Using Bootstrap to Test Mean-Variance Efficiency of a Given Portfolio*

Pin-Huang Chou
Department of Finance
National Central University
Chung Li, Taiwan 32054, ROC

September 4, 1996

Abstract

This paper proposes tests of unconditional mean-variance efficiency using bootstrap method that does not rely on specific distributional assumptions. We reject the mean-variance efficiency of the CRSP value-weighted stock index for five of the seven consecutive ten-year subperiods from 1926 to 1993, whereas the F-test of Gibbons, Ross, and Shanken (GRS, 1989) only rejects two of the seven subperiods. A further examination of the size of the tests reveals that, under various alternative distributional specifications for the error terms, the GRS test tends to over-reject the null hypothesis, while the bootstrap test has sizes close to the nominal levels. However, the GRS test has a slightly higher power than the bootstrap test.

JEL Classification: C13, C53, G14.
Key Words: Bootstrap hypothesis test, mean-variance efficiency, elliptical distributions, size and power.

Comments Welcome

*Very first draft. Tel: 886-3-4227151 ext. 6270. Fax: 886-3-4252961. E-mail address: chou@fin.mgt.ncu.edu.tw or chou@im.mgt.ncu.edu.tw.

1 Introduction

The Sharpe-Lintner Capital Asset Pricing Model (CAPM) is perhaps the most important asset pricing model in financial economics, and numerous efforts have been devoted to tests of the model during the past two decades. The CAPM is essentially untestable because the market portfolio is unobservable [Roll (1977)]. Hence, recently researchers have instead focused on tests of the mean-variance efficiency of the underlying portfolio under consideration (e.g., stock market index).

Under iid multivariate normality assumptions, Gibbons, Ross, and Shanken (GRS, 1989) show that the mean-variance efficiency of a portfolio, in the case where there exists a risk-free asset, can be tested with a Hotelling's T^2 statistic that has an exact F distribution. Based on the same assumptions, Harvey and Zhou (1991) investigate the Sharpe-Lintner CAPM in the Bayesian framework, and evaluate the hypothesis implied by the CAPM using Monte Carlo numerical integration. However, multivariate normality is not necessary to ensure the validity of the mean-variance framework, upon which the CAPM is derived. The largest class of distributions known thus far that validates the mean-variance framework is the class of elliptical distributions [see, e.g., Ingersoll (1987)]. Noting this fact, Zhou (1993) further proposes an approach based on Monte Carlo simulation to test the CAPM that allows the distribution of the (excess) asset returns and the market model error terms to be elliptical. The members of elliptical distributions include multivariate normal, mixture of multivariate normal, multivariate t, multivariate stable, and so on. Zhou (1993) finds that the GRS test is robust when excess returns follow a market model whose error term is not normally distributed but still retains the property of ellipticity. However, when the returns are elliptically, but not normally, distributed, in which case the market model error term no long has an elliptical distribution, the GRS test tends to over-reject the null hypothesis of portfolio efficiency too often. That is, the size of the GRS test is higher than the nominal significance level. When alternative distributional specifications are adopted for the asset returns, the mean-variance efficiency of the CRSP value-weighted index can no longer be rejected. It is worth noting that his analysis

is exact in that the statistical inference does not depend on the sample size. However, the inference is legitimate only if the underlying distributional specification is correct and provided that the random number generator for the assumed distribution is available. In addition, the distributions of the asset returns (or error terms) are never known in the real world. Hence, although Zhou's (1993) method extends the test of mean-variance efficiency to the case of elliptical distributions, his test requires that the returns follow a specific distribution.

MacKinlay and Richardson (1991) propose a generalized method of moments (GMM) approach to test portfolio efficiency that requires much weaker distributional assumptions. More specifically, their GMM approach allows the disturbance term can be serially dependent and conditionally heteroscedastic. Their test, however, is asymptotic and relies on the availability of large samples.

In this paper, we adopt a different approach, namely the bootstrap approach, to test the mean-variance efficiency of a portfolio that does not assume a specific distribution. Bootstrap method, originally proposed by Efron (1979), is basically a computer-based method that allows one to analyze the distribution of an estimator or test statistic of interest without needing to perform complicated analytical computations. In particular, bootstrap methods usually yield estimates of a higher accuracy than those obtained from classical asymptotic theories. Though the consistency of the bootstrap estimators and tests also often depends on large sample sizes, bootstrap estimators and tests generally have better performances than those obtained from canonical asymptotic theories in finite samples. We introduce the concept of bootstrap in more detail in section 3.

Given the advantages of the bootstrap methods, it may be somehow surprising that not many works have adopted bootstrap methods in the finance area [see Jeong and Maddala (1993) for a brief list of the bootstrap applications in finance]. Recently, however, there has been an increasing interest on the application of the bootstrap methods to empirical studies in finance. For example, Chatterjee and Pari (1990) use bootstrap to study the number of factors in the arbitrage pricing model. Goetzmann and Jorion (1993) use bootstrap to examine the predictive power of dividend yields.

More recently, Kramer (1995) proposes the use of bootstrap for event studies. In fact, hypothesis testing based on bootstrap methods is still an area not well developed yet. For complicated models, sometimes bootstrapping the data may be very difficult and extremely computer intensive. This may help explain, at least partially, why not many empirical studies adopt bootstrap methods.

