A New Test for Superior Predictive Ability

Zongwu Cai[a,b], Jiancheng Jiang[a] and Jingshuang Zhang[a]

[a]Department of Mathematics & Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA. E-mail addresses: zcai@uncc.edu (Z. Cai), jjiang1@uncc.edu (J. Jiang), & zhang.jshuang@gmail.com (J. Zhang).

[b]Wang Yanan Institute for Studies in Economics, MOE Key Laboratory of Econometrics, and Fujian Key Laboratory of Statistical Sciences, Xiamen University, Xiamen, Fujian 361005, China.

This Version: September 20, 2010

This paper provides a new methodology to test the superior predictive ability (SPA) of technical trading rules relative to the benchmark without potential data snooping bias. Unlike other previous methods, we explicitly approximate the covariance matrix through certain decomposition, which decreases the number of elements needed to be estimated. With the help of covariance matrix, we are able to exploit more information contained in the diagonal and off-diagonal terms and as a result, so that we improve the effectiveness of testing result. Due to the nuisance parameter in composite hypothesis, we choose the generalized likelihood ratio (GLR) test which is of uniform most power, to alleviate such problem and at the same time, to provide a pivotal distribution. Bootstrap procedure is employed in our simulation to obtain the power of the test. The result shows that the GLR test dominates the SPA test proposed by Hansen (2005) in terms of power and our GLR test is sensitive to the inclusion of superior models. Therefore, it increases the power faster than that of SPA test. The result also suggests that the GLR test is less conservative than SPA test.

Keywords: Covariance matrix estimation; Data snooping; Generalized likelihood Ratio test; Reality check; SPA test; Technical trading rules.

1 Introduction

Data snooping is practically unavoidable, especially in various applied fields such as finance and economics, in which only a single history of interest is available for analysis, such as stock price, interest rate, etc. A so-called "good" forecasting model with observed superior performance obtained under several specification searches is highly possible to come from pure luck instead of genuinely forecasting ability. White (2000) pointed out that "even when no exploitable forecasting relation exists, looking long enough and hard enough at a given set of data will often reveal one or more forecasting models that look good, but are in fact useless."

There is another example. Sullivan (1999) addressed a point that "Data snooping can result from a subtle survivorship bias operating on the entire universe of technical trading rules that have been considered historically. Suppose that, over time, investors have experimented with technical trading rules drawn from a very wide universe-in principle thousands of parameterizations of a variety of types of rules. As time progresses, the rules that happen to perform well historically receive more attention and are considered *serious contenders* by the investment community, and unsuccessful trading rules are more likely to be forgotten. If enough trading rules are considered over time, some rules are bound by pure luck, even in a very large sample, to produce superior performance even if they do not genuinely possess predictive power over asset returns. Of course, inference based solely on the subset of surviving trading rules may be misleading in this context because it does not account for the full set of initial trading rules, most of which are likely to have under-performed."

White (2000) looked at data snooping from the angle of data mining and pointed out that data snooping is equivalent to data mining, which is to extract valuable relationships from masses of messed data. The negative connotation, however, of data mining is from the ease with which naive practitioners may mistake the spurious for the substantive, which is familiar to econometricians and statisticians. Leamer (1978, 1983) was a leader in pointing out these dangers, proposing methods for evaluating the fragility of the relationships obtained by data mining.

Another concept in this field is the superior predictive ability (SPA). In general, SPA means there is a particular forecasting procedure that is capable of outperforming other alternatives. When testing for SPA, the question of interest is whether any alternative forecast is better than the benchmark forecast or, equivalently, whether the best alternative forecasting model is better than the benchmark. This question can be addressed by testing the null hypothesis that "the benchmark is not inferior to any alternative forecast." Diebold and Mariano (1995) and West (1996) proposed the tests for equal predictive ability (EPA), which means the forecasting ability of a model is the same as the benchmark. The framework of West (1996) can accommodate the situation where forecasts involve estimated parameters. White (2000) was the pioneer to formulate the null hypothesis of superior predictive ability and proposed the reality check (RC) test which takes into account the dependence of individual statistics, whereas Sullivan, Timmermann, and White (1999) applied the RC test to technical trading rules and found that they lose their predictive power for major U.S. stock indices after the mid 1980's. Later, Romano and Wolf (2005) introduced a RC-based stepwise test, hence, step-RC test, that is capable of identifying as many significant models as possible. Commenting on the framework of White (2000), Hansen (2003) suggested a new testing procedure for composite hypotheses incorporating additional sample information from nuisance parameter and similarity condition which is necessary for a test to be unbiased. Later, Hansen (2005) provided a test for SPA (known as SPA test) that invokes a sample-dependent null distribution to avoid the least favorable configuration. Recently, Hsu, Hsu and Kuan (2010) extended the SPA test to a stepwise SPA test that can identify predictive models in large-scale, multiple testing problems without data snooping bias. They employed the SPA test to find that technical rules have significant predictive ability prior to the inception of exchange traded funds (ETF) in U.S. growth markets.

