Empir Econ (2015) 48:389–405
DOI 10.1007/s00181-014-0887-6

A new semiparametric test for superior predictive ability

Zongwu Cai · Jiancheng Jiang · Jingshuang Zhang · Xibin Zhang

Received: 3 April 2013 / Accepted: 29 September 2014 / Published online: 3 December 2014

Abstract We propose a new method to test the superior predictive ability (SPA) of a benchmark model against a large group of alternative models. The proposed test is useful for reducing potential data snooping bias. Unlike previous methods, we model the covariance matrix by factor models and develop a generalized likelihood ratio (GLR) test statistic for the above testing problem. The GLR test is also extended to a stepwise GLR (step-GLR) test in the spirit of the step-RC test of Romano and Wolf (Econometrica 73(4):1237–1282, 2005) and step-SPA test of Hsu et al. (J Empir Financ 17(3):471–484, 2010). The step-GLR test can identify the most contributed predictive models to the rejection of the null hypothesis. A Monte Carlo simulation study shows that the GLR test is much more powerful and less conservative than

Z. Cai (✉)
Department of Economics, University of Kansas, Lawrence, KS 66045, USA
e-mail: caiz@ku.edu

Z. Cai
Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Sciences, Xiamen University, Xiamen 361005, Fujian, China

J. Jiang
Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
e-mail: jjiang1@uncc.edu

J. Zhang
Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen 361005, Fujian, China
e-mail: zhang.jshuang@gmail.com

X. Zhang
Department of Econometrics and Business Statistics, Monash University, Caulfield East, VIC 3145, Australia
e-mail: xibin.zhang@monash.edu

the SPA test of Hansen (J Bus Econ Stat 23(4):365–380, 2005). We also present an application to illustrate the use of the GLR test and make a comparison between our GLR and Hansen's SPA tests.

Keywords Data snooping · Generalized likelihood ratio · Reality check · Technical trading rules · Variance matrix estimation

JEL Classification C14 · C53

1 Introduction

Testing the superior predictive ability (SPA) of a specific forecasting procedure against a group of alternative forecasting procedures is of importance in business and economic forecasting. For example, in financial markets, quantitative analysis, such as a technical trading rule,[1] has been exhaustively used since W.P. Hamilton published a series of papers in *The Wall Street Journal* in 1902. A good forecasting model with observed superior performance may possibly come from pure luck instead of genuine 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." Therefore, it is important to test whether a particular forecasting model outperforms a large group of alternative competing models. For this purpose, Hansen (2005) proposed the so-called SPA test. In this paper, we propose a new test for testing SPA.

Although the conclusions about predictive power of technical analysis are mixed in the literature, there are numerous empirical studies to support using technical analysis. See, for example, the papers by Sweeney (1988), Blum et al. (1994), Brown et al. (1998), Gencay (1998), Lo et al. (2000), Savin et al. (2007), Hsu et al. (2010) and the references therein. However, such evidence is likely to be criticized due to their data snooping bias in testing SPA; see Lo and MacKinlay (1990), Brock et al. (1992), White (2000), Hsu et al. (2010), among others.

Data snooping occurs when a set of data are repeatedly used for the purpose of inference or model selection. This is due to the fact that when reusing such data, we create the possibility that any satisfactory results may simply be obtained by chance rather than by any merit inherent in the method yielding the results (White 2000). As noted by Sullivan et al. (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 issue of testing SPA can be addressed by testing the null hypothesis that the benchmark is not inferior to any alternative forecasting models. Diebold and Mariano (1995) and West (1996) proposed tests for equal predictive ability, which means that the forecasting ability of a model is the same as the benchmark. White (2000) formulated the test for SPA as a large-scale simultaneous test for data snooping and proposed the reality check (RC) test to solve the problem. Romano and Wolf

[1] For example, the moving average rules and the filter rules used in quantitative finance.

(2005) introduced a RC-based stepwise multiple test known as the stepM test, to identify as many significant models as possible in the sense that they outperform the benchmark. Based on the methodology proposed by White (2000), Hansen (2003) suggested a procedure to test composite hypotheses by incorporating additional sample information on nuisance parameters. Recently, Hsu et al. (2010) extended Hansen's SPA test to a stepwise SPA test which aims to identify predictive models in large-scale and multiple testing problems. They found that technical analysis has significant predictive ability prior to the inception of exchange traded funds in the US growth markets.

As in White (2000) and Hansen (2003, 2005), we also consider testing SPA by using the null hypothesis that the benchmark underperforms any alternative model. Our test is a large-scale simultaneous test for SPA and has some advantages in the following aspects.

First, both the RC test and Hansen's SPA test, as well as their variants, do not explicitly incorporate an estimate of the covariance matrix of the models or performance measures (see Sect. 2.1 for definition) into the test statistics. This may result in inefficient inference. With a belief that the dependence within a large number of models is driven by a small number of unobservable latent factors (to be decided by the data), we consider modeling the covariance matrix by a factor model, which assumes that the covariance matrix is contributed by unknown common background noise and underlying latent factors (see Sect. 3.1). This approach is particularly attractive when the number of forecasting models is large relative to the sample size. According to our simulations, Hansen's (2005) SPA has much less power than ours. This is possibly because Hansen's SPA test has a nonunique null distribution depending on nuisance parameters.

Second, by incorporating the covariance structure in our estimation, we extend the generalized likelihood ratio (GLR) test of Cai et al. (2000) and Fan et al. (2001) for testing SPA. As noted by Hansen (2003, 2005), his SPA test would be improved if there was a reliable way to incorporate information about the off-diagonal elements of the covariance matrix. Our modeling of the covariance matrix does not require distributional assumptions. Therefore, our approach is semiparametric in nature.

