } H_0 \quad \text{if } T^{RC} > c_\alpha^{RC*} \text{ and} \\ \text{Superior Predictive Ability Test (SPA):} \quad & \text{Reject } H_0 \quad \text{if } T^{SPA} > c_\alpha^{SPA*}, \end{aligned}$$

where c_α^{RC*} is the $(1-\alpha)$-quantile of $\{T_b^{RC*}\}_{b=1}^B$, with $T_b^{RC*} = \sqrt{n}\max_{m\in\mathbf{M}} \tilde{d}_b^*(m)$, and c_α^{SPA*} is the $(1-\alpha)$-quantile of $\{T_b^{SPA*}\}_{b=1}^B$ with $T_b^{SPA*} = \max_{m\in\mathbf{M}} \sqrt{n}\bar{d}_b^*(m)/\hat{\omega}(m)$, where

$$\bar{d}_b^*(m) = \hat{d}_b^*(m) - \hat{d}(m) \times 1\{\sqrt{n}\hat{d}(m)/\hat{\omega}(m) \geq -\sqrt{2\ln\ln n}\}.$$

Note that $\bar{d}_b^*(m)$ involves centering of $\hat{d}_b^*(m)$ selectively depending on whether $\hat{d}(m)$ is close to the boundary of the inequalities or not. The selective recentering is done to improve the power of the test by weeding out the forecasting methods that perform badly.

For the hybrid test which is the main proposal of this paper, define first the complementary test statistic:

$$T^S = \sqrt{n}\min\left\{\max_{m\in\mathbf{M}} \frac{\hat{d}(m)}{\hat{\omega}(m)}, \max_{m\in\mathbf{M}} -\frac{\hat{d}(m)}{\hat{\omega}(m)}\right\}.$$

As for T^K, we take $T^K = T^{SPA}$ defined in (11). As for critical values, construct

$$T_b^{S*} = \sqrt{n}\min\left\{\max_{m\in\mathbf{M}} \frac{\tilde{d}_b^*(m)}{\hat{\omega}(m)}, \max_{m\in\mathbf{M}} -\frac{\tilde{d}_b^*(m)}{\hat{\omega}(m)}\right\}.$$

Let c_α^{S*} be the $(1-\alpha/2)$-quantile of $\{T_b^{S*}\}_{b=1}^B$, and take c_α^{K*} to be the $(1-\alpha/2)$-quantile of $\{T_b^{K*}\}_{b=1}^B$, where $T_b^{K*} = T_b^{SPA*}1\{T_b^{S*} \leq c_\alpha^{S*}\}$. Choosing c_α^{S*} and c_α^{K*} as such reflects the choice of $\gamma = 1/2$. Then, the hybrid test is defined as

$$\begin{aligned} \text{Hybrid Test (Hyb):} \quad & \text{Reject } H_0 \quad \text{if } T^S > c_\alpha^{S*} \text{ or} \\ & \qquad\qquad \text{if } T^S \leq c_\alpha^{S*} \text{ and } T^{SPA} > c_\alpha^{K*}. \end{aligned}$$

In the simulation studies, the sample size n was 200 and the number of Monte Carlo simulations

and the bootstrap Monte Carlo simulations 2,000. The number (M) of candidate forecasting methods was chosen from $\{50, 100\}$.

4.2 Alternative Hypotheses

4.2.1 Alternatives with Local Positivity

Following Hansen (2005), we first consider the following alternatives with $\lambda(m)$:

$$\text{DGP A: } \lambda(m) = \begin{cases} 0, & \text{if } m = 0 \\ \lambda(1), & \text{if } m = 1 \\ \rho \cdot (m-1)/(M-2), & \text{if } m = 2, \cdots, M, \end{cases}$$

where ρ and $-\lambda(1)$ were chosen from $\{0, 1, 2, 3, 4\}$. The $M-1$ methods with $m = 2, \cdots, M$ are inferior to the benchmark method ($m = 0$). Their relative performance is ordered as $M \prec M-1 \prec \cdots \prec 2$.