Indeed, SPA is often more relevant for economic applications than EPA, because the existence of a better forecasting model is typically of more importance than the existence of a worse. For example, a fund manager is interested in whether the forecasting model in use is inferior to other models. The distinction between EPA and SPA is substantive. The former involves a simple null hypothesis while the latter leads to a composite hypothesis. Hansen (2003, 2005) pointed out that the main complication in composite hypotheses testing is that (asymptotic) distribution typically depends on nuisance parameters, such that the null distribution is not unique. Currently, there are two methodologies to make inference for superior predictive ability. They are RC test and SPA test respectively. The RC test handles with the ambiguity of null distribution by using the least favorable configuration (LFC), which is referred to as the point least favorable to the alternative. In turn, the LFC-based

RC test is accompanied with three problems. First, the test is much conservative. Second, the test is sensitive to the inclusion of poor models. The more poor models included, the less the power of RC until it is driven to 0. Last, RC test is biased. In contrast, SPA test alleviates above problems by studentizing the test statistic and by invoking a sample-dependent null distribution. The latter is based on a novel procedure that incorporates additional sample information to identify the *relevant* alternatives.

We conduct the superior predictive ability test under the null hypothesis proposed by White (2000) and Hansen (2003, 2005). That is the benchmark performs no inferior to any alternative models. Our paper makes contribution to the literature in four ways. First, no matter the RC test or SPA test, both employ a bootstrap procedure to circumvent an explicit estimation of a large covariance matrix. In our framework, a covariance matrix of error terms in factor model is introduced. It is approximated by a particular decomposition method that is partly similar to singular value decomposition (SVD), different from which background noise or systematic noise is considered and is able to be separated under our method. The approximation of covariance matrix is also applicable to the case when forecasting models exceeds the sample size, even in a large-scale. However, this situation is deemed to be infeasible by White (2000) and Hansen (2003, 2005).

Secondly, in SPA test, only the diagonal elements of the covariance matrix are used. In this paper, nearly all components in the matrix are utilized to obtain the knowledge in the matrix. It implies we take advantage of much more information including relationship among different models to make the test more powerful. The matrix consists of two types of information. The first one is real economic news which gives the performance of trading rules and is mainly used to gauge whether the predictive model is superior or not. The second one represents the background noise level, which will be separated from the real economic factors. The covariance is incorporated into our analysis through the error term in the representation of the so-called generalized likelihood ratio (GLR) test proposed in this paper. Indeed, Hansen (2003, 2005) pointed out that his SPA test may be improved if there is a reliable way to incorporate information about the off-diagonal elements .

Thirdly, as Hansen (2005) suggested that the testing problem of composite hypothesis is closely related to the problem of testing hypotheses in the presence of nuisance parameter, this paper utilizes generalized likelihood ratio test, which is of uniform most power and independent of nuisance parameter due to Wilks' phenomenon. The GLR test statistic follows distribution with certain degree of freedom. Further, this test statistic is asymptotically optimal in the sense that it achieves optimal rate of convergence. Under this test, the result is bound to be more persuasive.

Lastly, this paper details the bootstrap implementation step-by-step. We conduct the bootstrap in a way different from traditional bootstrap method since our bootstrap null distribution is generated under different samples while traditional bootstrap only involves only one sample. These samples follow the same data generating process (the same input parameter and from the same distribution). As a combination of results from different samples, it will be more representative and of generality. The main argument for our bootstrap procedure is to make the distribution more exact.

The rest of this paper is organized as follows. We review the existing tests in Section 2. In Section 3, a method for our test is described in detail and the detailed bootstrap implementation is contained. Section 4 includes a simulation to study the effectiveness of the proposed method and compares it with SPA. Section 5 gives a concluding remark.

2 Review of Existing Tests

2.1 Reality Check Test

In the framework of White (2000), the null hypothesis is set to express no predictive superiority over a benchmark and it can be expressed as follows:

$$H_0^k : E(d_{k,t}) \leq 0, \tag{2.1}$$

where $d_{k,t}$ ($k = 1, 2, \cdots, m$ and $t = 1, 2, \cdots, n$) denote their performance measure relative to a benchmark model over time. For each k, $E(d_{k,t}) = \mu_k$ for all t, and for each t, $d_{k,t}$ may be dependent across k. Data snooping arises when the inference for the null is drawn from the test of an individual hypothesis H_0^k. White (2000) circumvented the problem by invoking the RC test

$$\mathrm{RC}_n = \max_{1 \leq k \leq m} \sqrt{n}\bar{d}_k, \tag{2.2}$$

where $\bar{d}_k$ is the k-th element of $\bar{d}$ and $\bar{d} = \sum_{t=1}^n d_t/n$. The least favorable configuration (LFC) is that $\mu = 0$ is chosen to obtain the null distribution. Under some mild assumptions (see, Hansen (2005) and Hsu, Hsu and Kuan (2010) for details), the data obey a central limit theorem:

$$\sqrt{n}[\bar{d} - \mu] \ \to^{d} \ N(0, \Omega), \tag{2.3}$$

where $\Omega = \lim_{n\to\infty} \mathrm{Var}[\sqrt{n}(\bar{d} - E(d_t))] = (\omega_{ij})_{m\times m}$. The limiting distribution of RC_n under the null hypothesis is $\max_{1\leq k\leq m}\{N(0, \omega_{kk})\}$, which can be approximated via a bootstrap procedure. The null hypothesis is rejected when the bootstrap p-value is smaller than prespecified significance level. While the LFC is convenient to implement, the RC test also bears a few drawbacks. As Hansen (2003, 2005) pointed out, because it is a LFC-based test and the individual model statistics are non-standardized, the RC suffers two major drabacks: The first is that it is sensitive to the inclusion of poor and irrelevant models in the space of competing forecasting models. Since only binding constraints ($\mu = 0$) matter for the asymptotic distribution, the inclusion of poor model decreases the power of the test by increasing RC's p-value, which is based on $\max(\sqrt{n}\bar{d})$. The other one is that the power of the RC is unnecessarily low in most situations. In other words, it is relatively conservative whenever the number of binding constraints are small relative to the number of inequalities being tested.