Third, 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, and the null distribution of his SPA test depends on these nuisance parameters. In various scenarios, it is shown that the GLR test, as an extension to the likelihood ratio test, has asymptotic null distribution independent of nuisance parameters. This is referred to as the Wilks phenomenon (Wilks 1938). See, for example, Fan et al. (2001) and Fan and Jiang (2005) for details. Moreover, the GLR test is asymptotically optimal in the sense that it achieves the optimal rate of convergence in the context of semi- and nonparametric settings; see, for example, Fan and Jiang (2005, 2007) and Jiang et al. (2007). Therefore, it is reasonable to expect that our GLR test has the above properties.

Fourth, following the idea of the step-RC test of Romano and Wolf (2005) and the step-SPA test of Hsu et al. (2010), we extend the proposed GLR test to a stepwise version, which we call the step-GLR test. This allows us to sequentially identify the models that are superior to the benchmark.

Finally, a bootstrap method is used to implement the proposed GLR test. Our simulation shows that our test procedure is powerful and has the correct size.

The rest of this paper is organized as follows. In Sect. 2, we review the existing tests for predictive performance. Section 3 describes our testing procedure in detail. In Sect. 4, Monte Carlo simulation studies are presented to assess the effectiveness of the proposed method and to compare it with Hansen's SPA test. In Sect. 5, we present an application to illustrate the practical usefulness of our GLR and step-GLR tests. Section 6 concludes the paper.

2 A review of existing tests

2.1 Reality check test

Suppose we have m models for a forecasting purpose. Let $d_{k,t}$ be a performance measure of the kth model relative to a benchmark model at time t for $t = 1, 2, \ldots, n$ and $k = 1, \ldots, m$. For each t, $d_{k,t}$ may be dependent across k. For example, for a stock return r_t at time t, let $\delta_{k,t-1}$ be the trading signal with value 1 or -1, instructed by a trader based on the k-th trading rule at time $t-1$, where 1 and -1 correspond to long and short positions, respectively. Then, $\pi_{k,t} = r_t \delta_{k,t-1}$ is the profit yielded by the k-th trading rule. Let the benchmark model correspond to $k = 0$. Then, $d_{k,t} = \pi_{0,t} - \pi_{k,t}$. In the framework of White (2000), to determine whether there is a model with predictive superiority over the benchmark, one would like to test the null hypothesis:

$$H_0^k : \mu_k \leq 0, \quad \text{for} \quad k = 1, 2, \ldots, m, \tag{1}$$

where $\mu_k = E(d_{k,t})$. For the above trading example, the null hypothesis means that there is no trading rule bringing positive mean profit. Data snooping arises when 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 given by

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

where $\bar{d}_k = n^{-1} \sum_{t=1}^{n} d_{k,t}$ for $k = 1, 2, \ldots, m$. Let $d_t = (d_{1,t} \ldots, d_{m,t})'$, $\bar{d} = (\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_m)'$, and $\mu = E(d_t)$. If $\{d_t\}$ is stationary, then under Assumption 1 of Hsu et al. (2010), $\sqrt{n}(\bar{d} - \mu)$ converges in distribution to $N(0, \Omega)$, where $\bar{d} = \sum_{t=1}^{n} d_t / n$ and $\Omega = \lim_{n \to \infty} \mathrm{Var}(\sqrt{n}(\bar{d} - \mu))$. White (2000) used the least favorable configuration (LFC), i.e., $\mu = 0$, to derive the limiting null distribution, $\max\{N(0, \Omega)\}$, of RC_n, and proposed using a stationary bootstrap method to approximate the null distribution. At significance level α, the bootstrapped critical value is decided by the $(1-\alpha)$-th percentile of the bootstrap realizations of RC_n. Once RC_n is greater than the critical value, the null hypothesis (1) is rejected. As Hansen (2003, 2005) pointed out, the RC suffers from two major drawbacks. "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 that of the RC test, Hansen (2005) proposed a studentized test

$$\mathrm{SPA}_n = \max\left\{\max_{1\le k\le m} \sqrt{n}\bar{d}_k/\hat{\sigma}_k,\ 0\right\},$$

where $\hat{\sigma}_k^2$ is a consistent estimator of $\sigma_k^2 = \mathrm{Var}(d_{k,t})$. The null distribution of SPA_n can be obtained similarly to that of the RC test, but with the bootstrapped distribution re-centered. In fact, in Hansen's SPA test, the mean $E(d_k) = \mu_k$ is estimated by

$$\hat{\mu}_k = \bar{d}_k \cdot \mathbf{1}\left(\sqrt{n}\bar{d}_k/\hat{\sigma}_k \le -\sqrt{2\log\log n}\right), \quad k = 1, 2, \ldots, m, \tag{2}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function whose value is one for a true argument and zero otherwise. When $\mu_k = 0$, $\hat{\mu}_k = 0$ almost surely, and when $\mu_k < 0$, $\sqrt{n}\bar{d}_k/\hat{\sigma}_k \le \sqrt{2\log\log n}$ with probability approaching one. Hence, $\hat{\mu}_k$ converges in probability to μ_k under the null. Since $\sqrt{n}\bar{d}/\hat{\sigma} = \sqrt{n}(\bar{d}-\mu)/\hat{\sigma} - \sqrt{n}\mu/\hat{\sigma}$, where $\hat{\sigma} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_m)'$, and the division is operated componentwise. Hansen (2005) suggested adding $\sqrt{n}\hat{\mu}/\hat{\sigma}$ to the bootstrapped distribution of $\sqrt{n}(\bar{d}-\mu)/\hat{\sigma}$. This yields a better approximation to the null distribution of SPA_n, $\max\{N(0, \Omega_0), 0\}$, and higher power than the RC test.

Motivated by the extension of the RC test to the step-RC test in Romano and Wolf (2005), Hsu et al. (2010) extended Hansen's SPA test to the step-SPA test. The step-SPA allows for identifying significant models, yet it ought to be more powerful because its null distribution does not depend on the LFC. However, in the construction of the test statistic SPA_n, the covariance matrix of d_t is not used, which results in loss of power when the covariance matrix is not diagonal. This motivates us to propose a new testing procedure described below.

