ANNALS OF ECONOMICS AND FINANCE **2**, 217–249 (2006)

Using Bootstrap to Test Portfolio Efficiency *

Pin-Huang Chou

National Central University, Taiwan
E-mail: choup@cc.ncu.edu.tw

and

Guofu Zhou

Washington University in St. Louis, USA
E-mail: zhou@wustl.edu

To facilitate wide use of the bootstrap method in finance, this paper shows by intuitive arguments and by simulations how it can improve upon existing tests to allow less restrictive distributional assumptions on the data and to yield more reliable (higher-order accurate) asymptotic inference. In particular, we apply the method to examine the efficiency of CRSP value-weighted stock index, and to test the well-known Fama and French (1993) three-factor model. We find that existing tests tend to over-reject.

Key Words: Bootstrap; Efficiency; GMM test; Elliptical distribution.
JEL Classification Numbers: C13, C53, G14.

1. INTRODUCTION

A fundamental problem in finance is to examine the tradeoff between risk and return. Sharpe (1964) and Lintner's (1965) capital asset pricing model (CAPM) is perhaps one of the most important models in financial economics, which states that the expected return on a security is a linear function of its beta associated with the market portfolio. Question of its empirical validity has generated an enormous amount of research. Kandel

* We are grateful to Raymond Kan, Kin Lam, Carmela Quintos, Jun Shao, seminar participants at University of Wisconsin-Madison, participants of the Fourth Asian-Pacific Finance Conference in Kuala Lumpur and of the 1998 NTU (National Taiwan University) International Conference on Finance and an anonymous referee for many helpful comments. We are especially indebted to Joel Horowitz whose insightful comments lead to substantial improvements of the paper.

1529-7373/2006

and Stambaugh (1989) and Shanken (1996) provide an excellent survey of the earlier literature, while Shanken and Zhou (2006) analyze some of the recent issues.

As the market portfolio is unobservable (Roll (1977)), tests of the CAPM have been focused on testing the mean-variance efficiency of a given portfolio. With normality assumption on stock returns, Gibbons, Ross, and Shanken (1989) provide an exact test. Shanken (1987), Harvey and Zhou (1991), and Kandel, McCulloch and Stambaugh (1995) assess the efficiency in a Bayesian framework. As the normality assumption is usually rejected by the data, Zhou (1993) provides efficiency tests under elliptical assumption, of which the normality assumption is a special case. However, the inference is legitimate only if the underlying elliptical distribution is correctly specified. Because the distribution of asset returns is never known *a priori* in the real world, these studies may subject to a specification error as the conclusions may not hold if the specific distributional assumption is violated.

In this paper, we show that the bootstrap method can be used to obtain more robust and reliable asset pricing tests. Originally proposed by Efron (1979), it is a computation-intensive method for estimating the distribution of a test statistic or a parameter estimator by resampling the data. It usually generates an estimate of the distribution that is at least as accurate as that from standard asymptotic theory (first-order). Hence, it is particularly useful in cases where the asymptotic distribution is difficult to obtain or simply unknown. Moreover, the bootstrap method can often yield, in many applications, higher-order accurate estimates of the distribution that improves upon the usual asymptotic approximation. Because of these advantages, it is not surprising to find applications of the bootstrap method in finance. For examples, Ferson and Foerster (1994) and Kothari and Shanken (1997) apply it to asset pricing and Lyon, Barber and Tsai (1999), among others, use it in corporate finance. Despite its various applications, however, there appears a lack of appreciation as to why it should work and how it improves upon the usual asymptotic theory. Indeed, existing studies in finance consider only the independently and identically distributed case and some of them are problematic in designing the bootstrap.[1] Based on Shao and Tu (1995), and especially on Hall (1992) and Hall and Horowitz (1996), we provide simple intuitions underlying the bootstrap method that facilitates its design in applications.

In particular, we show how to obtain higher-order accurate confidence intervals for security betas and for some other functions of interest. We provide bootstrap tests for the mean-variance efficiency of a given port-

[1] Jeong and Maddala (1993), Vinod (1993), and especially Horowitz (1997), provide excellent surveys of the use and mis-use of the bootstrap in econometrics.

folio. The bootstrap tests allow two broad categories of distributions of the data. The first is that the stock returns are independently and identically distributed (iid) over time. In this case, bootstrap tests are easily constructed to have higher-order accuracy and to be more reliable than the usual asymptotic χ^2 test. The second category of distributions relaxes the iid assumption by allowing for serial dependence. In this case, existing studies, such as MacKinlay and Richardson (1991), usually rely on Hansen's (1982) GMM (generalized method of moments) method. Based on Hall and Horowitz (1996), we show how to obtain the bootstrap analogues of GMM tests. Our procedure seems the first in the finance literature that bootstraps a GMM test correctly in the presence of serial correlation of the data.[2] As the bootstrapped GMM test is obtained by using Hall and Horowitz's adjustments, it is of higher-order accuracy, and hence enjoys in general much better finite sample properties than the usual GMM χ^2 test. Therefore, it appears that it pays off to use the bootstrapped GMM test whenever possible. Because the GMM test is extensively used in finance, there seem wide applications of the suggested bootstrap procedures.

The rest of the paper is organized as follows. Section 2 outlines the standard framework for testing the mean-variance efficiency of a given portfolio. Section 3 introduces the bootstrap method, and shows how it can be used to provide better approximations for the distribution of the beta estimates and other functions of interest, as well as for the distribution of efficiency tests. Section 4 discusses how the bootstrap method can be used to assess the economic importance of portfolio inefficiency. Section 5 provides the empirical applications. To further investigate their performance under various alternative distributional assumptions, Section 6 performs Monte Carlo experiments to document that the bootstrap tests have reliable test sizes and they indeed offer important improvements over existing asymptotic approximations. The last section concludes the paper.

2. EFFICIENCY RESTRICTIONS AND THE GRS TEST

A test of the Sharpe-Lintner's CAPM is a test of the mean-variance efficiency of the market portfolio. Given a proxy or benchmark portfolio, this becomes a test of the efficiency of a given portfolio. In order to conduct such an efficiency test, a statistical model of asset returns has to be specified. Consider the standard market model regression:

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

[2]Although Ferson and Foerster (1994) show the advantage of using the bootstrap method in the GMM framework, their sampling procedure assumes iid residuals.

where

$$\begin{aligned} r_{it} &= \text{the excess return on asset } i \text{ in period } t, \\ r_{pt} &= \text{the excess return on a given portfolio } p \text{ in period } t, \\ &\quad \text{whose efficiency is tested,} \\ \varepsilon_{it} &= \text{the disturbance term for asset } i \text{ in period } t, \end{aligned}$$

N is the number of assets, and T is the number of time series observations. The error terms are assumed to be iid 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 iid assumption makes it easier for us to discuss Gibbons, Ross, and Shanken (1989) efficiency test (the GRS test) and review the literature in this section, but it will be relaxed to allow for serial dependence when we bootstrap the GMM test in subsection 3.4.

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

$$R_t = \alpha + \beta r_{pt} + \varepsilon_t, \qquad \varepsilon_t \sim P(0, \Sigma),\ t = 1, \ldots, T, \tag{2}$$

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

Mean-variance efficiency of the portfolio p implies $E(R_t) = \beta E(r_{pt})$. This pricing restriction translates into the standard testable joint hypothesis on the parameters of the market model regression:

$$H_0 : \qquad \alpha = 0. \tag{3}$$

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

Despite the iid assumption, it is well known that statistically efficient parameter estimates of the multivariate regression model (2) are the same as those under the assumption that the model residuals are normally distributed. Hence, the alphas and betas can be estimated by standard ordinary least squares (OLS) regression. In particular, the efficient estimator of asset i's alpha, $\hat{\alpha}_i$, is given by OLS regression in the i-th equation, and the efficient estimator of Σ is the average of the cross-product of the estimated model residuals. Asymptotically, the null hypothesis can be tested by using a Wald statistic that has an asymptotic chi-square distribution

with N degrees of freedom, i.e.,

$$W \equiv h^{-1}\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha} \overset{asy}{\sim} \chi^2_N, \tag{4}$$

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 the sample mean and sample standard deviation of r_p, respectively.[3] The asymptotic Wald test is valid under the iid assumption in large samples. However, under the normality assumption, Gibbons, Ross, and Shanken (1989) show that the Wald test tends to over-reject the null hypothesis in finite samples.

In fact, under the normality assumption, i.e., P is multivariate normal, Gibbons, Ross, and Shanken (1989) obtain an exact test, the well-known GRS test, that

$$GRS \equiv \frac{T-N-1}{N(T-2)} h^{-1}\hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha} \sim F_{N,T-N-1}(\lambda), \tag{5}$$

where $\lambda = h^{-1}\alpha'\Sigma^{-1}\alpha$ is the noncentrality parameter of the F distribution. The GRS test has a noncentral F distribution with degrees of freedom N and $T - N - 1$. Under the null hypothesis that $\alpha = 0$, the distribution of the GRS test reduces to the central F distribution. Hence, p-values of the statistic are easily calculated. In addition to its easy computation, Gibbons, Ross, and Shanken (1989) also provide an interesting economic interpretation of the GRS test that the test statistic measures how far the Sharpe measure of the given portfolio deviates from the Sharpe measure of the *ex post* efficient portfolio. As a result, it is not surprising that we reject the efficiency if the observed statistic is large. In comparison, the GRS test and the Wald test are clearly statistically equivalent in the sense that the exact p-values of the tests are identical. Furthermore, both are equivalent to the likelihood ratio test.

Because the GRS test has both an interesting economic interpretation and an easily computed exact distribution, it offers a nice solution to testing the efficiency of a given portfolio. However, it is valid theoretically only under the normality assumption. But, as shown by Richardson and Smith (1994) and Zhou (1993), among others, the normality assumption is strongly rejected by the data. On the other hand, the CAPM or mean-variance efficiency is valid under the elliptical distribution, of which the multivariate normal, multivariate t, and the mixture normal distributions are special cases. It is therefore of interest to consider how to test the efficiency without the normality assumption.