When $\lambda(1) = 0$, no alternative forecasting method strictly dominates the benchmark method, representing the null hypothesis. When $\lambda(1) < \lambda(0) = 0$, the method 1 performs better than the benchmark method, representing the alternative hypothesis. The magnitude ρ controls the extent to which the inequalities $\lambda(m) \geq \lambda(0) = 0$, $m = 2, \cdots, M$, lie away from binding. When $\rho = 0$, the remaining inequalities for methods 2 through M are binding, i.e., $d(m) = 0$ for all $m = 2, \cdots, M$.

Tables 1 and 2 show the empirical size of the tests under DGP A. The results show that the test RC has lower type I error as the design parameter ρ increases. For example, when $\rho = 2$, the rejection probability of the test RC is 0.0005 when the nominal size is 5% and $M = 50$. This extremely conservative size of the test RC is significantly improved by the test SPA of Hansen (2005) which shows the type I error of 0.0180. This improvement is made through two channels: the normalization by $\hat{\omega}(m)$ of the test statistic, and the trimming of poorly performing forecast methods. The hybrid approach shows a further improvement over the test SPA, yielding type I error of 0.0365 in this case.

Table 1: Empirical Size of Tests of Predictive Abilities under DGP A ($M = 50, n = 200$)

		$\alpha = .05$			$\alpha = .10$		
ρ	$\lambda(1)$	RC	SPA	Hyb	RC	SPA	Hyb
0	0	0.0520	0.0615	0.0620	0.1005	0.1180	0.1160
2	0	0.0005	0.0180	0.0365	0.0045	0.0330	0.0530
3	0	0.0005	0.0125	0.0230	0.0005	0.0210	0.0425
4	0	0.0000	0.0105	0.0220	0.0000	0.0230	0.0350

Table 2: Empirical Size of Tests of Predictive Abilities under DGP A ($M = 100, n = 200$)

		$\alpha = .05$			$\alpha = .10$		
ρ	$\lambda(1)$	RC	SPA	Hyb	RC	SPA	Hyb
0	0	0.0505	0.0710	0.0610	0.1060	0.1290	0.1210
2	0	0.0010	0.0125	0.0280	0.0040	0.0245	0.0485
3	0	0.0005	0.0080	0.0205	0.0015	0.0200	0.0355
4	0	0.0000	0.0120	0.0220	0.0015	0.0230	0.0315

Tables 3-4 show the power of the three tests. As for Hyb, the rejection probability is slightly lower than that of SPA in the case of $\rho = 0$. It is interesting to see that the rejection probability of Hyb is still better than RC when $\lambda(1) = -2, -3$ with $\rho = 0$. As the inequalities move farther away from binding while maintaining the violation of the null hypothesis (i.e. as ρ increases while $\lambda(1) < 0$), the performance of Hyb becomes prominently better than both RC and SPA.

To see how the power of Hyb can be better than that of SPA, recall that when the performance of SPA performs better than RC in finite samples, it is mainly because in computing critical values, SPA weeds out candidates that perform poorly. Given the same sample size n, the proportion of candidates weeded out tends to become larger as ρ increases. This explains the better performance of SPA over RC in Tables 3 and 4. When n increases so that $\sqrt{2 \ln \ln n}$ increases slowly yet $\sqrt{n}\hat{d}(m)/\hat{\omega}(m)$ is stable for many m's (as is the case with the simulation design with $\sqrt{n}$-converging Pitman local alternatives), the power-improvement by SPA is attenuated because there are many

Table 3: Empirical Power of Tests of Predictive Abilities under DGP A ($M = 50, n = 200$)

		$\alpha = .05$			$\alpha = .10$		
ρ	$\lambda(1)$	RC	SPA	Hyb	RC	SPA	Hyb
0	-2	0.1315	0.2995	0.2830	0.2180	0.4200	0.3950
	-3	0.4895	0.7785	0.7435	0.6210	0.8450	0.8265
2	-2	0.0115	0.2945	0.3770	0.0255	0.3810	0.4520
	-3	0.1135	0.7855	0.8395	0.2185	0.8490	0.8890
3	-2	0.0025	0.3055	0.4085	0.0095	0.3900	0.4690
	-3	0.0655	0.7745	0.8390	0.1360	0.8345	0.8770
4	-2	0.0030	0.3295	0.4125	0.0085	0.4185	0.4700
	-3	0.0475	0.8035	0.8625	0.0935	0.8655	0.8925

Table 4: Empirical Power of Tests of Predictive Abilities under DGP A ($M = 100, n = 200$)

