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: June 6, 2011

In this paper we propose a new method to test the superior predictive ability (SPA) of technical trading rules relative to a benchmark. The proposed test has no potential data snooping bias. Unlike previous methods, we model the covariance matrix by factor models and develop a generalized likelihood ratio (GLR) test for the above testing problem. The GLR test is then extended to a stepwise GLR (Step-GLR) test in the spirit of the Step-RC test of Romano and Wolf (2005) and Step-SPA test of Hsu et al. (2010), which can identify predictive models without potential data snooping bias. Asymptotic null dsitribution is approximated by resampling procedures. Our results show that the GLR test is much more powerful but less conservative than the SPA test of Hansen (2005).

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

1 Introduction

Testing superior predictive ability (SPA) of some forecasting procedures over a particular forecasting one is an important problem in economics and finance. In financial markets technical rules have been exhaustively used since W.P. Hamilton published a series of papers in The Wall Street Journal in 1902, but a good forecasting model with observed superior performance may possibly come from pure luck instead of genuinely forecasting ability. As White (2000) pointed out, "*it 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.*" While the SPA of technical rules are in controversy, there are numerous empirical results supporting them. See Sweeney (1988), Blume et al. (1994), Brown et al. (1998), Gencay (1998), Lo et al. (2000), Savin et al. (2007), and Hsu et al. (2010). These evidences indicate that the problem is not within the technical rules but the data snooping bias in testing SPA. (see for example, Lo and MacKinlay (1999), Brock et al. (2000), White (2000), Hus et al. (2010)).

Data snooping widely exists in practice, especially in various applied fields such as finance and economics, where only a single history of interest is available for analysis, such as stock price, interest rate, etc. As noted by Sullivan (1999), "*data snooping can result from a subtle survivorship bias operating on the entire universe of technical trading rules that have been considered historically.*"

In general, the testing problem 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. White (2000) formulated the test of SPA as a large-scale simultaneous test for data snooping and proposed the reality check (RC) test to attck the problem. Romano and Wolf (2005) introduced a RC-based stepwise (Step-RC) test to identify 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 et al. (2010) extended the SPA test to a stepwise SPA test that can identify predictive models in large-scale, multiple testing prob-

lems without data snooping bias. They found that technical rules have significant predictive ability prior to the inception of exchange traded funds (ETF) in U.S. growth markets.

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, which is a large-scale simultaneous test for SPA. Our work contributes in the following aspects:

(i) First, no matter the RC test or SPA test, both circumvent estimation of a large covariance matrix or ignore the dependence within the models, which may result in inefficient inference. We model the covariance matrix by factor models, where variance is contributed by common background noise and underlying economical factors. This approach is applicable to the case where the number of 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).

(ii) Secondly, we incorporate the covariance structure in our estimation and extend the generalized likelihood ratio (GLR) statistics for testing SPA. 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 of the covariance matrix.

(iii) Thirdly, as Hansen (2005) suggested, the testing problem of composite hypotheses is closely related to the problem of testing hypotheses in the presence of nuisance parameters. Typically a test will suffer from loss of powers when the number of nuisance parameters is very large. Our GLR test is a type of likelihood ratio tests independent of nuisance parameters due to Wilks' phenomenon (Wilks, 1937). For various situations it is shown that the GLR test statistic follows asymptotically a scaled χ^2-distribution with the scaling constants and the degrees of freedom independent of the nuisance parameters. Further, the GLR tests are asymptotically optimal in the sense that they achieves optimal rate of convergence (see Fan, Zhang and Zhang (2001) and Fan and Jiang (2005, 2007)). It can be expected that the proposed GLR test is more persuasive than the SPA test.

(iv) Following the idea of Step-RC test (Romano and Wolf (2005) and Step-SPA test (Hsu et al. (2010), we extend the GLR test to the Step-GLR test. This allows us to identify which models are superior to the benchmark.

(v) We provide a bootstrap method to implement the proposed GLR test.

The rest of this paper is organized as follows. In Section 2, we review the existing tests. In Section 3, we describe our testing procedure in detail. In Section 4, we conduct simulations to assess the effectiveness of the proposed method and to compare it with the SPA test. In Section 5, we give a concluding remark.