The reminder of the paper is organized as follows. Section 2 introduces the tests of mean-variance efficiency and roughly reviews the commonly adopted statistics and their distributions. Section 3 introduces the concept of bootstrap, and proposes a bootstrap test of portfolio efficiency. Section 4 contains the data and empirical results. To further investigate the performance of our bootstrap test under various alternative distributional assumptions, we perform several Monte Carlo experiments to compare the size and power of the bootstrap test and the GRS F-test in section 5. The last section briefly concludes the paper.

2 Tests of mean-variance efficiency

Consider the following system of excess-return market model:

$$r_{it} = \alpha_i + \beta_i r_{pt} + \varepsilon_{it}, \qquad i = 1, ..., N; \;\; t = 1, ..., T, \tag{1}$$

where

r_{it}	=	the excess return on asset i in period t,
r_{pt}	=	the excess return on portfolio p in period t, whose efficiency is being tested;
ε_{it}	=	the disturbance term for asset i in period t.

N refers to the number of assets under consideration, and T is the number of time series observations. The error terms are assumed to be independent and identically distributed with mean zero and a constant covariance matrix, i.e.,

$$E(\varepsilon_{is}, \varepsilon_{jt}) = \begin{cases} \sigma_{ij} & \text{if } s = t \\ 0 & \text{otherwise.} \end{cases}$$

The above model can be rewritten in a more compact form as follows:

$$R_t = \alpha + \beta r_{pt} + \varepsilon_t, \qquad \varepsilon_t \sim P(0, \Sigma), \tag{2}$$

where R_t is the $(N \times 1)$ vector of excess returns; $\alpha = (\alpha_1, ..., \alpha_N)'$; $\beta = (\beta_1, ..., \beta_N)'$; and $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$. The error term ε_t is assumed to be *iid* with unknown distribution function $P(0, \Sigma)$, whose mean is zero and the covariance matrix is Σ.

The Sharpe-Lintner CAPM asserts that if a portfolio p is mean-variance efficient, then the following linear relationship holds:

$$E(R_t) = \beta E(r_{pt}).$$

The relationship is referred to as the *security market line* (SML) in the literature. Hence, this restriction implies a testable joint hypothesis for testing the mean-variance efficiency of the underlying portfolio p [GRS (1989)]:

$$H_0 : \alpha = 0.$$

That is, if we regress the excess asset returns on those of an efficient portfolio, the resulting intercepts should be indistinguishable from zero.

It should be borne in mind, however, that the CAPM is derived based on the mean-variance framework. The class of distributions that validates the mean-variance approach known so far is the elliptical distributions, which include multivariate normal, multivariate t, mixture of multivariate normal, and so on [see discussion in Zhou (1993)]. Hence, it is not necessary to impose multivariate normality assumption. Though the multivariate normality of asset returns has been rejected in several studies [Zhou (1993), Richardson and Smith (1994) and Harvey (1995)], it does not necessarily invalidate the mean-variance framework and the CAPM, provided that the ellipticity of asset returns still holds.

The multivariate regression model (2) can be estimated by OLS, and the OLS estimate of the intercepts $\hat{\alpha}$ has a multivariate normal distribution asymptotically:

$$\hat{\alpha} \overset{A}{\sim} \mathcal{N}(\alpha, h\Sigma),$$

where $h = \frac{1}{T}(1 + \hat{\theta}_p^2)$, and $\hat{\theta}_p = \bar{r}_p / s_p$ is the observed Sharpe measure of the portfolio p; $\bar{r}_p$ and s_p are, respectively, the sample mean and sample standard deviation of r_p.[1]

[1] The sample standard deviation is not adjusted for degrees of freedom. Specifically, $s_p = \sqrt{\frac{1}{T}\sum (r_{pt} - \bar{r}_p)^2}$.

Asymptotically, the null hypothesis can be tested by a Wald statistic that has an asymptotic chi-square distribution with N degrees of freedom, i.e.,

$$h^{-1}\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha} \overset{a}{\sim} \chi^2_N$$

The asymptotic Wald test based on chi-square distribution, however, tends to over-reject to the null hypothesis in finite samples [see Table 1 of GRS (1989)].

If P is multivariate normal, however, the zero-intercept hypothesis can be tested by a statistic whose finite-sample distribution is known. Specifically, by exploiting the Hotelling's T^2 statistic, Gibbons, Ross, and Shanken (1989) show that the following statistic has a noncentral F distribution with degrees of freedom N and $T-N-1$:

$$GRS \equiv \frac{T(T-N-1)}{N(T-2)} h^{-1}\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha} \sim F_{N,T-N-1}(\lambda), \tag{3}$$

where $\lambda = h^{-1}\alpha'\Sigma^{-1}\alpha$ is the noncentrality parameter of the F distribution. Under the null hypothesis that $\alpha = 0$, the distribution reduces to the central F distribution. Hence, critical value and p value of the statistic are easily calculated.

MacKinlay (1987), however, documents that the ability of the GRS test to distinguish the CAPM against other asset pricing models (e.g., multifactor models) is poor. Affleck-Graves and McDonald (1989) find that when the sample nonnormalities are severe, the size and power of the GRS test can be seriously mis-stated.