		$\alpha = .05$			$\alpha = .10$		
ρ	$\lambda(1)$	RC	SPA	Hyb	RC	SPA	Hyb
0	-2	0.0945	0.2675	0.2475	0.1805	0.3765	0.3535
	-3	0.3945	0.7080	0.6755	0.5270	0.7885	0.7755
2	-2	0.0025	0.2315	0.3185	0.0115	0.3040	0.3830
	-3	0.0620	0.7040	0.7750	0.1230	0.7710	0.8220
3	-2	0.0000	0.2345	0.3245	0.0030	0.3110	0.3810
	-3	0.0350	0.7115	0.7865	0.0725	0.7770	0.8230
4	-2	0.0005	0.2385	0.3125	0.0035	0.3155	0.3615
	-3	0.0190	0.7220	0.7875	0.0435	0.7925	0.8250

methods that perform bad yet survive the truncation. In this situation, the power-improving effect of coupling by Hyb is still in force, because power reduction due to many bad forecasting methods that survive the truncation in SPA continues to be counteracted by the complementary test coupled in Hyb.

4.2.2 Alternatives with Local Positivity and Local Negativity

The hybrid test was shown to perform well relative to the other two tests under DGP A. However, DGP A mainly focuses on alternatives such that RC tends to have weak power. In this section, we

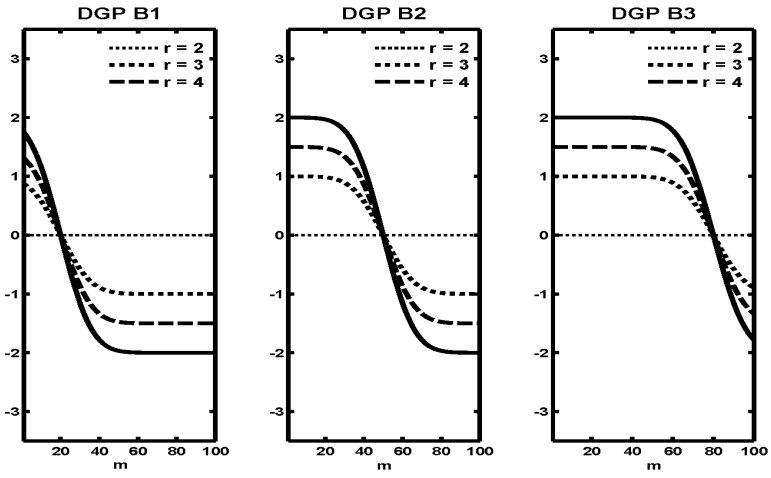

Figure 3: Three Designs of $\lambda(m)$ with $M = 100$. All the three designs represent different types of alternative hypotheses. In DGP B1, the forecasting methods with $m = 1$ through $m = 20$ outpeforms the benchmark method and in DGP B3, the forecasting methods with $m = 1$ through $m = 80$ outperforms the benchmark method.

consider the following alternative scheme: for each $m = 1, \cdots, M$,

$$\begin{aligned}
&\text{DGP B1:} \qquad \lambda(m) = r \times \{\Phi(-8m/M + 1/5) - 1/2\}, \\
&\text{DGP B2:} \qquad \lambda(m) = r \times \{\Phi(-8m/M + 2/5) - 1/2\}, \text{ and} \\
&\text{DGP B3:} \qquad \lambda(m) = r \times \{\Phi(-8m/M + 4/5) - 1/2\},
\end{aligned}$$

where Φ is a standard normal distribution function and r is a positive constant running in an equal spaced grid in $[0, 5]$. This scheme is depicted in Figure 3. In DGP B1, only a small portion of methods perform better than the benchmark method, and in DGP B3, a large portion of methods perform better than the benchmark method. The general discussion of this paper predicts that the hybrid test has relatively strong power against the alternatives under DGP B1 while it has relatively weak power against the alternatives under DGP B3.