2 Review of Existing Tests

2.1 Reality Check Test

Suppose we have m models for some variable. Let $d_{k,t}$ be the performance measure of the k-th model relative to a benchmark model at time t for $t = 1, 2, \cdots, n$. In the framework of White (2000), to determine if there is a model with predictive superiority over the benchmark, one would like to test the null hypothesis:

$$H_0^k : \mu_k \leq 0, \quad k = 1, \ldots, m, \tag{2.1}$$

where $\mu_k = E(d_{k,t})$. Then 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 the null hypothesis, the limiting distribution of RC_n can be approximated via a bootstrap procedure. The null hypothesis is rejected when the bootstrapped p-value is smaller than a pre-specified significance level. While the LFC is convenient to implement, the RC test also bears a few drawbacks. As Hansen (2003, 2005) pointed out, 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. 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

$$\mathrm{SPA}_n = \max\Big[\max_{1\le k\le m} \sqrt{n}\bar{d}_k/\hat{\sigma}_k,\ 0\Big], \tag{2.3}$$

where $\hat{\sigma}_k^2$ is a consistent estimator of $\sigma_k^2 = \omega_{kk}$. The main argument for the normalization is that it will improve the power typically. Since it uses a data-dependent choice for μ instead of $\mu = 0$ implied by the LFC condition, it usually leads to a more powerful test of composite hypotheses.

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 mean $E(d_k) = \mu_k$ is estimated by

$$\hat{\mu}_k = \bar{d}_k \cdot 1\{\sqrt{n}\bar{d}_k/\hat{\sigma}_k \le -\sqrt{2\log\log n}\}, \quad k = 1, 2, \cdots, m, \tag{2.4}$$

where $1\{\cdot\}$ denotes the indicator function. Advantages of the estimator $\hat{\mu}_k$ were clarified in Hansen (2005). We will use the estimator $\hat{\mu}_k$ when the null holds.

3 Methodology to Test Superior Predictive Ability

3.1 Estimation Procedure

Suppose we model the $d_t = (d_{1,t}, \ldots, d_{mt,})^T$ by

$$d_t = \mu + e_t = \mu + \Omega^{1/2}\varepsilon_t \quad t = 1, 2, \cdots, n, \tag{3.1}$$

where $\Omega = [\omega_{ij}]_{m\times m}$ is a definite positive covariance matrix and $\{\varepsilon_t\}_{t=1}^n$ are independent and identically distributed with mean 0 and covariance matrix $I_{m\times m}$, where $I_{m\times m}$ is the $m \times m$ identity matrix. This means that d_t is decomposed into two parts. The first part is the mean value of the trading rules during certain period, and the second one, the error term, is the systematic noise, which explains the variation of the trading rules. This decomposition allows us to explicitly model the covariance matrix later.

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). Specifically, we first make a singular value decomposition (SVD) of Ω,

$$\Omega = QDQ^T, \tag{3.2}$$

where $Q = (q_1, \cdots, q_m)$ is an $m \times m$ orthogonal matrix with $q_i^T q_j = 1$ for $i = j$, and $q_i^T q_j = 0$ for $i \neq j$, and D is an $m \times m$ diagonal matrix with decreasing entries on the diagonal. Then D is the variance matrix of the transformed data $Q^T d_t$.

Motivated by the idea of Liu et al. (2008) for clustering high-dimension, low-sample size data sets in gene expression microarray data, we decomposed the variance matrix D into two parts: one is caused by d ($d < m$) real economic factors, and the other is caused by a common noise with mean zero and variance δ^2. Therefore, it can be rewritten as

$$D = S_{m\times m} + \delta^2 I_{m\times m},$$

where $S_{m\times m} = \text{diag}(s_1, \cdots, s_d, 0, \ldots, 0)$ is determined by d real economical factors gathering the most important information specific to each trading rule, $\{s_d\}$ is arranged in a decreasing order, and the value of d is decided from data. In summary, the matrix D admits the following decomposition:

$$D_{m\times m} = \text{diag}\{\lambda_1, \cdots, \lambda_d, \delta^2, \cdots, \delta^2\}, \tag{3.3}$$

where $\lambda_j = s_j + \delta^2$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. The values of s_j reflect the strength of effects of real economic factors, and δ^2 reflects the level of background noise shared by all trading rules.