2.2 Superior Predictive Ability Test

Under the same null hypothesis as in RC test, Hansen (2005) proposed a studentized test

$$\mathrm{SPA}_n = \max\Big[\max_{1\leq k\leq m} \sqrt{n}\bar{d}_k/\hat{\sigma}_k, \ 0\Big], \tag{2.4}$$

where $\hat{\sigma}_k^2$ is a consistent estimator of $\sigma_k^2 = \omega_{kk}$. The main argument for the normalization is it will improve the power typically. However, it uses a data-dependent choice for μ instead of $\mu = 0$ implied by the LFC condition, which leads to a more powerful tests of composite hypothesis. The intuition of this method comes from the logarithm. Therefore, A proper test should reduce the influence of the poor models while preserving the influence of the models with $\mu_k = 0$. It may be tamping to simply exclude the alternative with $\bar{d}_k < 0$ from the analysis. But this approach does not lead to valid inference in general, because the models that are (or appear to be) a little worse than the benchmark can have a substantial influence on the distribution of the test statistic in finite samples. Therefore, based on the above discussions, we can construct our test in a way that incorporates all models, while reducing the influence of alternatives that the data suggests are poor.

While LFC-based RC test takes a supremum over the null hypothesis, the SPA test takes the supremum over a smaller confidence set chosen such that it contains the true parameter with a probability that converges to 1. In the SPA test, the estimator of $E(d_k) = \mu_k$ is suggested as

$$\hat{\mu}_k = \bar{d}_k \cdot 1\{\sqrt{n}\bar{d}_k/\hat{\sigma}_k \leq -\sqrt{2\log\log n}\}, \quad k = 1, 2, \cdots, m, \tag{2.5}$$

where $1\{\cdot\}$ denotes the indicator function. It can be seen that $\hat{\mu}_k = 0$ almost surely when $\mu_k = 0$. Moreover, if $\mu_k < 0$, $\sqrt{n}\bar{d}_k/\hat{\sigma}_k$ is smaller than the threshold rate $-\sqrt{2\log\log n}$ for sufficient large n, such that $\hat{\mu}_k \ll 0$ almost surely, where $x \ll y$ means that x is much smaller than y. Notice that the choice of the threshold value to be $-\sqrt{2\log\log n}$ is based on the strong law of large number. This estimator is used to well separate the poor trading models with $\mu_k < 0$ and models with mean zero and a little worse than zero since a poor model, $\mu_k < 0$, has an impact on the critical value whenever $\sqrt{n}\bar{d}_k/\hat{\omega}_k$ is only moderate negative, say between -1 and 0, and can not be simply omitted from analysis, especially in finite sample. In view of this, the LFC-based RC test is improved because there is sufficient information to determine exactly which inequalities are non-binding but still can be used to derive the null distribution.

3 Methodology to Test Superior Predictive Ability

3.1 Hypothesis of Interest

This question of interest can be addressed by testing the null hypothesis that the benchmark is not inferior to any alternative forecast. This objective can be interpreted as follows:

I. Performance of the kth trading strategy is measured by loss function relative to that of benchmark, instead of its absolute value, given by

$$d_{k,t} = L(\xi_t, \delta_{0,t-h}) - L(\xi_t, \delta_{k,t-h}), \quad k = 1, 2, \cdots, m, \tag{3.1}$$

where $L(\cdot, \cdot)$ is a loss function. The loss function is a function of two variables, i.e. $L(\xi_t, \delta_{k,t-h})$, $k = 1, 2, \cdots, m$, where ξ_t is a random variable that represents the aspects of the decision problem that are unknown at the time that the decision is made, and $\delta_{k,t-h}$ represents a possible decision rule which is made h periods in advance. If $k = 0$, $\delta_{0,t-h}$ is the decision made according to the benchmark trading strategy. Hansen (2005) gave an example, in which $\delta_{k,t-1}$ is assigned the value of -1 when a trader takes a short position, and the value of 1 if he/she takes a long position in an asset at time $t-1$. ξ_t is the return of asset in period t, i.e., $\xi_t = r_t$. The kth trading rule yields the profit $\pi_{k,t} = \delta_{k,t-h} r_t$. The loss function can be formulated as $L(\xi_t, \delta_{k,t-h}) = -\delta_{k,t-1}\xi_t$. We evaluate forecasts in terms of their expected loss, such as

$$E(d_k) = E[L(\xi_t, \delta_{0,t-h})] - E[L(\xi_t, \delta_{k,t-h})], \quad k = 1, 2, \cdots, m.$$

Therefore, we focus on d_k exclusively rather than the loss function itself.

II. The benchmark is the target to compare with. It is reflected in d_k as the performance of kth trading rule is net of that of a benchmark. For a fund manager who wants to know whether the performance of his portfolio beats the market, the benchmark can be the market rate of return. For a trader in above example, if $\delta_{0,t}$ is set to equal to 1 over time, then it is a buy and hold strategy. This benchmark is used by Sullivan et al. (1999, 2001).