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

Although one could use the GRS test without the normality assumption, 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. Thus, the asymptotic Wald test seems the only available alternative test under the iid assumption. However, since it is known that the test is unreliable under the normality assumption, its use in the iid case seems limited. To overcome this problem, we advocate the use of the bootstrap method. Like the asymptotic Wald test, the bootstrap test is also valid under the iid assumption, but it has higher-order accuracy than the asymptotic Wald test. As shown by simulations in Section 6, the bootstrap test is reliable in commonly used small sample sizes not only under the normality assumption, but also under plausible alternative assumptions. Indeed, it performs as well as the GRS test under the normality assumption, and performs much better than the asymptotic Wald test in general.

3. BOOTSTRAP TEST

In this section, we explain first the intuition why the bootstrap should work to deliver at least as accurate approximations as that of the asymptotic theory. Then, in subsection 3.2, we discuss why it improves upon the approximations of the usual asymptotic theory for a large class of estimators and statistics. Based on these intuitions, we develop, in subsection 3.3, a straightforward bootstrap procedure for testing the efficiency of a given portfolio under the iid assumption. Finally, in subsection 3.4, we show how to bootstrap the GMM test when the iid assumption is relaxed.

3.1. Why does bootstrap work?

The bootstrap method, introduced by Efron (1979), is a computation-intensive method for estimating the distribution of an estimator or a test statistic by resampling the data at hand. It treats the data as if they were the population. Under mild regularity conditions, the bootstrap method 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, e.g., Horowitz (1997)). In many instances, the sampling distribution of a statistic or an estimator can be very difficult, if not impossible, to derive even asymptotically, while the bootstrap method, on the other hand, may be easily applied to obtain the sampling distribution of the statistic via repeatedly resampling from the data. Hence, the bootstrap method can serve as an important and interesting alternative to the standard asymptotic theory.

Based on Shao and Tu (1995), we in what follows introduce the idea of bootstrap and explain the intuition why it works. This may be helpful in

applying the method widely and correctly to various models. However, as pointed out earlier, one of the major uses of the bootstrap method is to use it to improve upon the approximations of the usual asymptotic theory for a large class of estimators and statistics. To make the presentation more accessible, we focus here on the basic question concerning why the method should work at all, while leaving discussions on why it works even better than the asymptotic theory to the next subsection.

Let $x_1, \ldots, x_T$ be an iid random sample from an unknown distribution F with mean μ and variance σ^2. Then, the sample variance is

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2, \tag{6}$$

where $\bar{x}$ is the sample mean. The sample variance $\hat{\sigma}^2$ is an estimator of the unknown variance σ^2. Suppose we are interested in the accuracy of the estimation, and want to compute the standard error of $\hat{\sigma}^2$. To do so, consider first the second moment of $\hat{\sigma}^2$,

$$E(\hat{\sigma}^2)^2 = \int (\hat{\sigma}^2)^2 \, dF(x_1, \ldots, x_T), \tag{7}$$

where the integration is carried out over the joint distribution of $(x_1, \ldots, x_T)$. Notice that $\hat{\sigma}^2 = \hat{\sigma}^2(x_1, \ldots, x_T)$ is a function of $x_1, \ldots, x_T$. Because of the iid assumption, $dF(x_1, \ldots, x_T) = f(x_1) \cdots f(x_T)\, dx_1 \cdots dx_T$, where $f(x)$ is the unknown density function of the data. Hence, the second moment computation involves the integration of T variables. Since $f(x)$ or F is unknown, the above integral cannot be computed exactly.

Intuitively, if F is replaced by an estimator, $\hat{F}$, then $E(\hat{\sigma}^2)^2$ can be approximated by using $\hat{F}$ in the integral. Indeed, a simple non-parametric estimator of F is its empirical distribution:

$$\hat{F}(x \leq y) = \frac{1}{T}\sum_{t=1}^{T} I(x_i \leq y), \tag{8}$$

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

$$\hat{F}(x = x_i) = \frac{1}{T}, \qquad i = 1, \ldots, T. \tag{9}$$

With $\hat{F}$, the second moment of $\hat{\sigma}^2$ can be approximated as

$$E(\hat{\sigma}^2)^2 \approx \int (\hat{\sigma}^2)^2 \, d\hat{F}(x_1, \ldots, x_T). \tag{10}$$

Since $\hat{F}$ is known, the above integral can be computed by standard Monte Carlo integration. Samples $\{z_m = (x^*_{m1}, \ldots, x^*_{mT}) : \ m = 1, \ldots, M\}$ can be drawn from $\hat{F}$, and the integral is numerically close to

$$E(\hat{\sigma}^2)^2 \approx \frac{1}{M} \sum_{m=1}^{M} [\hat{\sigma}^2(z_m)]^2, \tag{11}$$

where M is the number of draws and $\hat{\sigma}^2(z_m)$ is the $\hat{\sigma}^2$ estimator valued at the sample $z_m = (x^*_{m1}, \ldots, x^*_{mT})$. This procedure gives an approximated value of the second moment $E(\hat{\sigma}^2)^2$.

Recall that we want to determine the standard error of $\hat{\sigma}^2$. This is the squared root of $E(\hat{\sigma}^2)^2$ minus the square of $E(\hat{\sigma}^2)$. So, what we need to do now is to compute the unknown mean $E(\hat{\sigma}^2) = \sigma^2$. At the above sample draws, σ^2 can clearly be computed as $\frac{1}{M}\sum_{m=1}^{M}[\frac{1}{T}\sum_{t=1}^{T}(x^*_{mt} - \bar{x}^*_m)^2]$, where $\bar{x}^*_m$ is the sample mean of the m-th draw. Hence, the standard error of $\hat{\sigma}^2$ is determined.

The above procedure is actually the standard bootstrap method for obtaining the standard error of $\hat{\sigma}^2$. Clearly, this standard error can also be obtained by using the standard asymptotic theory. In fact, there is theoretically no advantage in using the bootstrap method in this case since it does not offer better approximations than the standard asymptotic theory. The example here is merely for illustration of the idea behind the bootstrap. From this presentation, it is obvious that the bootstrap method also works (under certain regularity conditions) for any estimator or statistic other than $\hat{\sigma}^2$. This requires only replacing $\hat{\sigma}^2$ with the appropriate function in the integrand. In particular, the bootstrap method offers tractable solutions to the standard errors of Kandel, McCulloch and Stambaugh's (1995) portfolio efficiency measures and to our utility measures proposed later, for which it is not obvious at all how one can obtain the associated asymptotic approximations.

Our discussions show that the bootstrap method is a combination of two methods: an analogy principle and a numerical integration. The analogy principle (see Manski (1988)) replaces the unknown distribution F by $\hat{F}$, and the numerical integration is carried out by the Monte Carlo integration. Viewing the bootstrap method this way, countless versions can be proposed. First, various parametric or non-parametric estimators of F may be used. Second, different accelerated Monte Carlo integration methods, such as the control variate and antithetic techniques, may be used in the numerical integration. Because it is the combination of the analogy principle and the numerical integration, the bootstrap method has two sources of error. The first is from replacing F by $\hat{F}$. This error depends on the sample size and the problem at hand, i.e., the true but unknown distribution F. However, as the sample size gets large, $\hat{F}$ approaches F, and the error

approaches zero. The second error is caused from the numerical integration, but this error is controllable in the sense that one can choose M large enough to make potentially the error as small as one would like. Hence, the bootstrap error is primarily caused by the first error. In general, the accuracy of the bootstrap method is at least as good as the usual asymptotic theory, and so it can be very useful in cases where the asymptotic theory is difficult to obtain. In addition, as pointed out by Horowitz (1997) and others, the bootstrap method is still useful when the asymptotic theory is available because it serves as an indicator for the accuracy of the latter. When the two yield substantially different results, the asymptotic theory is usually unreliable. However, a major use of the bootstrap method is that it offers higher-order accurate results in many applications, as shown in the following subsection.

3.2. Why is bootstrap better?

Hall (1992) and Horowitz (1997) provide excellent discussions and analysis on situations where the bootstrap method actually offers higher-order approximations than the usual asymptotic theory. The improvements generally occur for "asymptotically pivotal" statistics whose asymptotic distributions are independent of any unknown population parameters. For example, the Wald statistic W is clearly asymptotically pivotal because its limiting distribution is χ^2 that is independent of any parameters. The reason for the improvements is that such a pivotal statistic, say J, usually admits an Edgeworth expansion of the form

$$P(J < x) = \Phi(x) + T^{-1/2} q(x)\phi(x) + O(T^{-1}), \tag{12}$$

where $\Phi(x)$ is the asymptotic distribution of J, $q(x)$ an even quadratic polynomial, $\phi(x)$ some density function, and T the sample size. Furthermore, under almost the same regularity conditions for (12), the bootstrap statistic, J^*, also admits a similar Edgeworth expansion

$$P(J^* < x) = \Phi(x) + T^{-1/2} \hat{q}(x)\phi(x) + O(T^{-1}), \tag{13}$$

where $\hat{q}(x)$ is an bootstrap estimate of $q(x)$. As it is often the case that $[\hat{q}(x) - q(x)] = O(T^{-1/2})$, it follows from (12) and (13) that the difference between the bootstrap probability, $P(J^* < x)$, and the true probability, $P(J < x)$, is of order $1/T$. In contrast, the asymptotic distribution $\Phi(x)$ has accuracy of only order $1/\sqrt{T}$. Furthermore, one can write out the $O(1/T)$ terms explicitly, and show that the bootstrap error can be reduced to $o(1/T)$, and, for symmetrical tests and iid data, even to a high order of $O(1/T^2)$.