In general, μ_k can be estimated by $\bar{d}_k$ which is in the form of $\bar{d}_k = n^{-1}\sum_{t=1}^{n} d_{k,t} \quad k = 1, 2, \cdots, m$. Hence, the residual from (3.1) is estimated by $\hat{e}_t = d_t - \bar{d}, \quad t = 1, 2, \cdots, n$. Based on $\{\hat{e}_t\}_{m\times 1}$, the residual sample covariance matrix is given by

$$\hat{\Omega} = \frac{1}{n-1}\Sigma_{t=1}^{n}\hat{e}_t\hat{e}_t'. \tag{3.4}$$

Note that the dimension m is generally large relative to the sample size n. $\hat{\Omega}$ cannot provide a good estimate of Ω, but it allows us to utilize the structure in (3.2).

Since δ^2 reflects the variance of the common background noise shared by all trading rules, it can be estimated by $\hat{\delta}^2 = \frac{1}{mn-1}\sum_{t=1}^{n}\|\hat{e}_t\|^2$. From covariance matrix $\hat{\Omega}$, we get its eigenvalues

$\{\lambda_i^*\}_{i=1}^m$ and the corresponding normalized eigenvectors $\{v_i\}_{i=1}^m$, where $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_m^*$. Let $\hat{Q} = (v_1, v_2, \cdots, v_m)$. Then by (3.3) the matrix D can be estimated by

$$\hat{D} = \text{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_d, \hat{\delta}^2, \ldots, \hat{\delta}^2\},$$

where $\hat{\lambda}_j = (\lambda_j^* - \hat{\delta}^2)\, 1(\lambda_j^* \geq \hat{\delta}^2)$ and $1(\cdot)$ is an indicator function. Therefore, by (3.2), the covariance matrix Ω is estimated by $\hat{\Omega}^* = \hat{Q}\hat{D}\hat{Q}^T$.

3.2 GLR Test

The testing problem in (2.1) is a high-dimensional nonparametric null hypothesis versus a high-dimensional nonparametric alternative. Since the distribution of ε_t is unspecified, we do not have a known likelihood function. Even though a likelihood is available when the distribution is given, the number of parameters in Ω is too large so that it is challenging to make an efficient inference for the parameters. To this end, we extend the GLR test to the current high-dimensional setting.

For any vector a, let $\|a\|$ be L_2-norm of a. Define the residual sum of squares under the null and alternative as follows:

$$\text{RSS}_0 = \sum_{t=1}^n \|\hat{\Omega}^{*-\frac{1}{2}}(d_t - \hat{\mu})\|^2$$

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

$$\text{RSS}_1 = \sum_{t=1}^n \|\hat{\Omega}^{*-\frac{1}{2}}(d_t - \bar{d})\|^2,$$

respectively. Then we define the GLR test statistic as

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

which compares the likelihood of the nearly best fitting under the alternative model with that under the null model. The null hypothesis is rejected when T_n is too large.

For finite-dimensional data in various scenarios, the GLR statistics follow asymptotically rescaled chi-squared distributions, with the scaling constants and the degrees of freedom independent of the nuisance parameters, which enables one to use simulations to approximate the null distribution. The resulting tests are powerful, since they are asymptotically optimal

in the sense that they can detect alternatives with optimal rates for nonparametric hypothesis testing according to the formulation of Ingster(1993) and Spokoiny (1996). For details, see Cai, Fan and Yao (2000), Cai and Tiwari (2000), Fan, Zhang and Zhang (2001) and Fan and Jiang (2005, 2007).

3.3 Calculating P-Value

We now introduce a bootstrap procedure to calculate the P-value of the GLR test. Given $d_{k,t}$, we use the proposed estimation method to compute the GLR test statistic T_n and the residuals from the alternative

$$\hat{\varepsilon}_t = \hat{\Omega}^{*-\frac{1}{2}}(d_t - \hat{\mu}),\ t = 1, 2, \ldots, n. \tag{3.6}$$

Then draw bootstrap residuals $\hat{\varepsilon}_t^*$ with size n from the centered empirical distribution of $\{\hat{\varepsilon}_t\}_{t=1}^n$ and compute

$$d_t^* = \hat{\mu} + \hat{\Omega}^{*-\frac{1}{2}}\hat{\varepsilon}_t^* \quad t = 1, 2, \cdots, n.$$

This forms a bootstrap sample $\{d_{k,t}^*\}$ $(k = 1 \ldots, m; t = 1, \ldots, n)$. Then use the bootstrap sample to obtain the GLR test statistic T_n^*. Repeat this procedure many times, then we get a sample of the GLR statistic T_n^*, which can be used to determine the quantiles of the test statistic under H_0. The P-value is the percentage of observations from the sample of T_n^* whose values are bigger than T_n.