III. The null hypothesis is postulated as follows: $H_0 : \mu \leq 0$, where $E(d_{k,t}) = \mu_k$, $d_t = (d_{1t}, d_{2t}, \cdots, d_{mt})'$ and $\mu = (\mu_1, \mu_2, \cdots, \mu_m)'$.

IV. Taking advantage of Hansen's estimator of from null hypothesis, we have

$$\hat{\mu}_k = \bar{d}_k \cdot 1\{\sqrt{n}\bar{d}_k/\hat{\sigma}_k \leq -\sqrt{2\log\log n}\}, \quad k = 1, 2, \cdots, m$$

where $\hat{\sigma}_k = \hat{\omega}_{kk}$ are the diagonal elements of covariance matrix Ω, which will be defined later.

3.2 Estimation Procedure

Indeed, the aforementioned work is in the spirit of the idea in Hansen (2005). Our first innovation is the introduction of factor model to give a clear expression to the approximation of covariance matrix.

$$d_t = \mu + e_t = \mu + \Omega^{1/2}\varepsilon_t \quad t = 1, 2, \cdots, n, \tag{3.2}$$

where $\Omega = [\omega_{ij}]_{m\times m}$ is a definite positive covariance matrix and

$$\varepsilon_t \sim iid \;\; N(0, I_m),$$

where I_m is the $m \times m$ identity matrix. After the decomposition, d_t is able to be expressed by two parts. The first part is the mean value of the trading rule during certain period, and the second one, the error term, is the systematic noise, which can not be explained by the mean. Another merit in the expression of d_t is from the convenience to explicitly analyze the covariance matrix after separation, which contains important information. With the covariance matrix, GLR test can be employed to explore d_t and obtain the null distribution independent of nuisance parameter.

In many applications, there are a ton of trading rules to be investigated so that m might be huge. For example, Sullivan et al. (1999) evaluated $7,846$ technical trading rules, and Hsu, Hsu, Kuan (2010) employed a total of $16,380$ rules. This means a sensible estimate of all elements of Ω is nearly infeasible, especially when competing trading strategies m exceeds the sample size n. Instead, we approximate the estimation of Ω using its most useful or important information, which is also in spirit of principle component analysis (PCA). The method to estimate covariance matrix is similar to singular value decomposition (SVD) which is a common technique for analysis of multivariate data without a systematic noise term. Since only the most important information of Ω is needed to be reported through estimation, the amount of elements in covariance matrix can decrease sharply.

We suppose the covariance matrix admits the following decomposition

$$\Omega = QDQ^T, \tag{3.3}$$

where $Q = (q_1, \cdots, q_m)$ is a $m\times m$ orthogonal matrix with $\{q_i\}_{i=1,\cdots,m}$ forming an orthogonal basis, so that $q_i^T q_j = 1$ for $i = j$, and $q_i^T q_j = 0$ otherwise, and matrix D consists of two components $S_{m\times m}$ and $N_{m\times m}$. We separate D into two parts in order to decrease the amount of values needed to be estimated, to simplify the estimation of covariance matrix. On the other hand, the two parts admit economic explanation, which an accommodate more applications in the real world. Here, $S_{m\times m}$, determined by real economical factors, gathers the most important information specific to each trading rules. The elements of S are only nonzero on the diagonal. Thus, $S = \text{diag}(s_1, \cdots, s_m)$. Furthermore, $s_k > 0$ for $1 \leq k \leq d$

and $s_k = 0$ for $(d+1) \leq k \leq m$. By convention, the ordering of s_k is determined by high-to-low sorting of its values, with the highest value in the upper left index of the S matrix. The other part, $N_{m\times m}$, is known as background noise or systematic noise. It represents the variance that is shared by all the components in Ω. It is also a diagonal matrix and additionally, all elements along diagonal have the same value denoted by δ^2, representing the background noise level. Specifically,

$$D_{m\times m} = S_{m\times m} + N_{m\times m} = \text{diag}\{\lambda_1, \cdots, \lambda_d, 0, \cdots, 0\},$$

where $\lambda_j = s_j + \delta^2$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$ are decided by real economic factors.

In general, under the null and alternatives, μ_k can be estimated by $\bar{d}_k$ which is in the form of

$$\bar{d}_k = n^{-1} \sum_{t=1}^{n} d_{k,t} \quad t = 1, 2, \cdots, n.$$

Hence, the residual from (3.2) is estimated by

$$\hat{e}_t = \hat{\Omega}^{1/2}\hat{\varepsilon}_t = d_t - \bar{d}_t \quad t = 1, 2, \cdots, n$$