The GMM approach of Richardson and Smith (1991) is robust against departure from normality, and allows the error term to be serially correlated and conditionally heteroscedastic. Like the asymptotic Wald test, the GMM tests relies on large samples to ensure the convergence of the test statistic to a normal or chi-square distribution. Zhou (1993) proposes a Monte Carlo simulation method that allows the asset returns and the market model disturbance terms to be elliptically distributed. Under a pre-specified distribution for the returns or the error terms, he can calculate the exact p value for the test of zero-intercept hypothesis. However, a weakness of his approach is that the parameters for the alternative distributions, such as the degree-of-freedom parameter in the multivariate t distribution and the mixing-probability parameter in the mixture-normal distribution, are unknown, and are obtained by trial and error in his paper.

As an approach complementing Zhou's method, our bootstrap approach does not have to assume a pre-specified distribution for the error terms. However, our bootstrap-based inference is not exact because no specific distribution is imposed, as in the study of GRS (1989) who impose normality or in Zhou (1989) who assumes several alternative elliptical distributions. Like most asymptotic test statistics, bootstrap estimators also require large samples to retain some large sample properties such as consistency. However, in finite samples the bootstrap methods often provide better approximation than those obtained from asymptotic theories. In the real world mostly researchers only have limited data and it is general difficult to identify how large a sample is a large sample. Hence, bootstrap methods appear to be a natural alternative that may offer researchers more information about the data at hand.

3 Bootstrap hypothesis test

The bootstrap, introduced by Efron (1979), is a computer-intensive method for estimating the distribution of an estimator or test statistic by resampling the data at hand. It treats the data as if they were the population. In fact, under mild regularity conditions, the bootstrap generally yields an approximation to the sampling distribution of an estimator or test statistic that is at least as accurate as the approximation obtained from traditional first-order asymptotic theory [see Horowitz (1995)]. In many instances the sampling distribution of a statistic may not be analytically available, while the bootstrap, on the other hand, obtains the sampling distribution of the statistic via repeatedly resampling from the sample at hand. Below we first introduce the concept of bootstrap, and then outline the procedure for testing the zero-intercept hypothesis of the market model.

3.1 Concept of bootstrap

Introduction of bootstrap has been available in several studies [Babu and Rao (1993), Efron and Tibshirani (1993), and Hall (1994)], and its application to econometrics

has also been discussed in some studies [Horowitz (1995), Jeong and Maddala (1993), and Vinod (1993)]. Here we shall only briefly introduce the concept of bootstrap.

Let $x = (x_1, ..., x_T)$ denote a random sample from an unknown distribution P. Suppose we are interested in the sampling distribution of a statistic $H_T(x)$. The statistic is used to infer a parameter of the population, denoted as $\theta(P)$. For example, if $\theta(P)$ is the mean of a random variable with distribution P, then $H_T(x)$ can be set as the sample mean. If P is normal, then $H_T(x)$ is also normally distributed, in which case statistical inference is simple and easy. If P is not normal, then the sample mean will not be normally distributed, at least in finite samples. In this case, the classical large-sample approximation may not be able to give a satisfactory analytical answer.

The bootstrap method may start by approximating the distribution P with its empirical counterpart $\hat{P}(y)$:

$$\hat{P}(y) = Pr(X \leq y) = \frac{1}{T} \sum_{t=1}^{T} I(x_i \leq y)$$

where $I(A)$ is an indicator function which takes value 1 if A is true, and 0 otherwise. This amounts to assign equal probability to each of the observation in the sample x:

$$\hat{P}(X = x_i) = \frac{1}{T}, \qquad i = 1, ..., T.$$

A bootstrap sample $z^* = (y_1^*, ..., y_T^*)$ is then constructed by randomly drawing from $\hat{P}(y)$, the empirical distribution. That is, one draws a sample z^* from x with replacement.

To analyze the sampling distribution of $H_T(x)$, one may, instead of analytically deriving the distribution, replicate the following procedure a large number of times to obtain a sample of $H_T(z^*)$'s.

1. Generate a bootstrap sample z^* from $\hat{P}(y)$.

2. Calculate $H_T(z^*)$.

Suppose one replicates the experiment k times. Thus, a sample of k observations for $H_T(z^*)$'s is obtained: $(H_T^{(1)}(z^*), ..., H_T^{(j)}(z^*), ..., H_T^{(k)}(z^*))$, where $H_T^{(j)}(z^*)$ refers

to the statistic calculated based on the jth bootstrap sample. The percentiles of the sample can be used to construct confidence interval for the parameter. Hypothesis testing can also be done by checking if a certain level of confidence interval contains the value of the statistic implied by the null hypothesis. For example, to test if the population mean is significantly different from zero, one may check if the 95% confidence interval, constructed based on the bootstrap statistics, covers the value zero.

In a bootstrap framework, there are two approaches for testing a hypothesis, on based on confidence intervals, and the other direct hypothesis testing. Direct bootstrap hypothesis test requires drawing a sample of the statistic of interest from an empirical distribution under the restrictions specified the null hypothesis, and then the achieved significance level (ASL), or the p value, can be calculated by comparing the observed statistic based on the original data and the sample of the statistic based on bootstrap samples. Usually a hypothesis can be tested by constructing an appropriate confidence set. However, direct bootstrap hypothesis test is sometimes easier when constructing a confidence set is complicated. Moreover, the bootstrap tests obtained directly may be better because they usually take account of the special nature of the hypothesis. See, e.g., Shao and Tu (1995, pages 175-188) for an in-depth explanation.