3.4 Step-GLR Test

To identify significant models, we extend the GLR test to the Step-GLR test using the ideas of Step-RC test of Romano and Wolf (2005) and Step-SPA test of Hsu (2010). The Step-GLR test should be more powerful since it inherits the advantages of the GLR test.

The idea is similar to the backward elimination method for variable selection and the one-case deleted procedure for regression diagnostics in linear models. First we calculate the GLR statistic T_n without using the data for i-th model $(i = 1, \ldots, m)$. To stress independence on the model i. We denote by $T_{n,-i}$ the resulting GLR statistic. Then we define

$$T_{n,-i} = \frac{mn}{2}(\mathrm{RSS}_{0,-i} - \mathrm{RSS}_{1,-i})/\mathrm{RSS}_{1,-i}. \tag{3.7}$$

where $RSS_{0,-i} = \sum_{t=1}^{n} \|\hat{\Omega}_{-i}^{*-\frac{1}{2}}(d_t - \hat{\mu})\|^2$, $RSS_{1,-i} = \sum_{t=1}^{n} \|\hat{\Omega}_{-i}^{*-\frac{1}{2}}(d_t - \bar{d})\|^2$, and $\hat{\Omega}_{-i}^{*}$ is defined the same as $\hat{\Omega}^*$ without using the data for the i-th model. The Step-GLR test with the pre-specified level α_0 then proceeds as follows.

1. Let $\Delta T_{ni} = T_n - T_{n,-i}$ for $i = 1, \ldots, m$. Re-arrange ΔT_{ni} in a descending order.

2. Reject the top model k if T_n is greater than the critical value T_α^* (all) from the full sample. Otherwise, the procedure stops.

3. Remove $\{d_{k,t}\}_{t=1}^n$ of the rejected models from the data. Let ΔT_{ni} (sub) be defined the same as ΔT_{ni} but with the remaining data as the new full sample. Reject the top model i in the sub-sample of remaining observations if ΔT_{ni} (sub) is greater than T_α^* (sub), the critical value from the sub-sample. If no model can be rejected, the procedure stops.

4. Repeat the Step 3 until no model can be rejected.

4 Mont Carlo Simulation Studies

4.1 Data Generating Process

In this section, we evaluate finite sample performance of the proposed method using Monte Carlo simulations. To this end, we consider the same data generating process as Hansen (2005) due to its genuine property and ease to compare with our test. We will use the notions in Hansen (2005).

Let performance of the kth trading strategy be measured by loss function relative to that of benchmark, instead of its absolute value:

$$d_{k,t} = L(\xi_t, \delta_{0,t-h}) - L(\xi_t, \delta_{k,t-h}), \quad k = 1, 2, \cdots, m, \tag{4.1}$$

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. For example, in Hansen (2005), $\delta_{k,t-1}$ 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

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

Therefore, we focus on $d_{k,t}$ exclusively rather than the loss function itself.

The benchmark is the target to compare with. It is reflected in $d_{k,t}$ as the performance of kth trading rule is net of that of a benchmark. For example, for a fund manager who cares about 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 was used by Sullivan et al. (1999, 2001) and will be used in this paper.

Loss function $L_{k,t} = L(\xi_t, \delta_{k,t-h})$ is generated from the model

$$L_{k,t} \sim iid \;\; N(\lambda_k/\sqrt{n}, \sigma_k^2) \quad k = 1, \cdots, m \text{ and } t = 1, \cdots, n, \tag{4.2}$$