In finance, there are several important cases where the bootstrap method provides higher-order accurate results than what the standard asymptotic

theory can provide. First, the method provides better confidence intervals for security betas and the associated t test. However, it does not provide asymptotic improvements for the estimation of the expected returns and for the estimation of the variance of the market model residuals, unless certain adjustments of the standard bootstrap method are made (see Hall (1992) for details). The second case, of our interest here, is that the bootstrapped Wald efficiency test, as provided in the next subsection, will enjoy better accuracy than the asymptotic theory. As most asset pricing tests are asymptotically pivotal and have simple chi-squared asymptotic distributions, it is expected that the bootstrap method can provide more reliable tests that are useful in a variety of contexts.

However, it should be pointed out that, although the accuracy of the bootstrap method is theoretically at least as high as that of the asymptotic theory in most cases, and it is of higher-order in certain cases, the bootstrap distributions are still approximations and the performance can be poor for some particular applications. As pointed out by Horowitz (1997) and others, for estimators whose asymptotic covariances are nearly singular, the bootstrap is likely to fail. Hence, despite its simplicity and good theoretical properties, the bootstrap method should not be used carelessly or uncritically.

3.3. Bootstrap test: iid case

With the intuitive reason why the bootstrap should work, it becomes almost straightforward to compute the bootstrap distributions for the parameter estimates in the multivariate regression model (2), and to develop a bootstrap efficiency test. Under the iid assumption, the sampling distributions of the bootstrapped tests are clearly obtained by resampling the data and re-estimating the model by the OLS regression. However, some care must be exerted in resampling the data. There are three possible assumptions one may make on the distribution of the data. The first assumption is that the model residuals are iid and the benchmark portfolio returns (the regressors) are treated as fixed constants. In this case, the fitted residuals should be resampled. The second assumption is that both the asset returns and the benchmark portfolio returns are jointly iid. Then, the returns should be jointly resampled. The third assumption is that the benchmark portfolio returns and the model residuals are jointly iid. In this case, one must resample the benchmark portfolio returns and the model residuals jointly. As explained in Hall (1992), the bootstrap approximations provide refinements over asymptotic distributions for the security betas under both the first and third assumptions, but not the second. However, the bootstrap test will always provide refinements over their asymptotic analogues regardless of the assumptions.

To develop the bootstrap analogue of the Wald test, we need to compute the statistic by resampling the data. Following the usual assumption that the errors are iid, we need resample the errors rather the returns. However, the errors are not directly observable, thus fitted residuals must be used. Either unrestricted or restricted residuals can be used because both of which provide approximations to the true but unknown residual distribution. However, in contrast to bootstrapping the distribution of the parameter estimates, we need to generate data from the model under the null hypothesis, which is *not* the same as the original data in most applications. This point, as emphasized by Shao and Tu (1995), seems obvious from the analysis of the simple bootstrap example in subsection 3.1, but is ignored in some existing studies. Formally, we can design the bootstrap efficiency test as follows:

1. Estimate the multivariate regression model (2) by using the OLS to obtain $\hat{\alpha}$ and $\hat{\varepsilon}_t$. Let $\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$. Calculate the Wald statistic:

$$W = h^{-1} \hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha}. \tag{14}$$

2. Estimate the model under the null hypothesis to get β_{res}, $\beta_{res}' = (r_p' r_p)^{-1} r_p' R$, where $r_p' = (r_{p1}, \ldots, r_{pT})$ and $R' = (R_1, \ldots, R_T)$.

3. Repeat the following steps a large number of times (we use 10,000).

(a) Draw ε_t^* $(t = 1, \ldots, T)$ from $\{\hat{\varepsilon}_t\}_{t=1}^{T}$ (with replacement). Then generate data from $R_t^* = \beta_{res} r_{pt}^* + \varepsilon_t^*$, $t = 1, \ldots, T$.

(b) Estimate the model by using the OLS to obtain the estimates $\hat{\alpha}^*$ and $\hat{\Sigma}^*$ with the bootstrapped data $R^{*\prime} = (R_1^*, \ldots, R_T^*)$ and r_p's. Calculate

$$W^* = h^{-1} \hat{\alpha}^{*\prime} \hat{\Sigma}^{*-1} \hat{\alpha}^*. \tag{15}$$

4. Calculate the percentage of W^*'s that are greater than W.

The percentage is the p-value of the bootstrap test. Theoretical justifications of this procedure can be found in Shao and Tu (1995, Chapter 7), and especially in Hall (1992), Horowitz (1997) and references therein. An alternative procedure is to use the residuals under the null for the bootstrap, rather than the unrestricted ones $\{\hat{\varepsilon}_t\}_{t=1}^{T}$. Since the restricted residuals in general do not have zero means, they must be "centered", i.e, de-meaned, before used as bootstrap samples. However, both procedures are equivalent asymptotically. To be focused, we use only the first procedure as outlined above in the rest of the paper.

3.4. Bootstrap test: Non-iid case

Theoretically, the CAPM, as a one-period model, still holds even if the stock returns are elliptically distributed at each period, irrespective of heteroscedasticity and serial dependence. However, in the presence of heteroscedasticity and serial dependence, the model is myopic. Nevertheless, as a fundamental model, it is still of interest to test its empirical implications under a very general statistical assumption. Based on the GMM framework of Hansen (1982), MacKinlay and Richardson (1991) provide a robust GMM test for the CAPM. Since the asymptotic GMM test often over-rejects, we refine this test below via the bootstrap method to yield a test of second-order accuracy.

The GMM estimator is obtained by minimizing a weighted quadratic form of the sample moments:

$$\min g_T'(\theta) W_T g_T(\theta), \tag{16}$$

over the parameter space $\theta = (\alpha, \beta)$, where $g_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t(\theta) \otimes Z_t$, $Z_t = (1, r_{pt}')'$, $\varepsilon_t(\theta) = R_t - \alpha - \beta r_{pt}$, and W_T, $2N \times 2N$, is the weighting matrix. As noted by MacKinlay and Richardson (1991), in the present model without imposing the null hypothesis, the GMM estimator is independent of the weighting matrix, and always coincides with the OLS estimator because there are $2N$ moments conditions and $2N$ parameters.

There are several versions of the GMM test. For simplicity, we consider first the Wald-type GMM test used by MacKinlay and Richardson (1991), and then a more general one that is based on the constrained moment conditions. The Wald-type GMM test is defined as

$$J_1 = T\hat{\alpha}' \left[\eta [D_T' S_T^{-1} D_T]^{-1} \eta'\right]^{-1} \hat{\alpha}, \tag{17}$$

where $\eta = I_N \otimes (1\ 0)$ such that $\eta\hat{\theta} = \hat{\alpha}$, $D_T = -\frac{1}{T}\sum_{t=1}^{T}[I_N \otimes Z_t Z_t']$, and $S_T = \frac{1}{T}\sum_{t=1}^{T}[\varepsilon_t \varepsilon_t' \otimes Z_t Z_t']$. This efficiency test has an asymptotic χ^2 distribution with degrees of freedom N.

To bootstrap the distribution of the J_1 statistic, we need to draw samples $\{\varepsilon_t^*\}$ from an approximate distribution of the residuals. In the presence of serial dependence, we cannot resample the residuals the same way as before, because the empirical distribution (see (8)) of the residuals no longer approximates the unknown residual distribution. This is evident because the empirical distribution ignores any time dependency of the data. It assumes time-independent distribution, and weighs them equally in computing the distribution at any time.

Various parametric or non-parametric density estimators of the residuals can be developed. But it is not a trivial matter to draw samples from them. Fortunately, there is a simple and elegant solution. As pointed

out by Shao and Tu (1995, Chapter 9), we can draw samples $\{\varepsilon_t^*\}$ from $\{\hat{\varepsilon}_t\}$ in *blocks*. To be more specific, we can divide the residual time series into, say l, blocks such that each block has m elements with $lm = T$. Both m and l may depend on T. For example, m may be equal to 2 or 5 percent of T. In this way, block 1 consists of $\epsilon_1, \ldots, \epsilon_m$, block 2 of $\epsilon_{m+1}, \ldots, \epsilon_{2m}$, etc. The bootstrap is implemented by drawing l blocks randomly with replacement from the empirical residuals. Like the iid case, the artificial returns can be generated from $R_t^* = \beta_{res} r_{pt}^* + \varepsilon_t^*$, $t = 1, \ldots, T$, upon which the bootstrapped J_1 statistic is easily calculated. Then, the empirical distribution is straightforward to obtain.

Intuitively, the block length captures dependence of the data up to the m-th lag. If some fixed m is already appropriate in capturing the data dependence, the blocks should be approximately independent, and each block draw approximates a draw from the joint distribution of the entire residuals. As l goes large, the blocks should converge to the joint distribution of the elements in each block. Then, the bootstrap draws are close to draws from the distribution of the entire data. For a complex dependence structure, a fixed m may not be appropriate. But as both m and l approach to infinity, the bootstrap draws should eventually capture the true distribution of the residuals. Theoretically, m may be chosen as proportional to some fixed power of T such that m/T goes to zero as T increases. The block procedure applies to many dependence structures. In particular, it works for m-dependence residuals, which assumes that the time series of residuals, $\{\ldots, \epsilon_{t-1}, \epsilon_t\}$ and $\{\epsilon_{t+m+1}, \epsilon_{t+m+2}, \ldots, \}$, are independent for all t. The m-dependence process is a simple dependence structure of the residuals, which includes the m-th order moving average time series model as a special case. The block draws are generally robust to the order of the moving averages, and to a stationary autoregressive process as well. Although the latter is a moving average process of an infinite order, a certain finite order moving average process will approximate it with negligible errors.

The above bootstrap procedure is simple and intuitive, and applies to tests of almost any regression models with instrumental variables. However, some important asset pricing models may not be cast into the regression framework, and a genuine GMM set-up may have to be used. More importantly, unless certain adjustments are made, the above procedure will not necessarily provide higher-order accurate approximations than the asymptotic distribution of the test. This is because the block draws may not exactly replicate the data generating process, and this effect must be taken into consideration. Hence, certain adjustments must be made to ensure that the bootstrap method works and does better, regardless of whether in regression models or in the more general GMM framework.