3 GLR test for SPA

3.1 Estimation

Consider the performance measures $\{d_{k,t}\}$, and let $d_t = (d_{1,t}, d_{2,t}, \ldots, d_{m,t})'$. Assume that $\{d_t\}_{t=1}^n$ is strictly stationary. Let $\mu = E(d_t)$, $\Omega = \mathrm{Var}(d_t)$, and $e_t = d_t - \mu$. Then, $\{e_t\}_{t=1}^n$ is also strictly stationary with mean 0 and covariance matrix Ω and there is a strictly stationary process $\{\varepsilon_t\}$ with marginal mean 0 and marginal covariance matrix $I_{m\times m}$, such that $e_t = \Omega^{1/2}\varepsilon_t$, where $I_{m\times m}$ is an $m \times m$ identity matrix. Thus, we can write d_t as

$$d_t = \mu + \Omega^{1/2}\varepsilon_t, \quad \text{for} \quad t = 1, 2, \ldots, n. \tag{3}$$

In many applications, such as testing for the superior predictive performance of trading rules in a stock market, there are a large number of trading rules to be investigated so that m might be large. For example, Sullivan et al. (1999) evaluated 7,846 technical trading rules, and Hsu et al. (2010) carried out their investigation based on a total of 16,380 trading rules. This means that a sensible estimate of all elements of Ω is infeasible, especially when m exceeds the sample size n.

In this paper, we propose to estimate Ω using its most useful information in the spirit of principal component analysis (PCA). Specifically, we make a singular value decomposition (SVD) of Ω,

$$\Omega = QDQ', \tag{4}$$

where $Q = (q_1, q_2, \ldots, q_m)$ is an $m \times m$ orthogonal matrix with $q_i' q_j = 1$, for $i = j$, and $q_i' q_j = 0$, for $i \neq j$, and D is an $m \times m$ diagonal matrix with decreasing diagonal elements, the decreased eigenvalues of Ω. Motivated by the idea of Liu et al. (2008) for clustering high-dimension and low sample size data in gene expression microarray data analysis, we model the diagonal matrix D as

$$D = S_{m\times m} + v^2 I_{m\times m}, \tag{5}$$

where $S_{m\times m} = \text{diag}\{s_1, \ldots, s_{d^*}, 0, \ldots, 0\}$, and d^* is an unknown positive integer determined from data. This model is appropriate if the dominated eigenvalues of Ω or equivalently the d^* major principal components are driven by d^* ($d^* < m$) latent factors, while the remaining is caused by a common noise with mean zero and variance v^2. Usually d^* is much smaller than m, according to many practical studies of PCA. Let $\gamma_i = s_i + v^2$, for $i = 1, 2, \ldots, d^*$. Then, model (5) is equivalent to

$$D = \text{diag}\left\{\gamma_1, \ldots, \gamma_{d^*}, v^2, \ldots, v^2\right\}, \tag{6}$$

where $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{d^*}$.

Model (6) can be understood as follows. Assume that d_t is generated from the following orthogonal factor model with d^* ($<m$) latent common factors

$$d_t - \mu = LF_t + \epsilon_t,$$

where d_t is an $m \times 1$ vector, L is an $m \times d^*$ loading matrix, F_t is a $d^* \times 1$ vector of factors, and ϵ_t is a $d^* \times 1$ common noise vector. Assume that F_t and ϵ_t are independent and that $E(F_t) = 0$, $\text{var}(F_t) = I$, $E(\epsilon_t) = 0$, and $\text{cov}(\epsilon_t) = v^2 I$. Then, $\text{var}(d_t) = LL^\top + v^2 I$. For the nonnegative definite matrix $LL^\top$, by the spectral decomposition theorem, there are an orthogonal matrix Q and a diagonal matrix $S = \text{diag}(s_1, \ldots, s_{d^*}, 0, \ldots, 0)$ such that $LL^\top = QSQ^\top$. Hence,

$$\text{Var}(d_t) = QSQ^\top + v^2 I = QDQ^\top,$$

where $D = S + v^2 I = \text{diag}\{\gamma_1, \ldots, \gamma_{d^*}, v^2, \ldots, v^2\}$ with $\gamma_j = s_j + v^2$ for $j = 1, \ldots, d^*$. This is exactly our model in (5) or equivalently in (6). Let $Q = (q_1, \ldots, q_m)$. Then, $QSQ^\top = \sum_{j=1}^{d^*} s_j q_j q_j^\top$ and hence

$$\mathrm{Var}(d_t) = \sum_{j=1}^{d^*} s_j q_j q_j^\top + v^2 I.$$

This is the model recently studied by Birbaum et al. (2013).

In general, μ_k can be estimated by $\bar{d}_k$, for $k = 1, 2, \ldots, m$. Therefore, the residuals from (3) can be calculated through $\hat{e}_t = d_t - \bar{d}$, for $t = 1, 2, \ldots, n$. Based on $\{\hat{e}_t\}$, the covariance matrix Ω may be estimated by $\hat{\Omega} = \frac{1}{n-1}\sum_{t=1}^{n} \hat{e}_t \hat{e}_t'$. However, in practice, the dimension m is generally large relative to the sample size n, and hence such $\hat{\Omega}$ cannot be a good estimator of Ω. In the following, we employ the SVD of Ω in (4) to derive a better estimate.