and the benchmark model has $\lambda_k = 0$ for all k. By the definition of loss function, $L_{k,t} > 0$ means that the kth model is worse than benchmark, and $L_{k,t} < 0$ means that 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 poor models are those with mean values being equally spaced between 0 and $\lambda_m = \Lambda_0$ (the worst model). That is, the values of λ_k's are are set as $\lambda_0 = 0$, $\lambda_1 = \Lambda_1$, $\lambda_k = (k-1)\Lambda_0/(m-1)$ for $2 \leq k \leq m$. We set Λ_0 to be 0, 1, 2, 5, and 10. The alternative models have $\Lambda_1 = 0$, -0.1, -0.2, -0.3, -0.4, and -0.5, respectively. Therefore, $\lambda_1 = \Lambda_1$ defines the local alternative that is being analyzed. When $\Lambda_1 = 0$, the null hypothesis is the same as the alternative. As Λ_1 deviates away from 0 on the left, the alternative get more and more away from the null. The variance reflects the quality of the model. The smaller the variance, the better the model. As Hansen (2005), we set $\sigma_k^2 = \exp(\arctan(\lambda_k))/2$, which indicates that the specification of variance is $\text{Var}(d_{k,t}) = \text{Var}(L_{0,t} - L_{k,t}) = 1/2 + \text{Var}(L_{k,t})$.

4.2 Simulation result

In this section we conduct a small simulation. We set $m = 100$ and $n = 200$ and 1000. To get the null distribution of the GLR test, we run 1000 simulations under the null model. For each simulation we generate 6 bootstrap samples of the GLR statistic T_n and then pool them together to get the p-value of T_n using the procedure introduced before. To evaluate the power of test, we run 1000 simulations which sample under the alternative. The rejection frequency in 1000 simulations are reported. Then the power of test is calculated as the relative rejection of frequency.

The results are reported under 5% and 10% level in Tables 1 and 2. In addition, for comparison SPA test results 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 much 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 example, when we set α=5%, the rejection probability is 3%, and if we change α to 10%, it becomes 8.8%, which are a little bit far away from the levels. This appears to be a small sample problem, because this problem is alleviated when the sample size increases to 1000. In Table 2, when $\Lambda_0 = \Lambda_1 = 0$, the probability of rejection is 4.9% for α=5% and 9% for $\alpha = 10\%$. It is seen that, with larger sample size, the speed of power to increase is higher. One can observe from Table 2 that, with larger sample than Table 1, our method gains power faster than that under small sample. For example, 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 cannot 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 the model is (which depends on 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. For example, it is only 0.007 when $(\Lambda_0, \Lambda_1) = (10, 0)$ which is far away from the nominal level 5%, but for our test it is around 5% with less extreme low values.

5 Concluding Remarks

We have proposed the GLR and Step-GLR tests to analyze superior predictive ability of multiple models over a benchmark. We explicitly approximate the covariance matrix by factoring the covariance matrix and by using PCA. 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). Simulations demonstrate that the power of the GLR test is much higher than the SPA test. This may be due to the nature of GLR test and the dependence structure of the alternative models used in estimation. Our results also indicate that the GLR test is less conservative than the SPA test. Further work includes deriving the asymptotic null distribution and the theoretical power of the GLR test, which is under investigation of our research project.

References

Brock, W., J. Lakonishok and B. Lebaron (1992). "Simple technical trading rules and the stochastic properties of stock returns". *Jounal of Finance* 47, 1731-1764.

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.

Gencay, R. (1998). "The predictability of security returns with simple technical trading rules". *Journal of Empirical Finance*, 5, 347-359.

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.

Ingster, Yu. I. (1993),"Asymptotic Minimax Hypothesis Testing for Nonparametric Alternatives ICIII." *Mathematical Methods in Statistics*, 2, 85C114; 3, 171C189; 4, 249C268.

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.

Lo, A.W. and C.A. MacKinlay (1990). "When are contrarian profits due to stock market overreaction?". *Review of Financial Studies*, 3, 175-206.

Lo, A.W. and Mamaysky, H. and Wang, J. (2000). "Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation". *Journal of Finance*, 55, 1705-1765.

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.

Savin, G., Weller, P., Zvingelis, J. (2007). "The predictive power of "head-and-shoulder" price patterns in the U.S. stock market. *Journal of Financial Econometrics*, 5, 243-265.

Spokoiny, V. G. (1996),"Adaptive Hypothesis Testing Using Wavelets." *The Annals of Statistics*, 24, 2477C2498.

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	SPA	Λ_1	GLR	Λ_1	SPA
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	SPA	Λ_1	GLR	Λ_1	SPA
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