Only the results for the cases DGP B1 and DGP B3 are shown in Figure 4 to save space. Under

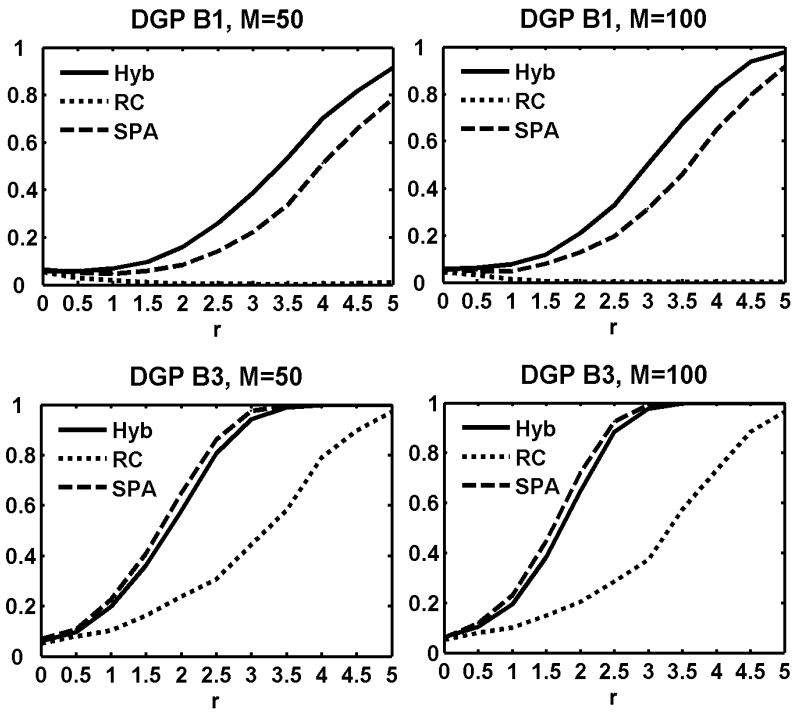

Figure 4: Finite Sample Rejection Probabilities for Testing Predictive Ability for Three Tests at Nominal Size 5%: Reality Check (RC) test of White (2000), Superior Predictive Ability (SPA) test of Hansen (2005), and Hybrid (Hyb) test of this paper against alternatives depicted in Figure 3. The result shows that Hyb performs conspicuously better than SPA and RC under DGP B1 that is generally associated with low power for tests, and it performs slightly worse than SPA under DGP B3 that is generally associated with high power for tests. Hence the result illustrates the robustified power behavior of the hybrid approach.

DGP B1, Hyb is shown to outperform the other tests. However, it shows a slight reduction in power (relative to SPA) under DGP B3. This result suggests that as long as the simulation designs used so far are concerned, the power gain by adopting the hybrid approach can be considerable under certain alternatives while its cost as a reduction in power under the other alternatives is only marginal.

5 Empirical Application: Reality Check Revisited

5.1 Testing Framework and Data

The empirical section of White (2000) investigates forecastability of excess returns using technical indicators. He demonstrated that unless the problem of data snooping is properly addressed, the best performing candidate forecasting method appears spuriously to perform better than the benchmark forecast based on a simple efficient market hypothesis. This section revisits his empirical study using recent S&P500 stock returns.

Similarly as in White (2000), this study considered 3,654 forecasts using technical indicators and adopted mean squared prediction error (MSPE) defined as follows: for $m = 1, \cdots, 3654$,

$$\text{MSPE} : \Lambda_{MSPE}(m) = \mathbf{E}\left[(Y_{m,T+1} - \hat{Y}_{m,T+1})^2\right],$$

where Y_T denotes the S&P500 return on day T, and $\hat{Y}_{T+1}$ its one day ahead forecast. (Given stock price P_t at t, the stock return is defined to be $Y_t = (P_t - P_{t-1})/P_{t-1}$.) See White (2000) for details about the construction of the forecasts.

S&P500 Stock Index closing prices were obtained from the Wharton Research Data Services (WRDS). The stock index returns data range from March 28, 2003 to July, 1, 2008. For each forecast method, we obtain 187 one-day head forecasts from October 4, 2007 to July 1, 2008. The data used for the estimation of the forecast models begin from the stock index return on March 28, 2003, and the sample size for estimation is 1,138. Hence the sample size for estimation is much larger than the sample used to produce forecasts, and it is expected that the normalized sum of the forecast error differences will be approximately normally distributed. (See Clark and McCracken (2001) for details.)