To explain the procedure of direct bootstrap test, consider the following example on test of population mean. Let $x_1, ..., x_T$ be an iid sample from P with mean θ, which may be univariate or multi-dimensional. Shao and Tu (1995, pages 180-181) suggest the following "standardized" statistic to test a hypothesis of the form: $H_0 : \theta = \theta_0$:

$$H_T = ||\sqrt{T}\hat{\Omega}^{-1/2}(\bar{x} - \theta_0)||$$

where $||\cdot||$ denotes a norm, and $\bar{x}$ and $\hat{\Omega}$ are, respectively, the estimates of the mean θ and the covariance matrix Ω. Suppose a Euclidean norm is used. If θ is univariate, the test statistic H_T is the usual t statistic. If θ is multivariate, the test statistic H_T is just the common Wald statistic, i.e., $H_T = T(\bar{x} - \theta_0)'\hat{\Omega}^{-1}(\bar{x} - \theta_0)$. To test the null hypothesis, one first estimates H_T, and then a bootstrap sample of the statistic H_T^* has to be drawn from the distribution subject to the restriction specified by the null

hypothesis. To do so, define a new random variable: $y_i = x_i - \bar{x} + \theta_0, i = 1, ..., T$. The empirical distribution under the null hypothesis can be specified as:

$$\hat{P}(Y = y_i) = \frac{1}{T}, \qquad i = 1, ..., T. \tag{4}$$

Clearly, y_i's have the same distribution as x_i's, except that the mean of the distribution is restricted to be θ_0. Then H_T^* can be calculated based on a bootstrap drawing from $\hat{P}(y_i)$. With a collection of H_T^*'s, the critical values or p value of the test statistic H_T can be calculated. For example, if θ is multivariate, then we can calculate the percentage of H_T^*'s that are larger than H_T, which is the achieved significance level, i.e., the p value. Specifically, the procedure for direct bootstrap test is as follows:

1. Calculate the statistic based on the data:

   $$H_T = T(\bar{x} - \theta_0)' \hat{\Omega}^{-1} (\bar{x} - \theta_0)$$

   Construct an empirical distribution that puts equal probability to each of the transformed variable $y_i = x_i - \bar{x} + \theta_0, i = 1, ..., T$; denote the distribution $\hat{P}(y)$.

2. Repeat the following steps k times.

   (a) Draw a bootstrap sample $y^* = (y_1^*, ..., y_T^*)$ from $\hat{P}(y)$. Calculate the mean and covariance of the bootstrap sample, $\bar{y}^*$ and $\hat{\Omega}^*$.

   (b) Calculate

   $$H_T^* = T(\bar{y}^* - \theta_0)' \hat{\Omega}^{*-1} (\bar{y}^* - \theta_0)$$

3. Calculate the percentage of H_T^*'s that are greater than H_T, the bootstrap achieved significance level.

The readers are referred to Shao and Tu (1995) for detailed discussion on the validity of the procedure on hypothesis testing. Our design of bootstrap test of portfolio efficiency is based on the same idea.

3.2 Bootstrap test of portfolio efficiency

Following the example given in last section, we adopt the following quadratic Wald-type statistic to test our null hypothesis of zero intercepts:

$$H_T = h^{-1}\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha}$$

The procedure for bootstrap hypothesis test is described below.

1. Estimate the model using OLS. Let $\hat{\alpha}$ and $\hat{\varepsilon}_t$ denote the OLS intercept estimate and the OLS residual, respectively. Denote $\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'$. Calculate the following:

   $$H_T = h^{-1}\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha}$$

   Calculate the restricted estimate for β: $\beta_{res}' = \sum_{t=1}^{T}(r_p'r_p)^{-1}r_p'R_t$, where $r_p' = (r_{p1}, ..., r_{pT})$.

2. Repeat the following steps 100,000 times.

   (a) Draw a bootstrap sample $\{\varepsilon_t^*\}$ from $\{\hat{\varepsilon}_t\}$. Let $R_t^* = \beta_{res}r_{pt} + \varepsilon_t^*$, $t = 1, ..., T$.

   (b) Calculate the OLS estimates for model parameters based on $(R_1^*, ..., R_T^*)$ and r_{pt}'s. Denote the estimates $\hat{\alpha}^*$ and $\hat{\Sigma}^*$.

   (c) Calculate

   $$H_T^* = h^{-1}\hat{\alpha}^{*\prime}\hat{\Sigma}^{*-1}\hat{\alpha}^*$$

3. Calculate the percentage of H_T^* that is greater than H_T, the bootstrap achieved significance level.

The achieved significance level is parallel to the p-value in classical statistical analysis.

4 Data and empirical results

In this study, we investigate the mean-variance efficiency of the CRSP (Center for the Research of Stock Prices) value-weighted index with respect to the ten CRSP size decile portfolios from 1926 to 1993 using monthly data. All returns are calculated in excess of the average one-month T-bill rate, obtained from the Fama bond file of the CRSP data base.