As v^2 reflects the variance of the common background noise shared by all alternative models, it can be estimated by $\hat{v}^2 = (m(n-1))^{-1}\sum_{t=1}^{n} \|\hat{e}_t\|^2$. For the estimated covariance matrix $\hat{\Omega}$, we obtain its eigenvalues $\{\gamma_i^*\}_{i=1}^m$ and the corresponding normalized eigenvectors $\{\hat{q}_i\}_{i=1}^m$, where $\gamma_1^* \geq \gamma_2^* \geq \cdots \geq \gamma_m^*$. Let $\hat{Q} = (\hat{q}_1, \ldots, \hat{q}_m)$. According to (6), the matrix D can be estimated by the following thresholding estimator:

$$\hat{D} = \mathrm{diag}\left\{\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_m\right\},$$

where $\hat{\gamma}_j = (\gamma_j^* - \hat{v}^2)\cdot \mathbf{1}(\gamma_j^* \geq \hat{v}^2) + \hat{v}^2$. Therefore, by (4), the covariance matrix Ω is estimated by $\hat{\Omega}^* = \hat{Q}\hat{D}\hat{Q}'$. Note that the number of latent factors, d^*, is automatically estimated as the number of $\hat{\gamma}_j$ greater than $\hat{v}^2$. Although d^* may not be estimated well, we do not seek to estimate it accurately. Our aim is to estimate Ω in an appropriate way, such that it would account for the majority of variability in the variable $\{d_t\}$ with d^* factors. This is in the same spirit as PCA. In fact, in our real data example, it is shown that the resulting test is not sensitive to 20 % perturbation on the thresholding value $\hat{v}^2$ for choosing the number of latent variables.

3.2 GLR test

The testing problem in (1) is a high-dimensional null hypothesis versus a high-dimensional alternative. As the distribution of ε_t is unspecified, we do not have a likelihood function and hence, the likelihood ratio test cannot be applied. Even though a likelihood is available when the error distribution is specified, there will be too many parameters in Ω and thus, it will be challenging to make efficient inferences for the parameters. Cai et al. (2000) and Fan et al. (2001) proposed the GLR test to deal with this problem. The GLR test has some good properties such as the Wilks phenomenon and the optimal rate in power. See Fan and Jiang (2005, 2007) for details. Here, we extend the GLR test to the current high-dimensional setting.

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

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

where $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_m)'$ with $\hat{\mu}_k$ defined in (2). Under the alternative, we define the residual sum of squares as

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

Following Cai et al. (2000) and Fan et al. (2001), we define our GLR test statistic as

$$T_n = \frac{mn}{2} (\text{RSS}_0 - \text{RSS}_1)/\text{RSS}_1, \tag{7}$$

which compares the likelihood under the alternative with that under the null. The null hypothesis is rejected when T_n is too large.

For various scenarios, the GLR statistic is asymptotically χ^2-distributed with degrees of freedom going to ∞ and independent of nuisance parameters. This property allows one to use bootstrap methods to approximate the null distribution of T_n. The GLR statistic in (7) uses the estimator $\hat{\Omega}^*$ of Ω to improve the power of test. If Ω is not well estimated or even a naive estimate such as $\text{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_m^2\}$ is used, the test could still be used, but at the price of losing some power. Since we do not assume any distribution for the GLR test, it can be used with any type of predicting models considered in Sect. 2.1, including semi- and nonparametric methods.

3.3 Calculating the p value

We now introduce a (wild) bootstrap procedure to calculate the p value of T_n. This method is similar in spirit to that in Cai et al. (2000). Given $d_{k,t}$, we use the proposed estimation method to compute the GLR statistic T_n and the residuals from the alternative hypothesis:

$$\hat{\varepsilon}_t = \left(\hat{\Omega}^*\right)^{1/2} \left(d_t - \bar{d}\right), \quad \text{for} \quad t = 1, 2, \ldots, n.$$

Then, we draw a sample $\{\hat{\varepsilon}_t^*\}_{i=1}^n$ from $\{\hat{\varepsilon}_t\}_{t=1}^n$ using the stationary bootstrap method as in Hsu et al. (2010) and compute

$$d_t^* = \hat{\mu} + \left(\hat{\Omega}^*\right)^{1/2} \hat{\varepsilon}_t^*, \quad \text{for} \quad t = 1, 2, \ldots, n,$$

where $\hat{\mu}$ is the estimator of μ in (2) under the null. This forms a bootstrap sample $\{d_t^*\}_{t=1}^n$. Then, we use it to obtain the GLR statistic, denoted by T_n^*, using the same approach as for T_n. Repeating this procedure s times, for example, $s = 600$, we obtain a bootstrap sample of the GLR statistic $\{T_n^{*(k)}\}_{k=1}^s$, which is then used to determine the quantiles of T_n under H_0. The p value of T_n is calculated by the relative frequency of $T_n^{*(k)}$ bigger than T_n. Note that the above procedure draws bootstrap samples from the residuals. Therefore, it is essentially a wild bootstrap method. Its consistency can be established similarly to that in Fan and Jiang (2005).

3.4 Stepwise GLR test

When the null hypothesis is rejected, it is often interesting to identify models that have contributed to the rejection. Following the ideas of the step-RC test of Romano and Wolf (2005) and the step-SPA test of Hsu et al. (2010), we extend the GLR test to a stepwise version which is termed as the step-GLR test. We expect that the step-GLR test is powerful because it inherits the advantages of the GLR test.

The idea is similar to the backward elimination method for variable selection or the case-deletion procedure for regression diagnostics in linear models. First, we calculate the GLR test statistic without using $d_{i,t}$:

$$T_{n,-i} = \frac{mn}{2}\left(\mathrm{RSS}_{0,-i} - \mathrm{RSS}_{1,-i}\right)/\mathrm{RSS}_{1,-i}, \tag{8}$$

for $i = 1, 2, \ldots, m$, where $\mathrm{RSS}_{0,-i} = \sum_{t=1}^{n} \left\| \left(\hat{\Omega}^*_{-i}\right)^{-1/2} (d_{t,-i} - \hat{\mu}_{-i}) \right\|^2$, $\mathrm{RSS}_{1,-i} = \sum_{t=1}^{n} \left\| \left(\hat{\Omega}^*_{-i}\right)^{-1/2} (d_{t,-i} - \bar{d}_{-i}) \right\|^2$, and "$-i$" means that the ith component of $d_{i,t}$ is not used. At a given significance level α_0, the step-GLR test proceeds as follows.