Hall and Horowitz (1996) provide adjustments of the standard bootstrap procedure in the general GMM framework such that the computed bootstrap distribution has, at least in theory, higher-order accuracy than the asymptotic theory. To apply Hall-Horowitz adjustments into testing portfolio efficiency, consider first the GMM estimation problem. Following Hansen (1982), the GMM parameter estimation and test are obtained in two steps. First, a fixed constant weighting matrix Ω, perhaps the identity matrix, is used to obtain the GMM estimator under the null, $\tilde{\theta}$, by solving

$$\min \left[\frac{1}{T}\sum_{t=1}^{T} f(X_t,\theta)\right]' \Omega \left[\frac{1}{T}\sum_{t=1}^{T} f(X_t,\theta)\right], \tag{18}$$

where $\theta = \beta$ is the parameter vector under the null, $f(X_t,\theta) = (R_t - \beta r_{pt}) \otimes Z_t$, $2N \times 1$, is the moment conditions of the model, and $X_t = (R_t', Z_t')'$, $(N+2) \times 1$, is the model variables and instruments. In the second step, the optimal GMM estimator, $\hat{\theta}$, is obtained from solving

$$\min \left[\frac{1}{T}\sum_{t=1}^{T} f(X_t,\theta)\right]' W_T \left[\frac{1}{T}\sum_{t=1}^{T} f(X_t,\theta)\right], \tag{19}$$

where the optimal weighting matrix $W_T = S_T^{-1}$, and S_T is a consistent estimator of the covariance matrix of the model residuals given by

$$S_T = \frac{1}{T}\sum_{t=0}^{T}\left[f(X_t,\tilde{\theta})f(X_t,\tilde{\theta})' + \sum_{s=1}^{\kappa} h(X_t, X_{t+s}, \tilde{\theta})\right], \tag{20}$$

and $h(X_t, X_{t+s}, \tilde{\theta}) = f(X_t,\tilde{\theta})f(X_{t+s},\tilde{\theta})' + f(X_{t+s},\tilde{\theta})f(X_t,\tilde{\theta})'$. The kappa, κ, is some integer such that, at the true parameter θ_0, $E[f(X_t,\theta_0)f(X_s,\theta_0)'] = 0$ when $|s-t| > \kappa$. It should be pointed out that this condition is not required and a consistent estimator of the residual covariance may have a more general form than (20) in the GMM setup of Hansen (1982). But it is needed here to ensure theoretically that the bootstrapped GMM test has a better approximation. However, the condition on the covariance of the moments is not as restrictive as it appears. In most asset pricing models, we have moment conditions like $E[f(X_t,\theta_0) \mid I_t] = 0$ conditional on information available at t. Suppose $s < t$ without loss of generality. Then X_s is contained in I_t. It follows by law of expectation that $E[f(X_t,\theta_0) \mid X_s] = 0$, and hence $E[f(X_t,\theta_0)f(X_s,\theta_0)'] = 0$, implying that $\kappa = 0$ in this case.

The GMM test of the efficiency hypothesis is the standard GMM overidentification test. The test statistic is the minimized quadratic form of the second-step estimation multiplied by T. It is distributed asymptotically

χ^2 with the degrees of freedom equal to the number of over-identification conditions. In general, one has to use non-linear numerical optimization techniques to solve the two-step GMM estimation problem in order to compute the GMM test statistic. However, in the present case of testing for efficiency, the estimation problem can be analytically solved. Let Z be a $T \times 2$ matrix of the instruments. Recall that R and r_p are a $T \times N$ matrix of the asset returns and a $T \times 1$ matrix of the benchmark returns, respectively. Then, the $\tilde{\theta}$ is given explicitly by

$$\tilde{\theta} = (x'\Omega x)^{-1} x'\Omega y, \tag{21}$$

where $y = \frac{1}{T}\text{vec}(Z'R)$ and $x = \frac{1}{T} I_N \otimes (Z'r_p)$, both of which are $2N$ vectors. Clearly, with Ω replaced by W_T, the same formula holds for the second-step GMM estimation $\hat{\theta}$. In addition, the GMM statistic can be written as

$$J_2 = Ty'[W_T - W_T x(x'W_T x)^{-1} x'W_T]y, \tag{22}$$

where W_T is the second-step weighting matrix, $W_T = S_T^{-1}$ with S_T given by (20).

Now we want to obtain the bootstrapped analogue of the GMM test J_2. Following Hall and Horowitz (1996), we can still sample the X_t's in blocks. For each of the bootstrap samples, we need to carry out the GMM estimation and compute the test statistic. The GMM estimation process will be similar to the previous real data case, but differs in one important aspect. The moment function has to be *centered* for the bootstrapped data, for it may not have zero expectation at all under the empirical probability. The centered moment function is

$$f^*(X, \theta) = f(X, \theta) - \frac{1}{T} \sum_{t=1}^{T} f(X_t, \hat{\theta}), \tag{23}$$

where $\hat{\theta}$ is the GMM estimator from the real data. With the centered moment function, we obtain, at each sampling draw, the bootstrapped GMM estimator in the same way as the usual GMM estimation. Specifically, let $X_t^* = \{(X_{i+1}^*, \ldots, X_{i+m}^*),\ i = 1, \ldots, l\}$ be a bootstrap sample drawn randomly with replacement from the l blocks of data, the first-step bootstrapped GMM estimator, $\tilde{\theta}^*$, is obtained by solving

$$\min \left[\frac{1}{T} \sum_{t=1}^{T} f^*(X_t^*, \theta) \right]' \Omega \left[\frac{1}{T} \sum_{t=1}^{T} f^*(X_t^*, \theta) \right], \tag{24}$$

where $f^*(X_t^*, \theta)$ is the centered moment function $f^*(X, \theta)$ valued at $X = X_t^*$. Fortunately, in our present application, the solution to (24) is analyt-

ically available,

$$\tilde{\theta}^* = (x^{*\prime}\Omega x^*)^{-1} x^{*\prime}\Omega y^*, \tag{25}$$

where $y^* = \frac{1}{T}\text{vec}(Z^{*\prime}R^*) - \bar{f}(\hat{\theta})$, $x^* = \frac{1}{T} I_N \otimes (Z^{*\prime} r_p^*)$, and $\bar{f}(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} f(X_t, \hat{\theta})$. Clearly this formula is the same as (21) except that the real data is replaced by the bootstrapped one, and an adjustment term of $\bar{f}(\hat{\theta})$ is present for the centering. The analytical solution makes it easy later to determine the empirical size of the bootstrapped GMM test. With $\tilde{\theta}^*$, the second-step weighting matrix, W_T^*, which is an analogue of W_T, can be obtained from the bootstrapped data. Then, with Ω replaced by W_T^*, the above formula yields the second-step bootstrapped GMM estimator, $\hat{\theta}^*$. Hence, the bootstrap analogue of J_2 is obtained, but its distribution is complex, and will not necessarily approximate the exact distribution of J_2 accurately.

To ensure that the computed bootstrap distribution has higher-order accuracy than the asymptotic one, one must apply Hall and Horowitz's (1996) adjustments to the bootstrap covariance estimators and to the eventual test statistic. Intuitively, the adjustments are necessary at least due in part to the centering of the moment conditions. There is also the fact that the block draw may not exactly replicate the dependence structure of the data. For example, adjacent blocks will not generally be independent, but maintain certain dependence structures at the ends. As a result, it is not surprising that the asymptotic covariance matrix of the bootstrapped model residuals has the following form,

$$\tilde{S}_T^* = \frac{1}{T}\sum_{i=0}^{l-1}\sum_{j=1}^{m}\sum_{k=1}^{m} f^*(X_{im+j}^*, \hat{\theta}) f^*(X_{im+k}^*, \hat{\theta})', \tag{26}$$

which is not S_T, the asymptotic covariance matrix of the true model residuals. Hence, the covariance matrix of the asymptotic distribution of $\sqrt{T}\sum_{t=1}^{T} f^*(X_t^*, \hat{\theta}^*)$ becomes (see, e.g., Hansen (1982, Lemma 4.1)),

$$\Psi_T^* = W_T^{*-1/2} N_T^* (W_T^{*1/2} \tilde{S}_T^* W_T^{*1/2}) N_T^* W_T^{*-1/2}, \tag{27}$$

where

$$N_T^* = I_{2N} - W_T^{*1/2} D_T^* (D_T^{*\prime} W_T^* D^*)^{-1} D_T^{*\prime} W_T^{*1/2}, \tag{28}$$

and D_T^* is the first order derivatives of the moment condition, $\frac{1}{T}\sum_{t=1}^{T} \partial f^*(X_t^*, \hat{\theta}^*)/\partial\theta$. Then, $W_T^{*1/2}\Psi_T^* W_T^{*\prime 1/2}$ is not even a diagonal matrix. Therefore, the minimized quadratic form multiplied by T will no longer have an asymptotic χ^2 distribution. Nevertheless, there are two ways to obtain a bootstrapped χ^2 statistic.