Table 1 shows that at the 5% significance level the GRS test rejects the efficiency of the CRSP value-weighted index only for two of the seven subperiods, whereas both the asymptotic Wald chi-square tests and the bootstrap test reject five of the seven subperiods. The results of the Wald test is expected because GRS (1989) and several studies have pointed out that in finite samples the Wald chi-square test tends to reject the null hypothesis too often, while generally the F test seems to be more conservative. For the full sample (26/4 - 93/12), however, all tests reject the efficiency of the value-weighted index at the 1% significance level.

The results of bootstrap test are very similar to those of the Wald test. However, the p values of the bootstrap test are always smaller than those of the Wald test, suggesting that the bootstrap may have also over-rejected the null hypothesis. If the data were really generated from multivariate normal, then both the Wald and bootstrap tests must have over-rejected the model. However, since the multivariate normality of the asset returns as well as the OLS market model residuals have been rejected by several studies [e.g., Richardson and Smith (1993)] and here each subperiod only contains about 120 observations, it is hard to conclude that bootstrap test has over-rejected the null hypothesis. To further investigate the performance of bootstrap test under various distributional assumptions, in next section we perform several experiments to compare the size and power of the bootstrap test and the GRS test. Our experiments show that bootstrap test has very reasonable size and power (see next section for detail), while the GRS test tends to be reject the null hypothesis too often, especially when the model error terms are generated from empirical distributions based on the OLS market model residuals. Hence, we conclude

that overall the mean-variance efficiency of the CRSP value-weighted index is rejected.

5 Size and power of the bootstrap test

This section examines and compares the size and power of the GRS test and the bootstrap test under various alternative distributions. We consider four different distributions for the error term: multivariate normal, mixture of multivariate normal, multivariate t, and empirical distributions constructed based on the OLS market model residuals. Use of the first three distributions is intended to investigate performances of the bootstrap test under elliptical errors. Since in theory the error terms are not required to follow normal or elliptical distributions, the last scenario based on empirical distributions allows us to investigate the robustness of the bootstrap and GRS tests against departures from assumed distributional structure.

The mixture of normal distributions considered in this paper is of the following form:

$$\varepsilon_t^* \sim w\mathcal{N}(0,\hat{\Sigma}) + (1-w)\mathcal{N}(0,\gamma\hat{\Sigma}),$$

where w $(0 \leq w \leq 1)$ is the mixing-probability parameter, and γ is a scale parameter. This is a mixture of two normal distributions. In this paper, we set $w = 0.7$ and $\gamma = 10$. The following two steps show how one can draw an observation ε_t^* from the mixture normals:

1. Draw u_t from uniform random number generator ranging between 0 and 1.
2. If $u_t \leq k$, draw $\varepsilon_t^* \sim \mathcal{N}(0,\hat{\Sigma})$. Otherwise, draw $u_t \leq k$, draw $\varepsilon_t^* \sim \mathcal{N}(0,\gamma\hat{\Sigma})$.

There are some different forms of multivariate t distributions. Here we follow the specification adopted in Zhou (1993). Let $\hat{\Sigma} = LL'$, then the multivariate t with ν degrees of freedom is generated according to the following:

1. Independently draw Z_t from an N-variate standard normal distribution: $\mathcal{N}(0, I_N)$ and c from chi-square distribution with ν degrees of freedom (χ^2_ν).

2. Set $\varepsilon_t^* = \left(\frac{c}{\nu}\right)^{-1/2} LZ_t$.

That is, $L^{-1}\varepsilon_t^*$ has a multivariate t distribution with ν degrees of freedom. Below we describe our experiment design on examining the size and power of the bootstrap and GRS tests.

5.1 Size

To compare the size of the GRS test and bootstrap test under various distributional specifications for the error terms, we repeat the following procedure 1,000 times for multivariate normal, mixture of multivariate normal, and multivariate t with 2, 10, and 20 degrees of freedom. The OLS estimates of β and Σ over the period 76/1 - 85/12 are used the model parameters. The intercepts are set to be zero to investigate the size of the tests.

1. Let $\hat{\beta}$ and $\hat{\Sigma}$ be the OLS estimate of β and Σ over the period 76/1 - 85/12. Then generate the returns series $\{R_t\}$ from the market model by imposing the intercepts to be zero:

$$R_t^* = \hat{\beta} r_{pt} + \varepsilon_t^*, \qquad t = 1, ..., T,$$

   where ε_t^* is generated from the distribution under consideration.

2. Calculate the p value of the GRS F-test based on $\{R_t^*, r_{pt}\}$.

3. Perform the bootstrap procedure described in section 3.2. Since the experiment is very time consuming, we only draw 1,000 bootstrap samples. Calculate the percentage of T_n^* that is greater than T_n, which is the bootstrap achieved significance level, the p-value based on bootstrap.

Based on the 1,000 replications, we calculate the rejection rates of the GRS test and the bootstrap test under 1%, 5%, and 10% significance levels. The results are reported in Table 2. Table 2 indicates that both the GRS and bootstrap tests have

sizes that are close to their nominal levels under different elliptical distribution assumptions for the error terms. The GRS test, however, performs slightly better than the bootstrap in almost all cases. The results confirm Zhou's (1993) findings that the GRS test is robust when error terms deviate from normality, but still retain the ellipticity property.