Using $\{\hat{e}_t\}_{m\times 1}$, the sample covariance matrix is estimated by

$$\hat{\Omega} = \begin{pmatrix} \hat{\omega}_{11} & \hat{\omega}_{12} & \cdots & \hat{\omega}_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\omega}_{m1} & \hat{\omega}_{m2} & \cdots & \hat{\omega}_{mm} \end{pmatrix} = \frac{1}{n-1} \Sigma_{t=1}^{n} \hat{e}_t \hat{e}_t', \tag{3.4}$$

where $\hat{\omega}_{ij}$ denotes the sample version of covariance between e_i and e_j . The background noise factor δ^2 is defined as the total variance of all elements in the matrix $d = \{d_{kt}\}_{m\times n}$ relative to their corresponding sample mean. The column vector of d is the performance of all trading rules at time t when the kth row vector of d is the performance measure of kth trading strategies over time. Mathematically, $\delta^2 = \text{Var}(d_{ij} - \mu_i) \quad i = 1, \cdots, m$ and $j = 1, \cdots, n$. From covariance matrix $\hat{\Omega}$, we get its associated eigenvectors $\{v_i\}_{i=1,\cdots,m}$ and eigenvalues $\{\lambda_i^*\}_{i=1,\cdots,m}$, where $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_m^*$. $\{v_i\}_{i=1,\cdots,m}$ satisfies the sufficient conditions of a orthogonal matrix and is the column vector of $\hat{Q} = (v_1, v_2, \cdots, v_m)$. Eigenvalues $\{\lambda_i^*\}_{i=1,\cdots,m}$ are used to determine matrix $\hat{D}$ according to the following rule:

$$\hat{\lambda}_j = \begin{cases} \lambda_j^* - \hat{\delta}^2, & \text{if } \lambda_j^* \geq \hat{\delta}^2; \\ 0, & \text{if } \lambda_j^* < \hat{\delta}^2 \end{cases} = (\lambda_j^* - \hat{\delta}^2)\, 1\{\lambda_j^* \geq \hat{\delta}^2\}, \qquad j = 1, 2, \cdots, m.$$

Then, the estimated covariance matrix $\hat{\Omega}$ is approximated by $\hat{\Omega}^* = \hat{Q}\hat{D}\hat{Q}^T$, where

$$\hat{D} = \mathrm{diag}\{\hat{\lambda}_1, \ddots, \hat{\lambda}_d, 0, \ddots, 0\} + \hat{\delta}^2 \, I_m.$$

3.3 Generalized Likelihood Ratio Test

The reason of using the generalized likelihood ratio test proposed by Cai, Fan and Yao (2000) is due to its great properties such as easy implementation and uniform most powerful test as well as the so-called Wilks phenomenon; see Fan, Zhang and Zhang (2001) and Fan, Jiang (2007) for details on these aspects. Note that the GLR test is also called the generalized F-test in Cai and Tiwari (2000). The existence of Wilks phenomenon in GLR test makes finite sample simulation feasible in determining the null distributions of the test statistics. Define the residual sum of squares under the null and alternative as follows:

$$\mathrm{RSS}_0 = \sum_{k=1}^{m}\sum_{k=1}^{n}\Big[\hat{\Omega}^{*-\frac{1}{2}}(d_{kt} - \hat{\mu}_k)\Big]^2 = \sum_{k=1}^{m}\sum_{k=1}^{n}\Big[\hat{\Omega}^{*-\frac{1}{2}}\hat{e}_{kt}\Big]^2 = \sum_{k=1}^{m}\sum_{k=1}^{n}\hat{\varepsilon}_{kt}^2,$$

where $\hat{\mu_k} = \bar{d_k} \; \{\sqrt{n}\bar{d}_k/\hat{\sigma}_k \leq -\sqrt{2\log\log n}\}$, $k = 1, 2, \cdots, m$, and

$$\mathrm{RSS}_1 = \sum_{k=1}^{m}\sum_{k=1}^{n}\Big[\hat{\Omega}^{*-\frac{1}{2}}(d_{kt} - \bar{d_k})\Big]^2 = \sum_{k=1}^{m}\sum_{k=1}^{n}\Big[\hat{\Omega}^{*-\frac{1}{2}}\hat{e}_{kt}\Big]^2 = \sum_{k=1}^{m}\sum_{k=1}^{n}\hat{\varepsilon}_{kt}^2,$$

respectively. Then the GLR test statistic is given by

$$T_n = \frac{mn}{2}(\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1. \tag{3.5}$$

We reject the null hypothesis for large T_n which might follow asymptotically a chi-square distribution with a large degree of freedom; see Cai, Fan and Yao (2000), Cai and Tiwari (2000), Fan, Zhang and Zhang (2001) and Fan, Jiang (2007) for details. It might not easy to derive the exact asymptotic distribution of T_n, which can be easily approximated by a Bootstrap approach, described in the next section.

3.4 Bootstrap Implementation

We now discuss step-by-step implementation of bootstrap procedure demonstrating its convenience. There are m trading rules operating on time from $t = 0$ to n. We suppose that the

$m \times n$ prediction observations are given. We also assume that a method for generating a collection of m model specifications has been specified. Next, we specify the times of bootstrap B for a single simulation, the number of simulations to the bootstrap is B_S and the times of replications is B_N to get the power of the GLR test. Then, we use the estimation model to get the estimates for vector $\{\hat{\mu}_{k0}\}_{k=1}^{m}$ under the null hypothesis and covariance matrix $\hat{\Omega}^{*-\frac{1}{2}}$ given the data. Generate the residuals from null by the equation

$$\hat{\varepsilon}_{kt0} = \hat{\Omega}^{*-\frac{1}{2}}(d_{kt} - \hat{\mu}_{k0}) \quad t = 1, 2, \cdots, n. \tag{3.6}$$