(i) If H_0 is rejected by T_n, we compute $\Delta T_{n,i} = T_n - T_{n,-i}$, for $i = 1, 2, \ldots, m$. The top model that contributes most to the rejection of H_0 is the one corresponding to the largest $\Delta T_{n,i}$, say $\Delta T_{n,p_1}$. Remove the top model p_1 from the collection of competitive models and denote $T_{n,-p_1}$ as T_{n-1}. Let $H_{0,-1}$ be the null hypothesis in (1) with $k \neq p_1$.

(ii) If $H_{0,-1}$ can be rejected by T_{n-1}, we compute $\Delta T_{n-1,i} = T_{n-1} - T_{n-1,-i}$, for $i = 1, 2, \ldots, m-1$. The top contributor to the rejection of $H_{0,-1}$ by T_{n-1} is the model with the largest $\Delta T_{n-1,i}$ and is denoted as p_2. Thus, the top contributed models to the rejection of the null hypotheses are models p_1 and p_2.

(iii) If $H_{0,-1}$ cannot be rejected by T_{n-1}, this stepwise procedure stops and the top contributed model is p_1.

(iv) Repeat (ii) and (iii) until all hypotheses are rejected.

If the above stepwise GLR testing procedure stops at the kth step, the top contributed models to the rejection of the null hypotheses are $p_1 \succ p_2 \succ \cdots \succ p_k$, where "$\succ$" represents a comparison between measurable contributions of two successive models.

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. We consider the same data generating process (DGP) as in Hansen (2005), because it leads to non-diagonal covariance matrix Ω, which may be better handled with our procedure. In the following, let us introduce the DGP. We follow the same notation as Hansen (2005).

Let the performance of the kth model relative to that of the benchmark model be measured by a loss function defined as

$$d_{k,t} = L\left(\xi_t, \delta_{0,t-h}\right) - L(\xi_t, \delta_{k,t-h}), \quad k = 1, 2, \ldots, m, \tag{9}$$

where $L(\cdot, \cdot)$ is a loss function with two arguments, ξ_t is a random variable representing the aspects of the decision problem that is unknown at the time when the decision is made, and $\delta_{k,t-h}$ is the kth decision rule that is made h periods in advance. In particular, $\delta_{0,t-h}$ is the decision based on the benchmark model. For the trading example in Sect. 2.1, $\delta_{k,t-1}$ equals 1(or -1) when a trader takes a long (or short) position at time $t-1$, and ξ_t is the return r_t of the underlying asset at the period t. The kth model (or trading rule) yields the profit $\pi_{k,t} = \delta_{k,t-h}\xi_t$. Therefore, the loss caused by the kth rule can be formulated as $L(\xi_t, \delta_{k,t-h}) = -\delta_{k,t-h}\xi_t$. Since we evaluate forecasts in terms of their expected loss given by

$$E(d_{k,t}) = E\left[L(\xi_t, \delta_{0,t-h})\right] - E\left[L(\xi_t, \delta_{k,t-h})\right], \quad \text{for} \quad k = 1, 2, \ldots, m,$$

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

Next, we generate $L_{k,t} = L(\xi_t, \delta_{k,t-h})$ from the model

$$L_{k,t} \sim iid \;\; N\left(\lambda_k/\sqrt{n}, \sigma_k^2\right), \quad \text{for} \quad k = 0, 1, \ldots, m, \quad \text{and} \quad t = 1, 2, \ldots, n.$$

The benchmark model has $\lambda_0 = 0$. When $\lambda_k > 0$, it means that the kth model is worse than the benchmark model; when $\lambda_k < 0$, the kth model is better than the benchmark. Naturally, $\{d_{k,t}\}$ in (9) are correlated across k and hence the covariance matrix Ω of d_t is not diagonal. As expected, our test procedure performs better than Hansen's SPA in this case.

The experiment is designed to control the value of λ_k, which is equivalent to choosing the poor and superior models. According to Hansen (2005), we set $\lambda_1 \leq 0$ and $\lambda_k \geq 0$, for $k = 2, \ldots, m$. Therefore, the first alternative ($k = 1$) defines whether the rejection probability corresponds to a type I error probability ($\lambda_1 = 0$) or power ($\lambda_1 < 0$). The poor models are those with mean values being evenly spaced between 0 and $\lambda_m = \Lambda_0$ (the worst model). This is to say that the values of λ_k are set as $\lambda_0 = 0$, $\lambda_1 = \Lambda_1$, and $\lambda_k = (k-1)\Lambda_0/(m-1)$, for $2 \leq k \leq m$. Following Hansen (2005), we set Λ_0 at 0, 1, 2, 5, and 10, respectively. For the alternative models, we set $\Lambda_1 = 0, -0.1, -0.2, -0.3, -0.4, and -0.5$ for our GLR test and $\Lambda_1 = 0, 1, 2, 3, 4, 5$ for Hansen's SPA. Therefore, $\lambda_1 = \Lambda_1$ defines the local alternative that is being analyzed. When $\Lambda_1 = 0$, the null hypothesis conforms with the alternative, and thus the alternative is not distinguishable from the null. As Λ_1 deviates away from 0 on the left, the alternative becomes further away from the null. It is worth pointing out that the values of Λ_1 for our test have a range much less than that for Hansen's SPA. Hence, it should be more difficult to differentiate the local alternatives from the null. The variance reflects the quality of the model. The smaller the variance, the better the model. Following Hansen (2005), we set $\sigma_k^2 = \exp(\arctan(\lambda_k))/2$, which indicates that the specification of the 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 results

We set $m = 100$ and $n = 200$ and 1,000. The bootstrap method is used to approximate the null distribution of T_n. Given a sample, we calculate the GLR statistic T_n and use $s = 600$ bootstrap replicates to get the p value of T_n according the procedure in Sect. 3.3. If the p value is less than the significance level, we reject the null hypothesis. To evaluate the power of the test, for each value of Λ_1 (i.e., each alternative), we compute the relative rejection frequencies of the null as the power of a test based on 1,000 simulations.