The last scenario considers the case where the error terms are generated from empirical distributions, constructed based on the market model residuals of the real data from 1926 to 1993. The purpose is to investigate how both tests perform facing real data. Similar to the procedure described above, the experiment is performed for each of the seven ten-year subperiods, except the first step is slightly different:

1. Let $\hat{\beta}$ and $\hat{\Sigma}$ be the OLS estimates of β and Σ, and $\{\hat{\varepsilon}_t\}$ the OLS market model residuals over the subperiod. Then generate the returns series $\{R_t^*\}$ from the restricted market model:

$$R_t^* = \hat{\beta} r_{pt} + \varepsilon_t^*, \qquad t = 1, ..., T,$$

where ε_t^* is generated from the empirical distribution constructed of $\{\hat{\varepsilon}_t\}$.

The results on size of the GRS and bootstrap tests are reported in Table 3. Table 3 indicates that the GRS test tends to over-reject the null hypothesis with respect to their nominal levels. The bootstrap test, on the contrary, has sizes much closer to their nominal levels in all cases.

In summary, our experiments show that both the GRS and bootstrap tests have reasonable sizes with respect to the usual nominal levels. The GRS test is robust when the error terms follow elliptical distributions, but tends to reject the null hypothesis too often when real data are used. The bootstrap test, on the other hand, is robust against different distributional specifications for the error terms, especially when the error terms are generated from the real data. The results suggest that the bootstrap test does not suffer from the potential problem of over-rejecting the null hypothesis. Hence, with the simulation results we are confident to reject the mean-variance efficiency of the CRSP value-weighted index.

5.2 Power

To investigate the power of the tests, we need to specify the alternative hypothesis. Following MacKinlay (1987), we consider an alternative hypothesis of the form: $H_A = \alpha = (1-\beta)\delta$. Originally, this alternative hypothesis is designed to evaluate the power of the GRS test in the case where the riskfree rate is measured with error. MacKinlay (1987) restricts the value of δ to be between 0 and 0.01. GRS (1989) show that the GRS F-statistic can be transformed as a ratio of the Sharpe measure of the underlying index (denoted $\hat{\theta}_p$ and that of the maximum obtainable Sharpe measure of the ex post efficient frontier (denoted $\hat{\theta}^*$, spanned by all assets in the market (including the index). It can be shown that $\hat{\theta}^{*2} = \hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha} - \hat{\theta}_p^2$. Hence, it is clear that different values of α give a different level of the ex post "price of risk." A larger $\hat{\theta}^*$ implies that the underlying index is relatively inefficient. Our alternative hypothesis, therefore, can be used to evaluate the power of tests given different degrees of inefficiency of the underlying index. We consider three values for δ: 0.01, 0.02, and 0.03. Presumably, the larger the value of δ, the less efficient of the underlying index, thereby the higher probability one will reject the null.

The results of power tests are reported in Tables 4 and 5. The results show that the GRS and the bootstrap tests have similar power in rejecting the null hypothesis under different distributional specifications. However, the GRS test has a slightly higher power in rejecting the null hypothesis than the bootstrap test almost in all cases. Also, the results show that when $\delta = 0.01$, which may be explained as a measurement error of 1\$ in riskfree rate, the power of both tests are generally smaller than 20%, implying that, as in MacKinlay (1987), both tests may not be sensitive to the use of instruments for riskfree rates.

Our empirical results imply that, if the alternative hypothesis is true, in general the GRS test has a higher probability to reject the null hypothesis than the bootstrap test. However, due to the tremendous computation loads in simulations, in each replication the bootstrap p value is calculated based on a sample of only 250 iterations. A larger number of iterations should be able to improve the performance of the bootstrap test.

6 Conclusion

This paper proposes tests of unconditional mean-variance efficiency using bootstrap method that does not rely on specific distributional assumptions. We reject the mean-variance efficiency of the CRSP value-weighted stock index for five of the seven consecutive ten-year subperiods from 1926 to 1993, whereas the F-test of Gibbons, Ross, and Shanken (GRS, 1989) only rejects two of the seven subperiods. A further examination on the size of the tests shows that the GRS test tends to over-reject the null hypothesis, while the bootstrap test has sizes close to the nominal levels. However, the GRS test has a slightly higher power than the bootstrap test.

Bootstrap methods have many potential applications in finance, especially when the sample size is limited, in which case traditional asymptotic theories may not provide good approximations. The test proposed here can be easily applied to the case of multifactor models. Extension of the bootstrap test to various multivariate linear models, such as multivariate event studies, should also be important and interesting.

We have restricted our test to the case where the error terms are *iid*. Some attempts have been made in the bootstrap literature that extend the model to the case where the sample is independent but not necessarily identically distributed. Extending the bootstrap testing framework that allows for serially correlated and heteroscedastic samples should be an interesting direction of further research.

References

[1] Affleck-Graves, John and Bill McDonald, 1989, Nonnormalities and tests of assets pricing theories, *Journal of Finance* 44, 889-908.

[2] Anderson, T.W., 1984, *An introduction to multivariate statistical Analysis.* 2 ed. New York: Wiley.