At this moment, from the observed sample points, we obtain the original GLR test statistic T_n based on $\{d_t, \hat{\varepsilon}_t, \hat{\varepsilon}_{t0}\}$ Further, we draw bootstrap residuals with size n from the empirical distribution of $\{\hat{\varepsilon}_{t0}\}_{t=1}^{n}$ selected under the same chance $1/n$ with replacement. Denote the new samples as

$$the\, b^{th}\, sample \equiv \{\hat{\varepsilon}_{t0}^{*(b)}\}_{t=1}^{n} \quad b = 1, 2, \cdots, B$$

and define the centered bootstrap residuals

$$\widetilde{\varepsilon}_{t0}^{*(b)} = \hat{\varepsilon}_{t0}^{*(b)} - \bar{\varepsilon}_{0}^{*(b)} \quad b = 1, 2, \cdots, B,$$

where $\bar{\varepsilon}_{0}^{*(b)} = \sum_{t=1}^{n} \hat{\varepsilon}_{t0}^{*(b)}/n$. Now, a new data set $\hat{d}_t^{*(b)}$ from the bth bootstrap is generated based on the sample $\{\hat{\mu}_0, \widetilde{\varepsilon}_{t0}^{*(b)}\}_{t=1}^{n}$

$$\hat{d}_t^{*(b)} = \hat{\mu}_0 + \hat{\Omega}^{*-\frac{1}{2}} \widetilde{\varepsilon}_{t0}^{*(b)} \quad t = 1, 2, \cdots, n.$$

In the use of the new sample, the GLR test statistic $\{T^{*(b)}\}_{b=1}^{B}$ is calculated

$$T_n^{*(b)} = \frac{mn}{2}(\mathrm{RSS}_0^{*(b)} - \mathrm{RSS}_1^{*(b)})/\mathrm{RSS}_1^{*(b)}.$$

Repeat the bootstrap procedure for B_S times and stack all the values of GLR test into vector $T_n^* = (T_n^{*(1)}, \cdots, T_n^{*(B)\times B_S})'$ in an ascending order to form the distribution of T_n^*. The null hypothesis H_0 is rejected when T_n from original sample is greater than the upper-α point of the conditional distribution of T_n^*, denoted by T_α^*, where α denotes the significance level.

Repeat to generate B_N original samples under the same model specification as for the GLR test statistic T_n above. As to each original sample, there is a new value of the test statistic $\{T_{i,n}\}_{i=1,\cdots,B_N}$. Taking advantage of $\{T_{i,n}\}_{i=1,\cdots,B_N}$, we define the power of our test as

$$\text{power} = \frac{\text{frequency of rejections}}{B_N} = \frac{\sum\limits_{i=1}^{B_N} 1\{T_{i,n} > T_\alpha^*\}}{B_N}.$$

The rejection decision is made by comparing $\{T_{i,n}\}_{i=1,\cdots,B_N}$ with upper-α point of the distribution of T_n^* obtained through above procedure.

4 Mont Carlo Simulation Studies

4.1 Data Generating Process

In this section, we evaluate the finite sample performance of the proposed method using Monte Carlo simulations. To this effect, we consider the same data generating process as Hansen (2005) due to its genuine property and ease to compare our result with that of SPA test. Loss function $L_{k,t}$ is generated under the following assumption

$$L_{k,t} \sim iid \;\; N(\lambda_k/\sqrt{n}, \sigma_k^2) \quad k = 1, \cdots, m \text{ and } t = 1, \cdots, n, \tag{4.1}$$

and the benchmark model has $\lambda_k = 0$ for all k. Recall the definition of loss function and we know that $L_{k,t} > 0$ corresponds to model that is worse than benchmark when $L_{k,t} < 0$ means it is better than the benchmark model.

The experiment is designed to control the value of λ_k which is equivalent to choosing the poor model and superior model. According to Hansen (2005), we have $\lambda_1 \leq 0$ and $\lambda_1 \geq 0$ for $k = 1, \cdots, m$, such that the first alternative (k=1) defines whether the rejection probability corresponds to a type I error ($\lambda_1 = 0$) or a power ($\lambda_1 < 0$). The performances of the "poor" models are such that their mean values are spread evenly between 0 and $\lambda_m = \Lambda_0$ (the worst model). Therefore, the vectors of the λ_k's are $\lambda_0 = 0$, $\lambda_1 = \Lambda_0$, $\lambda_k = (k-1)\Lambda_0/(m-1)$ for $2 \leq k \leq m$. We use $\Lambda_0 = 0$, 1, 2, 5, and 10 to control the extent to which the inequalities are binding with ($\Lambda_0 = 0$ corresponding to the case where all inequalities are binding). The alternative model has $\Lambda_1 = 0$, -0.1, -0.2, -0.3, -0.4, and -0.5 sequentially. Therefore, $\lambda_1 = \Lambda_1$ defines the local alternative that is being analyzed. $\Lambda_1 = 0$ then, conforms to null hypothesis, whereas $\Lambda_1 < 0$ violates the null. The variance reflects the "quality" of the model. The better the model, the smaller the variance is. Specifically, by setting $\sigma_k^2 = \exp(\arctan(\lambda_k))/2$, the specification of variance is $\mathrm{Var}(d_{k,t}) = \mathrm{Var}(L_{0,t} - L_{k,t}) = 1/2 + \mathrm{Var}(L_{k,t})$.