The results are reported in Tables 1 and 2 for the 5 and 10 % significance levels, respectively. For comparison, results from Hansen's SPA test are also reported. When $\Lambda_1 = 0$, in every panel in Tables 1 and 2, the relative frequencies of rejecting the null by the GLR test are all closer to the nominal sizes than those by Hansen's SPA test. Therefore, the type I error of the GLR test is better controlled than that of Hansen's SPA test.

Instead of using a relatively coarse measurement, $\Lambda_1 = 0, -1, -2, -3, -4$, and -5, in Hansen's test, we use $\Lambda_1 = 0, -0.1, -0.2, -0.3, -0.4, and -0.5$ for the GLR test. The results show that the power of GLR test approaches to one much faster than that of Hansen's SPA test. Regardless of the sizes and model specifications, the GLR test dominates Hansen's SPA test in terms of power.

In Table 1, the case where $\Lambda_0 = \Lambda_1 = 0$ refers to the situation that all the 100 inequalities are binding. It is the case discussed in White's (2000) RC test, where all the poor models are discarded. The relative rejection frequency is close to and less than the nominal levels. For example, when we set α at 5 %, the relative rejection frequency is 3 %, and when we set α at 10 %, the relative rejection frequency becomes 8.8 %. This may be caused by small sample sizes.

When the sample size n is increased from 200 to 1,000, the results become much better as shown in Table 2. When $\Lambda_0 = \Lambda_1 = 0$, the relative frequencies of rejecting H_0 are 4.9 % for $\alpha = 5$ and 9.0 % for $\alpha = 10$ %. Therefore, the GLR test has approximately correct size. It is also shown that Hansen's SPA test has approximately correct size as well. From Tables 1 and 2, one can see that, with the sample size increasing, the GLR test gains power at a faster speed for large samples than for small samples. For example, when the sample size is 1,000, the power of the GLR test is almost one in the case of $(\Lambda_0, \Lambda_1) = (0, -0.2)$. However, when the sample size is 200, the power of the GLR reaches one after the value of Λ_1 decreases to -0.5.

Hansen's SPA test cannot reject the null hypothesis when $\Lambda_1 = 1$, while the GLR test reaches power one 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 GLR test always outperforms Hansen's SPA test. Another important improvement is that our test is less conservative than Hansen's SPA test. For Hansen's SPA test, the probability of type I error shrinks fast as the value of Λ_0 increases. For example, it is only 0.007 when $(\Lambda_0, \Lambda_1) = (10, 0)$. The frequency value is far away from the nominal level 5 %. However, the probability of type I error for our GLR test is close to 5 %.

In summary, the above simulations demonstrate that the proposed GLR test is much more powerful and less conservative than Hansen's SPA test. This is possibly due to the nature of the GLR test and the correlation structure incorporated into the

Table 1 Relative frequencies of rejecting the null hypothesis under the null and alternative hypotheses ($m = 100$ and $n = 200$)

Level: $\alpha = 0.05$				Level: $\alpha = 0.10$			
Λ_1	GLR	Λ_1	SPA	Λ_1	GLR	Λ_1	SPA
Panel A: $\Lambda_0 = 0$							
0.0	0.030	0	0.060	0.0	0.088	0	0.11
−0.1	0.048	−1	0.074	−0.1	0.099	−1	0.129
−0.2	0.172	−2	0.280	−0.2	0.331	−2	0.389
−0.3	0.609	−3	0.764	−0.3	0.761	−3	0.845
−0.4	0.960	−4	0.979	−0.4	0.988	−4	0.99
−0.5	1	−5	1	−0.5	1	−5	1
Panel B: $\Lambda_0 = 1$							
0.0	0.052	0	0.022	0.0	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	0.920	−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: $\Lambda_0 = 2$							
0.0	0.048	0	0.012	0.0	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	0.690	−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: $\Lambda_0 = 5$							
0.0	0.054	0	0.007	0.0	0.107	0	0.013
−0.1	0.160	−1	0.031	−0.1	0.236	−1	0.054
−0.2	0.516	−2	0.273	−0.2	0.617	−2	0.370
−0.3	0.907	−3	0.787	−0.3	0.944	−3	0.860
−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: $\Lambda_0 = 10$							
0.0	0.020	0	0.007	0.0	0.081	0	0.015
−0.1	0.112	−1	0.043	−0.1	0.220	−1	0.073
−0.2	0.499	−2	0.340	−0.2	0.640	−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 Relative frequencies of rejecting the null hypothesis under the null and alternative hypotheses ($m = 100$ and $n = 1{,}000$)

Level: α=0.05				Level: $\alpha = 0.10$			
Λ_1	GLR	Λ_1	SPA	Λ_1	GLR	Λ_1	SPA
Panel A: $\Lambda_0 = 0$							
0.0	0.049	0	0.048	0.0	0.090	0	0.100
−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	0.390
−0.3	1	−3	0.762	−0.3	1	−3	0.840
−0.4	1	−4	0.980	−0.4	1	−4	0.990
−0.5	1	−5	1	−0.5	1	−5	1
Panel B: $\Lambda_0 = 1$							
0.0	0.070	0	0.017	0.0	0.226	0	0.039
−0.1	0.670	−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	0.740	−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: $\Lambda_0 = 2$							
0.0	0.067	0	0.009	0.0	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: $\Lambda_0 = 5$							
0.0	0.045	0	0.005	0.0	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: $\Lambda_0 = 10$							
0.0	0.017	0	0.005	0.0	0.098	0	0.005
−0.1	0.646	−1	0.042	−0.1	0.740	−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	0.980
−0.5	1	−5	1	−0.5	1	−5	1

GLR statistic. We have also conducted simulations in other settings, but the results are not reported here to save space. Our experience in simulations indicate that in large samples our method generally works better than Hansen's when the covariance matrix Ω is not diagonal and is favorably comparable to Hansen's when Ω is close to diagonal. However, in small samples both methods might not work satisfactorily for large dimensionality m.