The first approach, as suggested by Hall and Horowitz (1996), is to compute the Moore-Penrose generalized inverse, V_T^{-1}, of the matrix

$$V_T^* = N_T^* W_T^{*1/2} \tilde{S}_T^* W_T^{*1/2} N_T^*, \tag{29}$$

where W_T^* is the second-step optimal weighting matrix valued at $\hat{\theta}$, and D_T^*, $2N \times N$, is the first order derivatives of the moment condition, $\frac{1}{T}\sum_{t=1}^{T} \partial f(X_t^*, \hat{\theta})/\partial\theta$. Then, the bootstrapped GMM statistic is obtained as

$$J_2^* = T g(X^*, \hat{\theta}^*)' W_T^{*1/2} V_T^{-1} W_T^{*1/2} g(X^*, \hat{\theta}^*), \tag{30}$$

where $g(X^*, \hat{\theta}^*) = \frac{1}{T}\sum_{t=1}^{T} f^*(X_t^*, \hat{\theta}^*)$ and $\hat{\theta}^*$ is the bootstrapped GMM estimator with the second-step weighting matrix W_T^*. Clearly, the asymptotic covariance matrix of $W_T^{*1/2}\frac{1}{T}\sum_{t=1}^{T} f^*(X_t^*, \hat{\theta}^*)$ is $W_T^{*1/2}\Psi_T^* W_T^{*\prime 1/2} = V_T^*$. As $V_T^{-1/2} V_T^* V_T^{-\prime 1/2}$ is a diagonal matrix whose diagonal elements are zero and ones, it is not surprising that the adjusted statistic, J_2^*, is asymptotically χ^2 distributed. Furthermore, its empirical distribution provides asymptotic refinements over J_2. Indeed, under certain regularity conditions, Hall and Horowitz (1996) show that the critical value of J_2^* approximates the exact one with an error of order $1/T$ and the error is through $O(1/T)$, that is,

$$P(J_2 > z_\tau^*) = \tau + o(1/T), \tag{31}$$

where z_τ^* is the τ-level critical value of J_2^*, and $o(1/T)$ is the error term which after dividing by T goes to zero as the sample size goes to infinity. In contrast, the standard χ^2 approximation of the exact distribution of J_2 has an error of order $1/\sqrt{T}$. In practice, one can resample the data, say 1,000 times, to get 1,000 values of J_2^* as above. Then the largest τ percent of the 1,000 values determine the p-value of the test, which is generally more reliable than the asymptotic chi-squared p-value as it is of higher-order accuracy.

The second approach is to use a simple idea of Zhou (1994). As noted earlier, the covariance matrix of the asymptotic distribution of $\sqrt{T}\sum_{t=1}^{T} f^*(X_t^*, \hat{\theta}^*)$ is Ψ_T^*, which can be simplified as

$$\Psi_T^* = [I_{2N} - D_T^*(D_T^{*\prime} W_T^* D^*)^{-1} D_T^{*\prime} W_T^*] S_T^* [I_{2N} - D_T^*(D_T^{*\prime} W_T^* D^*)^{-1} D_T^{*\prime} W_T^*]. \tag{32}$$

This matrix must have rank $d = N$ under the null, where N is the number of over-identification conditions. Let $u_1 \geq \ldots \geq u_d$ be its nonzero eigenvalues of Ψ_T^*. Then there is a unique P such that:

$$\Psi_T^* = P' \,\mathrm{Diag}(u_1, \ldots, u_d, 0, \ldots 0)\, P = P' U P, \tag{33}$$

where $PP' = P'P = I_{2N}$. In fact, the i-th row of P is the standardized eigenvector corresponding to the i-th largest eigenvalue u_i. Therefore, the covariance matrix of $\sqrt{T}U^{-1/2}\sum_{t=1}^{T} f^*(X_t^*, \hat{\theta}^*)$ has an asymptotic covariance matrix $U^{-1/2}P\Psi_T^*P'U^{-1/2}$, which is simply $\text{Diag}(1, \ldots, 1, 0, \ldots 0)$. As a result,

$$J_2^{**} = Tg(X^*, \hat{\theta}^*)'PU^{-1}Pg(X^*, \hat{\theta}^*) \tag{34}$$

is asymptotically χ^2 distributed with degrees of freedom d. This can serve as an alternative bootstrapped GMM statistic. However, a careful examination of J_2^* and J_2^{**} shows that they are in fact identical numerically. Indeed, both J_2^* and J_2^{**} are quadratic forms of $g(X^*, \hat{\theta}^*)$. The weighting matrix of J_2^{**} is $P'U^{-1}P$, the Moore-Penrose generalized inverse of the matrix Ψ_T^*. The weighting matrix of J_2^* is $W_T^{*1/2}V_T^{-1}W_T^{*1/2}$. By definition, $\Psi_T^* = W_T^{*-1/2}V_T^*W_T^{*-1/2}$. A direct inversion yields $W_T^{*1/2}V_T^{-1}W_T^{*1/2}$. Therefore, J_2^* and J_2^{**} are identical. Henceforth, we use only notation J_2^*, and use only formula (34) for its computation. In practice, the latter is much faster to implement for at least $W_T^{*1/2}$ is not needed.

In comparison of J_2^* with the standard GMM test by using the bootstrap sample, the only difference is that we have now the extra term, V_T^{-1}, occurring in (30) to form the test statistic. This extra term, Hall and Horowitz's (1996) adjustment, is computed from (29) with $\tilde{S}_T^*$ determined from (26). As mentioned earlier, a major reason for the adjustment is that the block draw does not exactly replicate the data generating process, especially in the general serial dependence case. But if it is known that the data is iid, the adjustment of the GMM statistic will be unnecessary. The minimized quadratic from the bootstrap sample will serve directly as the bootstrapped GMM statistic. In addition, the block sampling can be reduced to $m = 1$, implying that one can draw the bootstrap samples as in standard iid applications. As shown by Hall and Horowitz (1996), the payoff of the complex adjustment is that the accuracy of the bootstrapped GMM tests is of order $1/T$ (actually through $O(1/T)$). In contrast, the accuracy of the standard asymptotic chi-squared tests is of order $1/\sqrt{T}$. One may think that the bootstrapped GMM tests are difficult to compute. But as demonstrated above, despite the appearance of complex algebra, the bootstrap statistic is conceptually easy to understand and computationally easy to implement.

4. MEASURES OF INEFFICIENCY

In practice, it is unlikely to find a portfolio that is perfectly efficient. In fact, even if a portfolio is *ex ante* exactly efficient, there will be deviations of its observed Sharpe measures from the *ex post* efficient portfolio simply because of sampling errors. The GRS test compares the observed Sharpe

measure of a given portfolio with that of the *ex post* efficient portfolio. The difference between the Sharpe measures indicates whether or not the *ex post* efficient portfolio is preferred to the given portfolio for all risk averse investors, but the degree of inefficiency, as measured by the difference between the Sharpe measures, is not necessarily the inefficiency measure of a given investor. In other words, while all investors prefer a portfolio that has a higher Sharpe measure, the economic consequence of a 10-percent higher Sharpe measure is unknown.

An investor's preference is fully characterized by his or her utility function. Without specifying the utility function, it is generally impossible to determine the importance of the difference between two portfolios. To assess the economic consequence, a specific form of the individual's utility function must be assumed. For purpose of illustration, we assume a simple quadratic utility,

$$u(W) = W - \frac{c}{2}W^2, \tag{35}$$

where the utility is defined over the end-of-period wealth W, and c is a constant parameter measuring the individual's risk aversion. For simplicity, we can assume that the initial wealth is one by re-scaling the unit of wealth, so that the argument of the utility is the gross return of investment.

Consider the scenario where the individual is interested in how to invest his wealth in light of current efficiency test of a given portfolio. Assume the riskfree rate is r_f. Then, in a standard one-period expected utility maximization framework, the investor's problem is to choose portfolio weights w to maximize his expected utility. If he accepts the hypothesis that the given portfolio r_p is efficient, he will allocate his wealth between r_p and the riskfree asset. Then, his maximized expected utility is

$$u_p = 1 + w_p\mu_p + r_f - \frac{c}{2}\left[(1 + w_p\mu_p + r_f)^2 + w_p^2\sigma_p^2\right], \tag{36}$$

where $w_p = \mu_p[(1 - c(1 + r_f)]/(c(\mu_p^2 + \sigma_p^2))$ is the optimal weight on r_p, μ_p is the expected return of r_p in excess of the riskfree rate, and σ_p^2 is the associated variance. If he does not accept the efficiency hypothesis, he will invest some of his money into the riskfree asset (or borrowing some money at the riskfree rate), and the rest into a portfolio of $R_1, \ldots, R_N$ and r_p. The maximized expected utility is

$$u_R = 1 + w_R'\mu + r_f - \frac{c}{2}\left[(1 + w_R'\mu + r_f)^2 + w_R'\Xi w_R\right], \tag{37}$$

where $w_R = [(1 - c(1 + r_f)](\mu\mu' + \Xi)^{-1}\mu/c$ is the optimal weights on the risky assets, μ is a vector of the expected returns on $R_1, \ldots, R_N, r_p$ in excess of r_f, and Ξ the covariance matrix. The economic importance of

the portfolio efficiency can only be answered if one can compute u_p and u_R. Now the parameters, μ_p, σ_p, μ and Ξ are unknown, and have to be estimated from the data. At a given estimate, u_p and u_R can be computed easily. However, the parameters have estimation errors, and hence the computed utility values are likely to be biased estimates. To overcome uncertainty of the parameter estimates, the correct values of u_p and u_R should be computed as the expectations taken over the distribution of the parameter estimates.[4] But even in the normality case, it is not an easy matter to compute such expectations.

As mentioned earlier, the bootstrap method can be used to compute the distribution of almost any statistic. In particular, it offers an effective and feasible way to compute u_p and u_R. The difference between u_p and u_R measures the economic importance of the decision making with respect to portfolio efficiency. Furthermore, a certainty equivalence amount of money can be easily computed to examine how much one needs to pay for the quadratic utility individual to accept efficiency.

5. EMPIRICAL RESULTS

In this section, we provide two empirical applications of the bootstrap method. First, we investigate the mean-variance efficiency of the CRSP (Center for the Research of Stock Prices) value-weighted index. This is examined by using the ten standard CRSP size decile portfolios on the New York Stock Exchange. The data is monthly from January 1926 to December 1995. All returns are calculated in excess of the average one-month T-bill rate, available from Ibbotson Associates and from the Fama bond file of the CRSP data base.

In the second application, we test the well-known Fama and French (1993) three-factor model which enjoys popularity with practitioners.[5] In this model, there are three common risk factors that are used to explain the average returns on 25 stock portfolios formed on size and book-to-market (B/M). The factors are an overall market factor, $RMRF_t$, as represented by the excess return on the weighted market portfolio; a size factor, SMB_t, as represented by the return on small stocks minus those on large stocks and a book-to-market factor, HML_t, as represented by the return on high B/M stocks minus those on low B/M stocks. The data are monthly from

[4] In a Bayesian framework, such as McCulloch and Rossi (1990), the uncertainty is incorporated into the posterior naturally. However, the disadvantage is that the residual distribution has to be restricted to a few tractable ones.