Table 5: Bootstrap p-Values for Testing Predictive Ability in terms of MSPE. The number q represents the tuning parameter that determines the random block sizes in the stationary bootstrap. See Politis and Romano (1994) and White (2000) for details.

	q	RC	SPA	Hyb
Data Snooping	0.10	0.1910	0.1598	0.0670
Taken into Account	0.25	0.2916	0.2736	0.0980
	0.50	0.3378	0.3206	0.1250
Data Snooping	0.10	0.0094	0.0094	0.0200
Ignored	0.25	0.0188	0.0188	0.0390
	0.50	0.0384	0.0384	0.0780

5.2 Results

The results are shown in Table 5. The number q in the tables represents the tuning parameter that determines the random block sizes in the stationary bootstrap of Politis and Romano (1994). (See White (2000) for details.) Note that in White (2000), $q = 0.5$ was used.

Table 5 presents p-values for the tests with data snooping taken into account, and for the tests with data snooping ignored. When data snooping is ignored, all the tests spuriously reject the null hypothesis at 10%. (Note that the results of RC and SPA are identical, because there is only one candidate forecast when data snooping is ignored, and this forecast is not trimmed out by the truncation involved in SPA.) This attests to one of the main messages of White (2000) that without proper consideration of data snooping, the best performing candidate forecasting method will appear to highly outperform the benchmark method.

Interestingly, the p-values from Hyb are conspicuously lower than those obtained from RC and SPA. For example, when $q = 0.10$, the p-values for RC and SPA are 0.1910 and 0.1598, but the p-value for Hyb is 0.0670. Hence the null hypothesis is rejected by Hyb while not by RC and SPA at 10% in this case. This illustrates the distinctive power behavior of Hyb. Note that Hyb can outperform RC even when SPA does not outperform it. Such a case may arise when for all the m's $\sqrt{n}\hat{d}(m)/\hat{\omega}(m)$ is greater than $-\sqrt{2\ln\ln n}$ with large probability, but for some m, $\sqrt{n}\hat{d}(m)/\hat{\omega}(m)$

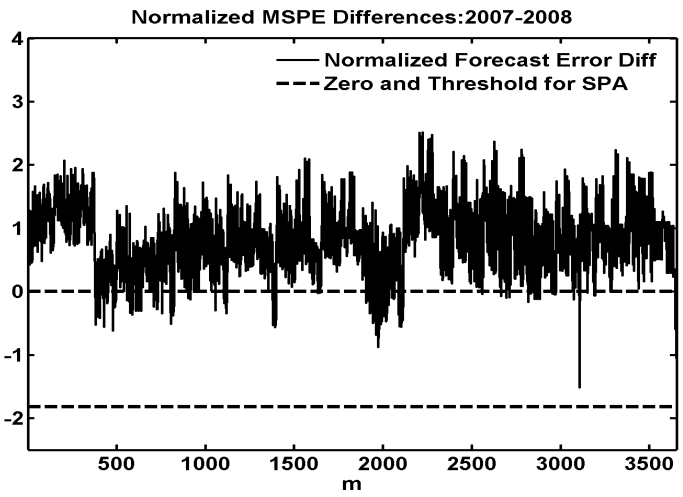

Figure 5: Normalized Estimated Mean Forecast Error Differences in terms of MSPE (i.e., $\sqrt{n}\hat{e}(m)/\hat{\omega}(m)$). The threshold is $-\sqrt{2\ln\ln n}$. The x-axis is the forecasting method index m running from 1 to 3654. The forecast error differences lie above the threshold value (represented by the lower dashed line). This means that no forecast method was truncated by SPA in this case. Therefore, the difference between RC and SPA is solely due to the normalization by $\hat{\omega}(m)$ in SPA. Even in this case, Hyb can still improve the power of the sup test through the power-enhancing effect from the complementary test.

still tends to take a fairly negative value relative to the critical value c_α of RC.

For example, see Figure 5 that plots the normalized MSPE differences, i.e., $\sqrt{n}\hat{d}(m)/\hat{\omega}(m)$, for $m = 1, \cdots, 3654$. It is interesting to see that no forecasting method is truncated by the truncation scheme of SPA. This is indicated by the fact that all the normalized forecast error differences are above the threshold value (lower dashed line). This perhaps explains similar p-values for RC and SPA. The difference between the results from RC and SPA is solely due to the fact that SPA involves normalization by $\hat{\omega}(m)$ while RC does not. Even in this case, Hyb continues to counteract its power-reducing effect by coupling with the complementary test statistic.