5 An application of the GLR test

In this section, our aim is to compare the performance of GLR and SPA tests with a real example. To this end, we consider the linear relationship between the monthly return of the S&P 500 index denoted as Y, and a set of explanatory variables, which are the monthly money supply (X_1), federal funds rate (X_2), unemployment rate (X_3), earnings–price (E–P) ratio of the S&P 500 index (X_4), and dividend–price ratio of the S&P 500 index (X_5). These monthly data series are from February 2001 to October 2011.

The benchmark model is the model under the null hypothesis. The benchmark model is either the random walk model or a model specifying Y_t at a fixed value, which is chosen to be 0, -0.5, 0.5, -1, and 1 %, respectively.

The collection of alternative models contains 31 linear regressions of the following types:

$$
\begin{aligned}
&Y \sim X_i, \quad \text{for} \quad i = 1, 2, 3, 4, 5,\\
&Y \sim (X_i, X_j), \quad \text{for} \quad i \neq j, \quad \text{and} \quad i, j = 1, 2, 3, 4, 5,\\
&Y \sim (X_i, X_j, X_k), \quad \text{for} \quad i \neq j \neq k, \quad \text{and} \quad i, j, k = 1, 2, 3, 4, 5,\\
&Y \sim (X_i, X_j, X_k, X_l), \quad \text{for} \quad i \neq j \neq k \neq l, \quad \text{and} \quad i, j, k, l = 1, 2, 3, 4, 5,\\
&Y \sim (X_1, X_2, X_3, X_4, X_5)\,.
\end{aligned}
$$

We also include the lagged value(s) of Y_t as explanatory variable(s) in each of the aforementioned 31 models with the maximum lag order being six. When one lagged variable is included as a regressor, it must be Y_{t-1}; when two lagged variables are included as regressors, they must be Y_{t-1} and Y_{t-2}; and similarly, when five lagged variables are included as regressors, they must be Y_{t-i}, for $i = 1, 2, \ldots, 5$. Therefore, there are 223 models in total.

We examine the performance of the GLR and SPA tests through rolling samples. The rolling sample size is 24, which means that the sample contains two-year monthly data. All models are fitted to this sample and are used to forecast the one-step-ahead monthly return. The next sample for estimation is obtained by rolling the former sample forward by one step. Completion of this rolling sample procedure would result in a collection of 124 forecast values of monthly returns under each model. As expected, the collection of these forecast values is enough to evaluate forecasting performance of the two tests. When the benchmark model is a random walk process, the one-step-ahead forecast is simply the sample mean. The p values of GLR and SPA tests under benchmark models are given in Table 3.

Table 3 p values of the GLR and SPA tests

Null hypothesis	GLR	SPA
$H_0 : -0.5\,\%$	0.1940	0.3780
$H_0 : -1.0\,\%$	0.0380	0.0200
$H_0 : 0.0\,\%$	0.3580	0.7420
$H_0 : 0.5\,\%$	0.5340	0.8400
$H_0 : 1.0\,\%$	0.2300	0.4900
H_0 : random walk	0.3340	0.6200

Table 4 Mean and standard deviation of the performance measures for the top five contributed models in the GLR test

Model	Performance measures	
	Mean	SD
$Y_t \sim (Y_{t-1}, Y_{t-2})$	0.0036	0.0127
$Y_t \sim (Y_{t-1})$	0.0037	0.0113
$Y_t \sim (Y_{t-1}, Y_{t-2}, Y_{t-3})$	0.0038	0.0135
$Y_t \sim (X_2, X_4, Y_{t-1})$	0.0028	0.0112
$Y_t \sim (X_2, X_3, X_5, Y_{t-1})$	0.0006	0.0182

Table 5 Mean and standard deviation of the performance measures for the contributed models in the SPA test

Model	Performance measures	
	Mean	SD
$Y_t \sim (X_2, X_4, X_5)$	0.0055	0.0143
$Y_t \sim (X_2, X_3, X_4, X_5, Y_{t-1})$	0.0054	0.0159
$Y_t \sim (X_2, X_3, X_4, X_5)$	0.0054	0.0142
$Y_t \sim (X_2, X_4, X_5, Y_{t-1})$	0.0052	0.0152

Under the null hypothesis that no alternative is better than the above random walk process, both tests cannot reject the null hypothesis. This finding is consistent with a general belief of financial analysts about technical analysis. Most analysts tend to believe that most models cannot outperform the moving average method in terms of forecasting performance.

When the benchmark model specifies $-1\,\%$ as the monthly return, both the GLR and SPA tests reject the null hypothesis at the $5\,\%$ significance level. In this situation, we conduct the step-GLR test and find 13 models which contribute to the rejection of null hypothesis. The top five contributed models, as well as their mean and standard deviation of performance measures, are presented in Table 4. By conducting the step-SPA test, we find that four models have contributed to the rejection of null hypothesis. These four contributed models, as well as their mean and standard deviation of performance measures, are presented in Table 5.