4.2 Simulation result

We consider two experiments. First, we set $m = 100$ and $n = 200$, then increase the sample size n to 1000. In the second experiment, we have $B \times B_S = 6000$ values to generate the bootstrap distribution of GLR estimator. The rejection frequencies we report are based on $B_N = 1000$ simulations. The results are reported under 5% and 10% level in Tables 1 and 2. Furthermore, SPA test results under the same significance level and size are also exhibited. When $\Lambda_1 = 0$ in every panel in Tables 1 and 2, all the alternatives conform to null hypothesis. Consequently, the rejection frequencies correspond to type I error. In other cases, as $\Lambda_1 < 0$, the rejection frequencies are the power of the test. In contrast to SPA test which uses a relative coarse measurement, say $\Lambda_1 = 0$, -1, -2, -3, -4, and -5, we change it into $\Lambda_1 = 0$, -0.1, -0.2, -0.3, -0.4, and -0.5. It is easy to find that our method approaches 100% power at a faster speed. No matter whatever the sizes and model specifications, our method dominates SPA test in terms of power.

In Table 1, $\Lambda_0 = \Lambda_1 = 0$ refers to the situation that all the 100 inequalities are binding. It is the case in White's LFC-based RC test where all the poor models are discarded. The rejection probability is close to and less than the nominal levels. For exampel, when we set α=5%, the rejection probability is 3%, and if we change α to 10%, the probability to reject is 8.8%. It appears to be a small sample problem because this problem is alleviated when the sample size increases to 1000. The power increases when the model performs better and better in its mean relative to benchmark. In Table 2, we choose $\Lambda_0 = \Lambda_1 = 0$, α=5%, the probability of rejection is 4.9% and it increases to 9% if the α is set to be 10%. Furthermore, with large sample size, the speed to increase is higher. One can observe from Table 2 that, within large sample, our method gains power faster than that under small sample. In the case of $(\Lambda_0, \Lambda_1) = (0, -0.2)$, the power goes to almost 100% while in Table 1 the first time to reach full power happens at the point $(\Lambda_0, \Lambda_1) = (0, -0.5)$ in the panel of $\Lambda_0 = 0$. This may be due to the positive correlation across alternatives, $\mathrm{Cov}(d_{i,t}, d_{j,t}) > 0$.

Comparing with SPA test which nearly can not reject the null hypothesis when $\Lambda_1 = 1$ except the case of $\Lambda_0 = 0$, our test reaches 100% power even when $\Lambda_1 = -0.5$. Similarly, we find that no matter how poor model we choose (the level of Λ_0), our method always dominates SPA test. Another important improvement is that our test is less conservative than SPA. In SPA, the type I error shrinks fast with the increase of Λ_0, such that it is only

0.007 when $(\Lambda_0, \Lambda_1) = (10, 0)$ far away from nominal level 5%. Under our test with the values of Λ_0 and Λ_1, it is around 5% with less extreme low values.

5 Concluding Remarks

This paper proposes a new method to analyze superior predictive ability of multiple models over a benchmark. We explicitly approximate the covariance matrix by invoking certain decomposition, which is simplified via decreasing the number of estimated elements. Such approximating covariance matrix is even applicable to the case that competing models exceeds the sample size, which is considered to be infeasible to estimate by Hansen (2003, 2005). With more information from the diagonal and off-diagonal terms, the power increases dramatically comparing with SPA test which only takes into account the diagonal elements. That is because the dependence of each models contains knowledge useful to forecast. This is illustrated when we use a large sample size, say $n = 1000$, where the type I error generated by controlling the value of input parameters, approaches the nominal level.

Due to the uniform most power property of generalized likelihood ratio test, we use it instead of t-test to control the nuisance parameter problem in composite hypothesis and the convergence rate. It follows a pivotal distribution – distribution with certain degree of freedom and it is convenient to use.

Our Monte Carlo simulations show that the GLR test dominates the SPA test proposed by Hansen (2005) in terms of power and our GLR test is sensitive to the inclusion of superior models. Therefore, it increases the power faster than that of SPA test. The result also suggests that the GLR test is less conservative than SPA test.

References

Cai, Z., J. Fan and Q. Yao (2000). "Functional-coefficient regression models for nonlinear time series". *Journal of American Statistical Association* 95, 941-956.

Cai, Z. and R. Tiwari (2000). "Application of a local linear autoregressive model to BOD time series. *Environmetrics*, 11, 341-350.

Diebold, F.X., and Mariano, R.S. (1995). "Comparing predictive accuracy". *Journal of Business & Economic Statistics*, 13, 353-367.

Fan, J. and J. Jiang (2005). "Nonparametric inference with generalized likelihood ratio tests". *Test*, 16, 471-478.

Fan, J., C. Zhang and J. Zhang (2000). "Generalized likelihood ratio statistics and Wilks phenomenon". *The Annals of Statistics*, 29, 153-193.

Hansen, P.R. (2003). "Asymptotic tests of composite hypotheses". *http://www.stanford.edu/people/peter.hansen.*

Hansen, P.R. (2005). "A test for superior predictive ability". *Journal of Business & Economic Statistics*, 23, 365-380.

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

Leamer, E. (1978). *Specification searches: Ad hoc inference with nonexperimental data.* New York: Wiley.

Leamer, E. (1983). "Let's take the con out econometrics". *American Economic Review*, 73, 31-43.