[5] See, e.g., page 41 of *Cost of Capital*, 1998 yearbook published by Ibbotson Associates.

January 1964 to December 1993, a total of thirty years data, 372 observations.[6]

We test the efficiency under two separate assumptions on the stochastic behavior of the market model residuals over time. First, we assume the residuals are iid. In this case, there are three tests available, the GRS test, the asymptotic Wald χ^2 test and the proposed bootstrap test. Since the monthly returns span seventy years, the market model structure is unlikely to stay constant over the entire period. Therefore, concerning about parameter stability, we, following most studies, apply the tests to ten-year subperiods. This implies $T = 120$ in our application. Although the efficiency is tested directly only in the ten-year subperiods, an indirect test of the model over the entire seventy years can be provided. Based on Shanken (1985), an overall p-value of the efficiency hypothesis can be computed which aggregates the p-values over subperiods. The novelty of this aggregation test is that it allows one to make inferences over the whole sampling period without assuming the stationarity of the parameters during that period. However, it is still necessary to assume stationarity within each of the subperiods as well as the independence of the model residuals across subperiods.

The second assumption on the model residuals is weaker than the first one. It allows for serial dependence and a general non-parametric density for the data. In this case, there are four tests available, the GMM tests, J_1 and J_2, and their bootstrapped analogues, J_1^* and J_2^*. For comparison with the iid case, we also apply these tests to the ten-year subperiods.

The empirical results are reported in Panel A of Table 1. Under the iid assumption, the GRS test rejects the efficiency of the CRSP index in one of the seven subperiods at the 5 percent significance level. In contrast, the asymptotic Wald test, W, has lower p-values than the GRS test for all the subperiods, and it rejects the efficiency in three of the subperiods. This is expected because Gibbons, Ross and Shanken (1989), among others, have pointed out that the asymptotic Wald test tends to reject more often than the GRS test (under normality). If normality holds, the over-rejection will be incorrect. However, due perhaps to non-normality, whether or not the asymptotic Wald test is reliable needs further examination.

As shown in the table, the bootstrapped Wald test, W^*, does not reject the efficiency at all.[7] Whether the data follow a multivariate normal or a non-normal distribution, given the good performance of the bootstrap test under various distributions to be shown in the next section, it appears plausible to rely more on the bootstrap test than on the GRS test in reaching a conclusion about efficiency. As the bootstrap test does not reject

[6]See Fama and French (1993) for a detailed description of the data. The authors are grateful to C. Harvey and R. Kan for permission and forward of their data.

[7]The bootstrap p-values of this section are based on 10,000 bootstrap draws.

TABLE 1.

Efficiency Tests

	iid case			non-iid case			
Period	GRS	Wald	W^*	J_1	J_1^*	J_2	J_2^*
			Panel A: CRSP index				
1926/1-1935/12	0.0406	0.0173	0.0792	0.0084	0.0587	0.0370	0.2229
1936/1-1945/12	0.8875	0.8629	0.8765	0.8255	0.8539	0.8788	0.9217
1946/1-1955/12	0.0715	0.0362	0.0886	0.1046	0.2093	0.3493	0.4817
1956/1-1965/12	0.0786	0.0410	0.0873	0.0271	0.0779	0.1832	0.2594
1966/1-1975/12	0.4903	0.4150	0.5495	0.4085	0.5506	0.4973	0.6215
1976/1-1985/12	0.1951	0.1310	0.2107	0.1029	0.1920	0.1747	0.2304
1986/1-1995/12	0.2843	0.2104	0.3800	0.3917	0.5706	0.8116	0.8824
1926/1-1995/12	0.0330	0.0057	0.0680	0.0077	0.1057	0.2286	0.6075
			Panel B: Fama-French Factors				
1964/1-1993/12	0.0001	< 0.0001	0.0003	< 0.0001	0.0055	0.0015	0.0065

We examine first the efficiency of the CRSP value-weighted index in the standard market model:

$$R_t = \alpha + \beta r_{pt} + \varepsilon_t, \qquad t = 1, \cdots, T,$$

where R_t is a vector of returns on 10 CRSP size decile portfolios in excess of the 30-day T-bill rate. The asset pricing restrictions,

$$H_0 : \alpha = 0,$$

are tested by using Gibbons, Ross, and Shanken's (1989) test (GRS), the asymptotic Wald χ^2 test (Wald), the bootstrapped Wald test (W^*), the GMM tests (J_1 and J_2) and the bootstrapped ones (J_1^* and J_2^*). Panel A of the table provides the p-values. The bootstrap p-values are based on 100,000 resampling of the data. The p-values for the whole period are computed by using Shanken's (1985) normal density aggregation technique. We also examine the joint efficiency of the Fama and French's (1993) factors in

$$R_{it} - r_{ft} = \alpha_i + \beta_i RMRF_t + s_i SMB_t + h_i HML_t + \varepsilon_{it},$$

where R_{it}'s are monthly returns on 25 stock portfolios formed on size and book-to-market (B/M), $RMRF_t$ is the excess return on a market index, SMB_t is the return on small stocks minus those on large stocks, HML_t is the return on high B/M stocks minus those on low B/M stocks, and r_{ft} is the 30-day T-bill rate. The same asset pricing restrictions are tested and the p-values are provided in Panel B of the table.

efficiency in any of the subperiods, and the aggregated p-value is 0.0680, we may conclude that the tests do not reject the mean-variance efficiency of the CRSP value-weighted index in the case where the asset returns are assumed to follow iid distributions.

If we relax the iid assumption and allow the data to have serial dependence, then the mean-variance efficiency can be examined by using the GMM tests, J_1 and J_2, and the bootstrapped GMM tests, J_1^* and J_2^*. Notice that J_1 is just the non-normality and dependence adjusted asymptotic Wald test, but J_2 is the GMM over-identification test. It is interesting that, after adjusting for non-normality and serial dependence, the new Wald test, J_1, rejects the efficiency only in two of the three subperiods rejected earlier based on W. However, the bootstrapped test, J_1^*, cannot reject in any of the subperiods, suggesting that there is some tendency for J_1 to over-reject the null. The over-identification test J_2 and its bootstrap analogue J_2^* seem to echo the conclusions reached by J_1 and J_1^*. For example, J_1 has the lowest p-value, 0.0084, in the first subperiod, and J_2 also has its lowest p-value, 0.0587, in the same subperiod. However, J_2 in general has higher p-values than J_1. Like J_1^*, the bootstrapped test J_2^* has greater p-values than its analogue J_2. Overall, the bootstrap tests do not reject the efficiency. However, their p-values appear to depend heavily on the assumptions we made on the model residuals. Imposing the non-iid assumption generally makes the tests to have higher p-values than before. For example, in the last subperiod, W^* has a p-value of 38.00 percent, whereas J_1^* and J_2^* have p-values 57.06 and 88.24 percent, respectively. This may suggest that one should be cautious in interpreting studies that impose either normality or iid assumption on the data.

Now we apply the same procedure to the Fama and French (1993) model. The results are reported in Panel B of Table 1. It is seen that the p-value from the GRS test is 0.0001, suggesting a rather strong rejection. This is also echoed by W and W^*. Relaxing the iid assumption does not help because the J tests suggest rejection as well. However, it is interesting to notice that the bootstrap tests always have greater p-values than the non-bootstrap ones. For example, J_1 has a p-value less than 0.0001, but J_1^* gets 0.0055, more than 50 times greater. One may ask why the three-factor model is rejected whereas the earlier one-factor model is not. This is because different asset returns are used in the two models. Intuitively, Fama and French's (1993) asset returns are sorted to have greater cross-sectional differences in expected returns, and, as a result, a greater number of factors are needed to model or explain the returns.

The above statistical tests assess how the model fits the data given efficiency. It may not convey the concept of utility loss resulted from investment decisions based on whether or not the benchmark portfolio is efficient. Hence, it is of interest to examine the economic measures of inefficiency

which may shed some light on why the efficiency is not rejected. Following the framework of Section 4, we examine the values of the maximized expected utility under two scenarios. The first is to accept efficiency. The maximized expected utility, u_p, depends on our specification of the risk-aversion parameter c as well as the riskfree rate r_f. For simplicity, we take r_f as the average riskfree rate over the relevant time period, then it is straightforward to apply the bootstrap procedure to compute the maximized utility for any given value of c. As reported in Panel A of Table 2, given the efficiency of the CRSP index, u_p equals 0.8064, 0.7535, 0.7019, 0.6509 and 0.6004 for $c = 0.4$, 0.5, 0.6, 0.7 and 0.8 in the first subperiod. It is seen that u_p decreases as risk aversion increases. This is true not only for the first subperiod, but also for all the subperiods and the entire sample period. The second scenario is to reject efficiency. In this case, the maximized expected utility in the first subperiod, u_R, is 0.9060, 0.8088, 0.7314, 0.6652 and 0.6059 for the previous c values. Like u_p, u_R is also a decreasing function of c. Moreover, at a given value of c, u_R is greater than u_p. This is expected as the opportunity set for maximizing u_R is greater than that for maximizing u_p.

The economic importance of the efficiency of the given portfolio is clearly indicated by the percentage gain in utility, $(u_R - u_p)/u_p$. For example, in the first subperiod, the gain in utility is 10.9279 percent for individuals with risk aversion $c = 0.4$. However, the gain reduces substantially to only 0.2220 percent for those with risk aversion $c = 0.9$. To provide some understanding on the degree of risk aversion, the third column of Table 2 reports the optimal portfolio weight, w_p, on the given benchmark risky portfolio which is obtained by allocating investments between the risky asset and a riskless one. For $c = 0.4$, $w_p = 1.2112$, which suggests an aggressive investment strategy that borrows money to invest in the risky asset, an unlikely event for most of the investors in the real world. On the other hand, a risk aversion of $c = 0.8$ implies an allocation of only 20.69 percent of wealth into the risky asset, another unlikely event for the average investor (many brokerages' recommendations for stock investments are close to 50 percent or exceed it by a small amount). Hence, in the first subperiod, the plausible value for c is 0.6 at which the gain in utility is only 4.0198 percent. This echoes the previous p-value analysis that the GRS test rejects the efficiency only slightly, whereas the bootstrap tests suggest no rejections at all. Also consistent with the p-value analysis, the utility gains are the greatest in the first subperiod where the p-values are the lowest. However, at later subperiods, the plausible c values are higher than 0.6. As c increases, the utility gain becomes smaller. Intuitively, the more risk-averse an investor, the less impact the risky asset allocation. As a result, he cares less about efficiency, implying less utility gain if he rejects efficiency. Finally, it is observed that w_p has negative values in the

TABLE 2.