6 Closing Remarks

This paper has shown that the one-sided sup tests of predictive ability can be severely asymptotically biased in a general set-up. To alleviate this problem, this paper proposes the approach of hybrid tests where we couple the one-sided sup test with a symmetrized complementary test. Through simulations, it is shown that this approach yields a test with robust power behavior. The hybrid approach can be applied to numerous other tests of inequalities beyond predictive ability tests. The question of which modification or extension is suitable often depends on the context of application.

7 Acknowledgement

I thank Werner Ploberger and Frank Schorfheide for valuable comments and advice. Part of this paper began with a draft titled, "Testing Distributional Inequalities and Asymptotic Bias." The comments of Co-Editor, Associate Editor, and a referee were valuable and helped improving the paper substantially. I thank them for their comments.

References

[1] Amisano G. and R. Giacomini (2007), "Comparing density forecasts via weighted likelihood ratio tests," *Journal of Business and Economic Statistics*, 25, 177-190.

[2] Andrews, D. W. K. (2011), "Similar-on-the-boundary tests for moment inequalities exist but have poor power," CFDP 1815.

[3] Andrews, D. W. K. and X. Shi (2010), "Inference based on conditional moment inequalities," CFDP 1761.

[4] Andrews, D. W. K. and G. Soares (2007), "Inference for parameters defined by moment inequalities using generalized moment selection," CFDP 1631.

[5] Bao, Y., T-H., Lee, and B. Saltoğlu (2007), "Comparing density forecast models," *Journal of Forecasting*, 26, 203-225.

[6] Barret, G. F. and S. G. Donald (2003), "Consistent tests for stochastic dominance," *Econometrica*, 71, 71-104.

[7] Bugni, F. (2010), "Bootstrap inference in partially identified models defined by moment inequalities: coverage of the identified set," *Econometrica*, 78, 735-753.

[8] Canay, I. A. (2010), "EL inference for partially identified models: large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156, 408-425.

[9] Christoffersen, P. F. (1998), "Evaluating interval forecasts," *International Economic Review*, 39, 841-861.

[10] Clark, T, E., and M. W. McCracken (2001), "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105, 85-110.

[11] Diebold, F. X., T. A. Gunther, and A. S. Tay (1999), "Evaluating density forecasts with applications to financial risk management," *International Economic Review*, 39, 863-883.

[12] Diebold, F. X., J. Hahn, and A. S. Tay (1999), "Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange," *Review of Economics and Statistics*, 81, 661-673.

[13] Diebold, F. X. and R. S. Mariano (1995), "Comparing predictive accuracy," *Journal of Business, and Economic Statistics*, 13, 253-263.

[14] Giacomini, R. and H. White (2006), "Tests of conditional predictive ability," *Econometrica*, 74, 1545-1578.

[15] Hansen, P. R. (2005), "Testing superior predictive ability," *Journal of Business and Economic Statistics*, 23, 365-379.

[16] Lee S., K. Song, and Y-J. Whang (2011), "Testing functional inequalities," Cemmap Working Paper, CWP12/11.

[17] Linton, O., E. Maasoumi, and Y-J. Whang (2005), "Consistent testing for stochastic dominance under general sampling schemes," *Review of Economic Studies* 72, 735-765.

[18] Linton, O., K. Song, and Y-J. Whang (2009), "An improved bootstrap test of stochastic dominance," *Journal of Econometrics* 154, 186-202.

[19] Politis, D. N. and J. P. Romano (1994), "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.

[20] Sullivan, R., A. Timmermann, and H. White (1998), "Data snooping, technical trading rule performance, and the bootstrap," *Journal of Finance*, 54, 1647 1692.

[21] Stinchcombe, M. B. and H. White (1998): "Consistent specification testing when the nuisance parameters present only under the alternative," *Econometric Theory*, 14, 295-325.

[22] West, K. D. (1996), "Asymptotic inference about predictive ability," *Econometrica*, 64, 1067-1084.

[23] West, K. D. (2006), "Forecast evaluation," *Handbook of Economic Forecasting,* Chapter 3, eds. by G. Elliot, C.W.J. Granger, and A. Timmermann, North-Holland.

[24] White, H. (2000), "A reality check for data snooping," *Econometrica*, 68, 1097-1126.