[3] Babu, Gutti Jogesh, and C. R. Rao, 1993, Bootstrap methodology, Handbook of Statistics 9, edited by C. R. Rao.

[4] Chatterjee, S. and R. A. Pari, 1990, Bootstrapping the numbers of factors in the arbitrage pricing theory, *Journal of Financial Research* 13, 13-51.

[5] Efron, B., 1979, Bootstrap methods: Another look at the jacknife, *Annals of Statistics* 7, 1-26.

[6] Efron, B. and R. J. Tibshirani, 1993, *An introduction to the bootstrap* New York: Chapman & Hall.

[7] Gibbons, Michael, Stephen A. Ross, and Jay Shanken, 1989, Testing the efficiency of a given portfolio. *Econometrica* 57, 1121-1152.

[8] Goetzmann, W. N. and P. Jorion, Testing the predictive power of dividend yields, *Journal of Fiance* 48, 663-679.

[9] Hall, Peter, 1992, *The bootstrap and edgeworth expansion*, New York: Springer-Verlag.

[10] Hall, Peter, 1994, Methodology and theory for the bootstrap, Handbook of Econometrics IV, Edited by R.F. Engle and D. L. McFadden, Elsevier Science.

[11] Harvey, C. R., 1995, Predictable risk and returns in emerging markets, *Review of Financial Studies* 8, 773-816.

[12] Harvey, C. R. and G. Zhou, 1990, Bayesian inference in asset pricing tests, Journal of Financial Economics 26, 221-254.

[13] Horowitz, J. L., Bootstrap methods in econometrics: Theory and numerical performance, working paper, Department of Economics, University of Iowa.

[14] Ingersoll, J. E. Jr., 1987, *Theory of financial decision making.* Maryland: Rowman & Littlefield.

[15] Jeong, J and G. S. Maddala, 1993, A perspective on application of bootstrap methods in econometrics, Handbook of Statistics 11, Edited by G. S. Maddala, C. R. Rao, and H.D. Vinod. 573-600.

[16] Jobson, J.D., 1982, A multivariate linear regression test for the Arbitrage Pricing Theory *Journal of Finance* 37, 1037-1042.

[17] Kramer, Lisa, 1996, The bootstrap in event studies, working paper, University of British Columbia.

[18] Lamoureux, C. G. and W. D. Lastrapes, 1990, Persistence in variance, structural change and the GARCH model, *Journal of Business Economics and Statistics* 8, 225-234.

[19] MacKinlay, A. Craig, 1987, On multivariate tests of the CAPM, *Journal of Financial Economics* 18, 341-371.

[20] MacKinlay, A. Craig and Matthew P. Richardson, 1991, Using generalized method of moments to test mean-variance efficiency, *Journal of Finance* 46, 511-527.

[21] Marais, M. L., 1984, An application of the bootstrap method to the analysis of squared, standardized market model prediction errors, *Journal of Accounting Research* 22, 34-54.

[22] Muirhead, R.J., 1982, *Aspects of multivariate statistical Theory.* New York: Wiley.

[23] Richardson, Matthew, and Tom Smith, 1993, A test for multivariate normality in stock returns. *Journal of Business* 66, 295-321.

[24] Roll, Richard. 1977 A critique of the asset pricing theory's tests: On past and potential testability of the theory, *Journal of Financial Economics* 4, 129-176.

[25] Shao, Jun and Dongsheng Tu, 1995, *The jacknife and bootstrap*, New York: Springer Verlag.

[26] Vinod, H. D., 1993, Bootstrap methods: Applications in econometrics, Handbook of Statistics 11, Edited by G. S. Maddala, C. R. Rao, and H.D. Vinod. 629-661.

[27] Zhou, Guofu, 1993, Asset pricing tests under alternative distributions. *Journal of Finance* 48, 1925-1942.

Table 1: Test of Efficiency of the CRSP Value-weighted Index

The efficiency is examined by using the market model:

$$R_t = \alpha + \beta r_{pt} + \varepsilon_t, \qquad \varepsilon_t \sim P(0, \Sigma),$$

where R_t is the $(N \times 1)$ vector of excess returns; $\alpha = (\alpha_1, ..., \alpha_N)'$; $\beta = (\beta_1, ..., \beta_N)'$; and $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$. The error term ε_t is assumed to be *iid* with an unknown distribution function $P(0, \Sigma)$, with mean zero and covariance matrix Σ. We test the following hypothesis:

$$H_0 : \alpha = 0.$$

We report the p-values based on the test of Gibbons, Ross, and Shanken (1989) (denoted GRS), asymptotic Wald χ^2 test (denoted Wald), and bootstrap test (denoted Bootstrap). The p value of the bootstrap test is calculated based on a simulation of 100,000 replications.