Utility of Accepting/Rejecting Efficiency

Period	c	w_P	u_P	u_R	$\frac{u_R - u_P}{u_P} \times 100\%$
		Panel A: CRSP value-weighted index			
26/01-35/12	0.4	1.2112	0.8064	0.9060	10.9279
	0.5	0.8036	0.7535	0.8088	6.8024
	0.6	0.5278	0.7019	0.7314	4.0198
	0.7	0.3377	0.6509	0.6652	2.1412
	0.8	0.1966	0.6004	0.6059	0.9159
36/01-45/12	0.4	3.7724	0.8142	0.8641	5.7540
	0.5	2.5124	0.7579	0.7858	3.5373
	0.6	1.6702	0.7042	0.7190	2.0455
	0.7	1.0630	0.6520	0.6591	1.0761
	0.8	0.6153	0.6008	0.6036	0.4610
46/01-55/12	0.4	10.7208	0.8380	0.9213	8.9813
	0.5	7.1143	0.7713	0.8178	5.6607
	0.6	4.7789	0.7117	0.7362	3.3216
	0.7	3.0708	0.6558	0.6676	1.7672
	0.8	1.7787	0.6024	0.6070	0.7525
56/01-65/12	0.4	8.2469	0.8213	0.9091	9.6042
	0.5	5.5001	0.7624	0.8110	5.9731
	0.6	3.6531	0.7070	0.7330	3.5338
	0.7	2.3642	0.6538	0.6662	1.8612
	0.8	1.3496	0.6017	0.6065	0.7842
66/01-75/12	0.4	-0.8892	0.8069	0.8769	7.9104
	0.5	-0.5885	0.7547	0.7933	4.8390
	0.6	-0.3740	0.7031	0.7235	2.8133
	0.7	-0.2418	0.6520	0.6618	1.4695
	0.8	-0.1455	0.6012	0.6049	0.6156

(Table 2 continued)

Period	c	w_P	u_P	u_R	$\frac{u_R-u_P}{u_P} \times 100\%$
76/01-85/12	0.4	5.2558	0.8173	0.8972	8.8495
	0.5	3.4387	0.7605	0.8049	5.5010
	0.6	2.3022	0.7065	0.7299	3.2087
	0.7	1.4756	0.6538	0.6650	1.6883
	0.8	0.8428	0.6020	0.6062	0.7031
86/01-95/12	0.4	6.4329	0.8207	0.8937	8.1065
	0.5	4.2986	0.7623	0.8027	5.0153
	0.6	2.8262	0.7071	0.7286	2.9380
	0.7	1.8221	0.6539	0.6642	1.5471
	0.8	1.0510	0.6019	0.6059	0.6504
26/01-95/12	0.4	2.3488	0.8068	0.8176	1.3176
	0.5	1.5480	0.7545	0.7605	0.7833
	0.6	1.0370	0.7031	0.7063	0.4502
	0.7	0.6602	0.6520	0.6535	0.2304
	0.8	0.3834	0.6012	0.6018	0.0967
			Panel B: Fama-French Factors		
64/01-93/12	0.4	4.32577	0.8083	0.9175	11.8674
	0.5	2.89317	0.7555	0.8157	7.3661
	0.6	1.92737	0.7036	0.7356	4.3357
	0.7	1.22616	0.6523	0.6675	2.2819
	0.8	0.70778	0.6013	0.6072	0.9618

Assume a simple quadratic utility for investors,

$$u(W) = W - \frac{c}{2}W^2,$$

where the utility is defined over the end-of-period wealth W and c is a constant parameter measuring the individual's risk aversion. If an investor accepts the hypothesis that the CRSP value-weighted index is efficient, he obtains his maximized expected utility, u_p, by optimally allocating his wealth (one unit) between the riskfree asset and the index, with w_p being the weight on the index. If he rejects the efficiency, his maximized expected utility, u_R, is obtained by optimally allocating his wealth between the riskfree asset and a mix of the index and size portfolios. The table reports w_p, u_p, u_R and $(u_R - u_p)/u_p$ (in percent) for a wide range of risk aversion parameter values. Panel A of the table reports the results. Panel B of the table provides similar results for the joint efficiency of the Fama and French's (1993) factors.

subperiod of January 1966 to December 1975. This is because the stock market in this decade has basically zig-zag movements. As a result, the excess return is −0.1206 percent per month. Hence, in an *ex post* optimal portfolio choice, investors would short the market portfolio proxy r_p and hence have the negative weight. Clearly, in this decade, r_p is inefficient and is dominated by the riskfree asset. However, this is not reflected by the previous p-value analysis. The reason is that the p-values essentially compare the squared Sharpe measure of the given portfolio, r_p, with that of the *ex post* efficient portfolio, and hence they are invariant to the signs of the excess returns on r_p.[8]

For the Fama and French's (1993) model, the efficiency implies the joint efficiency of the factors, i.e., a portfolio of the factors lies on the efficient frontier. The utility measures can be computed as easily as before when they applied to the joint efficient portfolio. As reported in Panel B of Table 2, the percentage gain in utility is only 0.9618 for $c = 0.8$, and 2.2819 for $c = 0.7$. The magnitude of w_p seems to suggest that $c = 0.8$ is a plausible value for the Fama-French model. Then, interestingly, there do not seem any gains in terms of investor's utility despite of earlier statistical rejection of the efficiency.

6. SIZE OF THE BOOTSTRAP TESTS

In this section, we examine first the size of the GRS test, W^*, J_1, J_1^*, J_2 and J_2^* under three iid distributions: the standard multivariate normal, multivariate t and multivariate mixture normal. Then, we study the size of the tests under alternative distributions with conditional heteroscedasticity and serial dependence.

6.1. iid case

To compare the size of the tests, we need to generate the model residuals and artificial returns from a given distribution. Choose arbitrarily the parameter estimates of β and Σ over the subperiod from January 1976 to December 1985 as the true parameters. With the given parameters, it is straightforward to generate the data from the multivariate normal distribution. For the multivariate t distribution, we choose 5 and 10 as the degrees of freedom to reflect a varying degree of kurtosis.

The mixture normal distribution is of the following form:

$$\varepsilon_t \sim w\mathcal{N}(0, \Sigma) + (1 - w)\mathcal{N}(\eta, \gamma\Sigma), \tag{38}$$

[8] As pointed out by Gibbons, Ross, and Shanken (1989), the zero-intercept hypothesis only allows one to test the *necessary* condition that the underlying portfolio is mean-variance efficient.

where w $(0 \leq w \leq 1)$ is a mixing-probability parameter, γ is a scale parameter, and η is a non-zero vector. The distribution is skewed and has non-zero mean. To generate data from the mixture distribution, we need to specify w, γ and η. We set $w = 0.7$ and $\gamma = 5$. This implies that a weight of 70 percent is given to data from the standard normal distribution, and 30 percent to data from a non-zero mean normal distribution with 5 times the covariance matrix. To reflect the skewness of the observed return data, η is set as a vector of the standard deviations multiplied by minus one. Since the residuals have zero means, the residuals drawn from the mixture normal distribution will be de-meaned before used in the market model to generate returns data.

The design of the Monte Carlo experiment can be summarized as follows.

1. Let $\hat{\beta}$ and $\hat{\Sigma}$ be a parameter estimate of β and Σ, say from the true data over January 1976 to December 1985. Generate the model residuals from one of the alternative distributions. Then, artificial returns can be computed from

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

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

As for each of the artificial data set, we need to bootstrap the distribution. If there were 10,000 data sets and 10,000 bootstrap draws, there will be $10,000 \times 10,000$ OLS estimations involved in the bootstrap, which is a very time consuming process.[9] As a result, we use only 250 bootstrap samples to conduct the size study. The number of Monte Carlo draws of the data is set to be 1,000.[10]

[9]Calculation of the bootstrap rejection rates of the GMM tests in Table 3 is very time demanding. Each of the distributional scenarios takes more than twenty-four hours on a SUN SPARCstation 20.

[10]The rejection rates even with 100 bootstrap draws tend to be reliable because the data is also generated randomly. Many studies, such as Horowitz (1995), use only 100 bootstrap samples and 1,000 data sets.

TABLE 3.

Size of Tests under Various Distributions

distribution	GRS			W^*			J_1			J_1^*			J_2			J_2^*		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
Panel A: iid case																		
Normal	.010	.051	.105	.008	.047	.102	.032	.126	.207	.008	.047	.102	.007	.045	.091	.010	.055	.103
t(10)	.004	.052	.095	.009	.037	.086	.025	.105	.176	.008	.038	.083	.001	.034	.084	.003	.041	.088
t(5)	.012	.051	.098	.014	.039	.085	.026	.100	.179	.013	.041	.086	.006	.045	.090	.008	.043	.091
Mixture Normal	.026	.093	.148	.012	.050	.102	.069	.159	.249	.011	.053	.095	.025	.081	.143	.014	.049	.097
Panel B: non-iid case																		
MVT10	.016	.057	.113	.005	.037	.081	.038	.111	.180	.004	.033	.080	.006	.047	.088	.006	.035	.070
MVT5	.006	.046	.107	.001	.021	.081	.027	.099	.190	.000	.022	.078	.003	.044	.102	.006	.039	.082
VAR(1)	.048	.118	.170	.023	.083	.148	.095	.186	.243	.024	.081	.139	.033	.071	.104	.015	.048	.080

Consider the market model:

$$R_t = \alpha + \beta r_{pt} + \varepsilon_t; \quad t = 1, \cdots, T;$$

where R_t is a vector of excess returns and r_{pt} the return on the CRSP index. The following asset pricing restrictions,

$$H_0 : \alpha = 0;$$

are tested by using Gibbons, Ross, and Shanken's (1989) test, GRS, the bootstrapped Wald W^* test, the GMM tests, J_1 and J_2, and the bootstrapped GMM tests J_1^* and J_2^*. There are 1,000 data sets for each replication of the bootstrap draw. The bootstrap rejection rates are calculated based on 250 bootstrap draws. For iid model residuals, four distributions are considered: Normal, t distribution with 10 and 5 degrees of freedom, and mixture of normals. For the non-iid case, multivariate distributions with 10 and 5 degrees of freedom and a VAR(1) process are used.