In terms of the GLR test, the top three contributed models are all autoregressive models with lag orders up to three, and none of the exogenous variables contributed to the rejection of the null hypothesis. In contrast, the four contributed models for the SPA test in rejecting the null hypothesis include only one lagged variable, together

Table 6 Sensitivity analysis of the p value of the GLR test with respect to the number of components

Null hypothesis	$\hat{v}^2$		$\hat{v}^2(1+20\,\%)$		$\hat{v}^2(1-20\,\%)$	
	p value	Factors	p value	Factors	p value	Factors
$H_0 : -0.5\,\%$	0.1940	19	0.1920	17	0.1920	20
$H_0 : -1.0\,\%$	0.0380	17	0.0380	16	0.0380	20
$H_0 : 0.0\,\%$	0.3580	20	0.3580	17	0.3580	21
$H_0 : 0.5\,\%$	0.5340	18	0.5380	16	0.5320	20
$H_0 : 1.0\,\%$	0.2300	17	0.2260	16	0.2320	19
H_0 : randomwalk	0.3340	20	0.3420	18	0.3360	22

with X_2, X_3, X_3 and X_4 as explanatory variables. More specifically, it can be clearly seen that only Y_{t-1} is included in all the top five contributed models under the GLR test and explanatory variables X_2, X_4 and X_5 are contained in all the four contributed models under the SPA test.

Finally, as the GLR test is constructed based on the principal components of the variance–covariance matrix Ω, one might be interested in analyzing the sensitivity of this test with respect to the number of components. The number of components used by the test is determined by $\hat{v}^2$. When this value is used, the number of factors selected under each null hypothesis is presented in the third column of Table 6. When the value of $\hat{v}^2$ is increased or decreased by 20 %, the number of factors selected under each null hypothesis is given in the fifth or seventh column, respectively. It is clear to see that the derived p values are not sensitive to a 20 % change in the threshold value $\hat{v}^2$ for choosing the number of latent variables. Therefore, this test is quite robust against to the choice of $\hat{v}^2$ for determining the number of latent variables.

6 Conclusion

We have proposed the GLR and step-GLR tests to check the superior predictive ability of a benchmark model against a large group of alternative models. To model the performance of all alternative models relative to a benchmark model, we have explicitly approximated the covariance matrix by model (6). This method is applicable when the number of alternative models exceeds the sample size. A Monte Carlo simulation study demonstrates that the power of GLR test is much higher than that of Hansen's SPA test. Our simulation results also show that the GLR test is less conservative than Hansen's SPA test. In addition to examining the predictive ability of technical trading rules and empirical models, our test is also applicable to other multiple testing problems for the predictive ability of various econometric models as well as the performance of mutual funds and corporate managers. Our method is expected to work well in large samples when the performance measures are correlated cross-individual models, funds, or managers. However, if the sample size is small or the stationarity condition for $\{d_t\}$ is not valid, the GLR test would not be expected to perform well. Further work such as

deriving the asymptotic null distribution and the theoretical power of T_n is warranted and is left as a future research topic.

Acknowledgments We extend our sincere thanks to the editor, Subal Kumbhakar, Christopher Parmeter, and two referees for their very insightful comments and suggestions that have substantially improved our paper. Cai's research was supported, in part, by the National Nature Science Foundation of China Grants #71131008 (Key Project) and #70971113. Jiang's research was supported, in part, by the NSF grant DMS-09-06482 and the NSFC Grant #71361010. Xibin Zhang's research was supported under the Australian Research Council's *Discovery Projects* funding scheme (Project Nos. DP1095838 and DP130104229).

References

Birbaum A, Johnstone IM, Nadler B, Paul D (2013) Minimax bounds for sparse PCA with noisy high-dimensional data. Ann Stat 41(3):1055–1084

Blume L, Easley D, O'Hara M (1994) Market statistics and technical analysis: the role of volume. J Financ 49(1):153–181

Brock W, Lakonishok J, LeBaron B (1992) Simple technical trading rules and the stochastic properties of stock returns. J Financ 47(5):1731–1764

Brown SJ, Goetzmann WN, Kumar A (1998) The Dow theory: William Peter Hamilton's track record reconsidered. J Financ 53(4):1311–1333

Cai Z, Fan J, Yao Q (2000) Functional-coefficient regression models for nonlinear time series. J Am Stat Assoc 95(451):941–956

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):353–367

Fan J, Jiang J (2005) Nonparametric inferences for additive models. J Am Stat Assoc 100(471):890–907

Fan J, Jiang J (2007) Nonparametric inference with generalized likelihood ratio tests. Test 16(3):409–444

Fan J, Zhang C, Zhang J (2001) Generalized likelihood ratio statistics and Wilks phenomenon. Ann Stat 29(1):153–193

Gencay R (1998) The predictability of security returns with simple technical trading rules. J Empir Financ 5(4):347–359

Hansen PR (2003) Asymptotic tests of composite hypotheses. Working Paper, Stanford University. http://www.stanford.edu/people/peter.hansen

Hansen PR (2005) A test for superior predictive ability. J Bus Econ Stat 23(4):365–380

Hsu PH, Hsu YC, Kuan CM (2010) Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. J Empir Financ 17(3):471–484

Jiang J, Zhou H, Jiang X, Peng J (2007) Generalized likelihood ratio tests for the structure of semiparametric additive models. Can J Stat 35(3):381–398

Liu Y, Hayes DN, Nobel A, Marron JS (2008) Statistical significance of clustering for high-dimension, low-sample size data. J Am Stat Assoc 103(483):1281–1293

Lo AW, MacKinlay AC (1990) When are contrarian profits due to stock market overreaction? Rev Financ Stud 3(2):175–205

Lo AW, Mamaysky H, Wang J (2000) Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. J Financ 55(4):1705–1770

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

Savin G, Weller P, Zvingelis J (2007) The predictive power of "head-and-shoulders" price patterns in the US stock market. J Financ Econ 5(2):243–265

Sullivan R, Timmermann A, White H (1999) Data-snooping, technical trading rule performance, and the bootstrap. J Financ 54(5):1647–1691

Sweeney RJ (1988) Some new filter rule tests: methods and results. J Financ Quant Anal 23(3):285–300

West KD (1996) Asymptotic inference about predictive ability. Econometrica 64(5):1067–1084

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

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9(1):60–62