Period	N	GRS	Wald	Bootstrap
1926/4 - 1935/12	117	0.07837	0.04001	0.01486
1936/1 - 1945/12	120	0.88848	0.86404	0.82851
1946/1 - 1955/12	120	0.07131	0.03661	0.01226
1956/1 - 1965/12	120	0.39878	0.32102	0.22121
1966/1 - 1975/12	120	0.01441	0.00437	0.00122
1976/1 - 1985/12	120	0.02697	0.01010	0.00243
1986/1 - 1993/12	96	0.05626	0.02109	0.00570
1926/1 - 1993/12	813	0.00760	0.00635	0.00513

Table 2: Size of GRS and Bootstrap Tests under Various Distributions

This table presents the empirical results on the power of the GRS test and the bootstrap test for the zero-intercept hypothesis of the market model based on 1,000 replications. In each replication, the bootstrap rejection rate is calculated based on a sample of 1,000 iterations. The error term is generated based on multivariate normal, mixture normal, and multivariate t distributions with 2, 10, and 20 degrees of freedom. The average rejection rates of 1,000 replications under 1%, 5%, and 10% significance levels are reported for both the GRS and the bootstrap tests under different distributional specifications.

	GRS			Bootstrap		
Distribution	1%	5%	10%	1%	5%	10%
Normal	0.011	0.050	0.097	0.008	0.047	0.085
t(20)	0.015	0.062	0.103	0.012	0.054	0.102
t(10)	0.009	0.052	0.106	0.008	0.047	0.096
t(2)	0.014	0.041	0.094	0.010	0.035	0.086
Mixture Normal	0.007	0.038	0.095	0.006	0.037	0.086

Table 3: Size of GRS and Bootstrap Tests under Bootstrap Errors

This table presents the empirical results on the power of the GRS test and the bootstrap test for the zero-intercept hypothesis of the market model based on 1,000 replications. In each replication, the bootstrap rejection rate is calculated based on a sample of 1,000 iterations. The error term is generated based on the OLS residuals of the true sample. The average rejection rates of 1,000 replications under 1%, 5%, and 10% significance levels are reported for both the GRS and the bootstrap tests for each subperiod.

	GRS			Bootstrap		
Sample period[a]	1%	5%	10%	1%	5%	10%
1926/4 - 1935/12	0.023	0.087	0.155	0.010	0.045	0.101
1936/1 - 1945/12	0.027	0.084	0.142	0.010	0.048	0.106
1946/1 - 1955/12	0.014	0.060	0.114	0.010	0.052	0.097
1956/1 - 1965/12	0.018	0.057	0.103	0.010	0.042	0.089
1966/1 - 1975/12	0.021	0.071	0.138	0.005	0.046	0.091
1976/1 - 1985/12	0.023	0.070	0.122	0.013	0.052	0.090
1986/1 - 1993/12	0.014	0.080	0.130	0.006	0.045	0.097
Average	0.0200	0.0727	0.1291	0.0091	0.0471	0.0959
(Std Dev)	(0.0049)	(0.0570)	(0.0177)	(0.0027)	(0.0038)	(0.0063)

[a] The sample period refers to the time period for which the sample estimates of β are used as the parameter values and the OLS market model residuals are used to construct the empirical distribution of the error terms.

Table 4: Power of GRS and Bootstrap Tests under Various Distributions

This table presents the empirical results on the power of the GRS test and the bootstrap test for the zero-intercept hypothesis of the market model based on 1,000 replications. In each replication, the bootstrap rejection rate is calculated based on a sample of 250 iterations. The error term is generated based on multivariate normal, mixture normal, and multivariate t distributions with 2, 10, and 20 degrees of freedom. We consider the alternative hypothesis of the form: $H_A : \alpha = (1 - \beta)\delta$, and three different values for δ are examined: 0.01, 0.02, and 0.03. 5% significance level is used.

	δ =0.01		δ =0.02		δ =0.03	
Distribution	GRS	Bootstrap	GRS	Bootstrap	GRS	Bootstrap
Normal	0.125	0.117	0.501	0.504	0.898	0.884
t(20)	0.114	0.114	0.484	0.465	0.874	0.857
t(10)	0.141	0.124	0.460	0.438	0.842	0.835
t(2)	0.142	0.142	0.448	0.433	0.664	0.649
Mixture normal	0.140	0.131	0.510	0.486	0.843	0.885

Table 5: Power of GRS and Bootstrap Tests under Bootstrap Distributions

This table presents the empirical results on the power of the GRS test and the bootstrap test for the zero-intercept hypothesis of the market model based on 1,000 replications. In each replication, the bootstrap rejection rate is calculated based on a sample of 250 iterations. The error term is generated based on OLS market model residuals. We consider the alternative hypothesis of the form: $H_A : \alpha = (1 - \beta)\delta$, and three different values for δ are examined: 0.01, 0.02, and 0.03. 5% significance level is used.

	δ =0.01		δ =0.02		δ =0.03	
Sample Period[a]	GRS	Bootstrap	GRS	Bootstrap	GRS	Bootstrap
1926/4 - 1935/12	0.202	0.139	0.401	0.309	0.676	0.567
1936/1 - 1945/12	0.425	0.331	0.926	0.861	0.999	0.996
1946/1 - 1955/12	0.210	0.188	0.821	0.799	0.990	0.995
1956/1 - 1965/12	0.171	0.154	0.642	0.601	0.960	0.945
1966/1 - 1975/12	0.155	0.109	0.464	0.382	0.845	0.768
1976/1 - 1985/12	0.177	0.132	0.549	0.457	0.909	0.854
1986/1 - 1993/12	0.133	0.098	0.346	0.248	0.658	0.547

[a] The sample period refers to the time period for which the sample estimates of α and β are used as the parameter values and the OLS market model residuals are used to construct the empirical distribution of the error terms.