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

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

Sullivan, R., A. Timmermann, and H. White (1999). "Data-snooping, technical trading rule performance, and the bootstrap". *Journal of Finance*, 54, 1647

White, H. (2000). "A reality check for data snooping". *Econometrica*, 68, 2079-1126.

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

Table 1: Rejection Frequencies under the Null and Alternative (m=100 and n=200)

Level: α=0.05				Level: α=0.10			
Λ_1	GLR	Λ_1	SPAC	Λ_1	GLR	Λ_1	SPAC
Panel A: Λ_0=0							
0	0.03	0	0.06	0	0.088	0	0.11
-0.1	0.048	-1	0.074	-0.1	0.099	-1	0.129
-0.2	0.172	-2	0.28	-0.2	0.331	-2	0.389
-0.3	0.609	-3	0.764	-0.3	0.761	-3	0.845
-0.4	0.96	-4	0.979	-0.4	0.988	-4	0.99
-0.5	1	-5	1	-0.5	1	-5	1
Panel B: Λ_0=1							
0	0.052	0	0.022	0	0.153	0	0.044
-0.1	0.123	-1	0.041	-0.1	0.288	-1	0.072
-0.2	0.409	-2	0.252	-0.2	0.613	-2	0.345
-0.3	0.789	-3	0.744	-0.3	0.92	-3	0.829
-0.4	0.977	-4	0.977	-0.4	0.993	-4	0.989
-0.5	0.999	-5	1	-0.5	1	-5	1
Panel C: Λ_0=2							
0	0.048	0	0.012	0	0.151	0	0.026
-0.1	0.118	-1	0.032	-0.1	0.261	-1	0.058
-0.2	0.421	-2	0.244	-0.2	0.69	-2	0.336
-0.3	0.849	-3	0.745	-0.3	0.933	-3	0.827
-0.4	0.994	-4	0.978	-0.4	1	-4	0.989
-0.5	1	-5	1	-0.5	1	-5	1
Panel D: Λ_0=5							
0	0.054	0	0.007	0	0.107	0	0.013
-0.1	0.16	-1	0.031	-0.1	0.236	-1	0.054
-0.2	0.516	-2	0.273	-0.2	0.617	-2	0.37
-0.3	0.907	-3	0.787	-0.3	0.944	-3	0.86
-0.4	0.999	-4	0.986	-0.4	0.999	-4	0.995
-0.5	1	-5	1	-0.5	1	-5	1
Panel E: Λ_0=10							
0	0.02	0	0.007	0	0.081	0	0.015
-0.1	0.112	-1	0.043	-0.1	0.22	-1	0.073
-0.2	0.499	-2	0.34	-0.2	0.64	-2	0.455
-0.3	0.913	-3	0.843	-0.3	0.956	-3	0.907
-0.4	1	-4	0.992	-0.4	1	-4	0.998
-0.5	1	-5	1	-0.5	1	-5	1

Table 2: Rejection Frequencies under the Null and Alternative (m=100 and n=1,000)

Level: α=0.05				Level: α=0.10			
Λ_1	GLR	Λ_1	SPAC	Λ_1	GLR	Λ_1	SPAC
Panel A: Λ_0=0							
0	0.049	0	0.048	0	0.09	0	0.1
-0.1	0.326	-1	0.064	-0.1	0.495	-1	0.122
-0.2	0.998	-2	0.282	-0.2	0.999	-2	0.39
-0.3	1	-3	0.762	-0.3	1	-3	0.84
-0.4	1	-4	0.98	-0.4	1	-4	0.99
-0.5	1	-5	1	-0.5	1	-5	1
Panel B: Λ_0=1							
0	0.07	0	0.017	0	0.226	0	0.039
-0.1	0.67	-1	0.036	-0.1	0.822	-1	0.069
-0.2	1	-2	0.252	-0.2	1	-2	0.342
-0.3	1	-3	0.74	-0.3	1	-3	0.814
-0.4	1	-4	0.978	-0.4	1	-4	0.985
-0.5	1	-5	1	-0.5	1	-5	1
Panel C: Λ_0=2							
0	0.067	0	0.009	0	0.146	0	0.021
-0.1	0.689	-1	0.029	-0.1	0.802	-1	0.054
-0.2	1	-2	0.242	-0.2	1	-2	0.322
-0.3	1	-3	0.737	-0.3	1	-3	0.798
-0.4	1	-4	0.979	-0.4	1	-4	0.983
-0.5	1	-5	1	-0.5	1	-5	1
Panel D: Λ_0=5							
0	0.045	0	0.005	0	0.085	0	0.008
-0.1	0.666	-1	0.028	-0.1	0.828	-1	0.042
-0.2	1	-2	0.267	-0.2	1	-2	0.306
-0.3	1	-3	0.777	-0.3	1	-3	0.784
-0.4	1	-4	0.987	-0.4	1	-4	0.981
-0.5	1	-5	1	-0.5	1	-5	1
Panel E: Λ_0=10							
0	0.017	0	0.005	0	0.098	0	0.005
-0.1	0.646	-1	0.042	-0.1	0.74	-1	0.039
-0.2	1	-2	0.335	-0.2	1	-2	0.299
-0.3	1	-3	0.835	-0.3	1	-3	0.778
-0.4	1	-4	0.994	-0.4	1	-4	0.98
-0.5	1	-5	1	-0.5	1	-5	1