Panel A of Table 3 reports the rejection rates of the GRS test, W^*, J_1, J_1^*, J_2 and J_2^*, at the 1, 5, and 10 percent significance levels, respectively. It is observed that both the GRS and W^* have sizes that are close to their nominal levels under the normal distribution and t distribution with degrees of freedom 10. For example, at the usual 5 percent level, the rejection rate of the GRS test is 5.1 percent, virtually identical to 4.7 percent of the W^* test. However, for the t distribution with degrees of freedom 5 (high kurtosis), both GRS and W^* have close rejection rates, but tend to under-reject at the 10 percent level. In contrast, for the skewed mixture normal distribution, the rejection rates of the GRS test are higher and farther away from the nominal levels. In comparison, W^* has rejection rates very close to the nominal levels. Overall, the bootstrap test W^* is reliable for all the alternative distributions, while the GRS appears sensitive to skewness. The Wald test, W, is not reported in the table as it is well-known that it over-rejects the null. For the GMM test J_1, the rejection rates are about twice the nominal levels under normality, and substantially higher under the mixture normal distribution. In comparison, the bootstrapped test, J_1^*, has very reliable rejection rates for all the alternative distributions, indicating it pays off to bootstrap the GMM test. The blocks for the GMM bootstrap have 3 percent length. A 10 percent length yields similar results. In contrast with J_1, the alternative GMM test, J_2, performs much better for all the distributions. However, J_2 still over-rejects under the mixture normal distribution. As predicted by the theoretical analysis, J_2^* provides reliable rejection rates in all scenarios. In summary, the bootstrap tests do offer reliable refinements over existing tests which might otherwise be unreliable.

6.2. Non-iid case

The GMM tests hold theoretically without making an assumption on a particular distribution of the residuals as long as their distribution satisfies certain regularity conditions. However, to generate the residuals, we must specify a particular form of conditional heteroscedasticity and serial dependence. There are many possible and complex specifications, but we use only a few simple ones here. First, we assume R_t and r_{pt}, $(r_{pt}, R_t')'$, are jointly (iid) multivariate t distributed with the degrees of freedom ν, mean (μ_1, μ_2) and a non-singular covariance matrix V. Partition V as

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

In this case, we obtain a particular form of conditional heteroscedasticity as it is easy to see that the covariance matrix of the asset returns conditional

on the benchmark portfolio is

$$\mathrm{Var}(R_t|r_{pt}) = c\left[1 + (r_{pt} - \mu_1)' V_{11}^{-1}(r_{pt} - \mu_1)/(\nu - 2)\right] V_{22}. \tag{39}$$

where $c = (\nu - 2)/(\nu - 1)$. Notice that the matrix plays the role of the covariance matrix of the residuals in the market model (1). Now, the covariance matrix of the residuals is no longer constant, but heteroscedastic or time-varying in the particular fashion of (39). As ν increases, the multivariate t distribution approaches the multivariate normal distribution and $\mathrm{Var}(R_t|r_{pt})$ converges to the constant matrix $V_{22} = \Sigma$, making the heteroscedasticity less important. In the extreme case of $\nu = +\infty$, it collapses to the iid normality case.

Like the iid case, we use data over January 1976 to December 1985 to estimate (μ_1, μ_2) and V. Given the estimated parameters, Monte Carlo samples are easily generated. However, we have to simulate the returns under the null hypothesis to study the size. To impose the null hypothesis, μ_2 used in the simulations is computed from $\mu_2 = V_{21}V_{11}^{-1}\mu_1$. Then, we can compute the two GMM tests, J_1 and J_2. For interest of comparison, we also compute the GRS test and W^*, although they are not valid theoretically.

Panel B of Table 3 provides the results. It is surprising that both the W^* and the GRS test have rejection rates fairly close to their nominal levels, indicating, at least for the specifications here, that they are fairly robust to the conditional heteroscedasticity introduced by the t distribution. To analyze further the performance, we introduce serial correlations in the model residuals. For simplicity, we use a VAR(1) process estimated from the actual market model. Under the VAR(1) specification, both the GRS test and W^* over-reject the model substantially. Furthermore, although theoretically valid, the GMM test J_1 performs fairly poorly, even worse than the GRS test and W^* test. This appears to be a problem with the sample size. As sample size increases, J_1 must converge theoretically to its exact distribution, and the rejection rates have to approach to their nominal levels. In contrast, J_1^*, performs much better. It is somewhat surprising that J_2 also performs very well. As a result, J_2^* does not offer too much improvement. Overall, it is clear that it pays off to use the bootstrap method whenever possible.

7. CONCLUSION

This paper proposes bootstrap asset pricing tests in the context of testing portfolio efficiency. In contrast to the well-known Gibbons, Ross, and Shanken's (1989) test (GRS), the bootstrap test does not rely on any specific distributional assumption on asset returns. In addition, the bootstrap method provides useful assessment of economic measures of portfolio in-

efficiency. Using monthly returns grouped by size from January 1926 to December 1995, the bootstrap analogues of Wald test and GMM tests do not reject the efficiency of the CRSP value-weighted stock index with respect to the standard ten size portfolios. We also apply the methods to test the well-known Fama and French (1993) three-factor model and find that existing tests tend to over-reject. In either of the applications, rejecting efficiency produces little gain in expected utility maximization if the utility function is quadratic.

There are potentially a great number of applications of the bootstrap method in finance. The method is useful when asset returns have unknown distributions, and useful when the sample size is small. Especially when the sample size is limited, asymptotic distributions of an asset pricing test may not provide good approximations of the true distribution, and may not be available at all. But, as shown by Hall (1992) and Hall and Horowitz (1996), the bootstrap method can often provide approximations with higher-order accuracy than existing asymptotic tests. The bootstrap procedures may also be appealingly applied to multi-factor models, such as the APT model, and to multivariate event studies. In addition, the bootstrapped GMM test, like the GMM test itself, can be widely used for testing stochastic discount factor models and term structure models.

REFERENCES

Affleck-Graves, J. and B. McDonald, 1989. Nonnormalities and tests of assets pricing theories. *Journal of Finance* **44**, 889–908.

Efron, B., 1979. Bootstrap methods: Another look at the Jacknife. *Annals of Statistics* **7**, 1–26.

Ferson, W. E. and S. R. Foerster, 1994. Finite sample properties of the generalized method of moments in tests of conditional asset pricing models. *Journal of Financial Economics* **36**, 29–55.

Gibbons, M., S. A. Ross and J. Shanken, 1989. Testing the efficiency of a given portfolio. *Econometrica* **57**, 1121–1152.

Hall, P., 1992. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Hall, P. and J. Horowitz, 1996. Methodology and theory for the Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* **64**, 891–916.

Hansen, L. P., 1982. Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

Harvey, C. R. and G. Zhou, 1990. Bayesian inference in asset pricing tests. *Journal of Financial Economics* **26**, 221–254.

Horowitz, J., 1995. Bootstrap methods in econometrics: Theory and numerical performance. In: *Advances in Economics and Econometrics: Theory and Applications* **III**, edited by D. M. Kreps and K. F. Walls, 188–222, Cambridge University Press.

Jeong, J. and G. S. Maddala, 1993. A perspective on application of bootstrap methods in econometrics. In: *Handbook of Statistics* **11**, Edited by G. S. Maddala, C. R. Rao, and H.D. Vinod, 573–600, North-Holland.

Kandel, S. and R. F. Stambaugh, 1989. A mean-variance framework for tests of asset pricing models. *Review of Financial Studies* **2**, 125–156.

Kandel, S., R. McCulloch and R. F. Stambaugh, 1995. Bayesian inference and portfolio efficiency. *Review of Financial Studies* **8**, 1–53.

Kothari, S. P. and J. Shanken, 1997. Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics* **44**, 169–203.

Lyon, J., B. Barber and C. Tsai, 1999. Improved methods for tests of long-run abnormal returns. *Journal of Finance* **54**, 165–201.

MacKinlay, A. C. and M. Richardson, 1991. Using generalized method of moments to test mean-variance efficiency. *Journal of Finance* **46**, 511–527.

Manski, C. F., 1988. *Analog Estimation Methods in Econometrics*. Chapman and Hall, New York.

McCulloch, R. and P. Rossi, 1990. Posterior, predictive, and utility-based approaches for testing the arbitrage pricing theory. *Journal of Financial Economics* **28**, 7–38.

Richardson, M. and T. Smith, 1993. A test for multivariate normality in stock returns. *Journal of Business* **66**, 295–321.

Roll, R., 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.

Shanken, J., 1985. Multivariate tests of the zero-beta CAPM. *Journal of Financial Economics* **14**, 327–348.

Shanken, J., 1987. A Bayesian approach to testing portfolio efficiency, *Journal of Financial Economics* **19**, 195–216.

Shanken, J., 1990. Intertemporal asset pricing: An empirical investigation. *Journal of Econometrics* **45**, 99–120.

Shanken, J., 1996. Statistical methods in tests of portfolio efficiency: a synthesis. In: *Handbooks of Statistics: Statistical Methods in Finance* **14**, edited by G. S. Maddala and C. R. Rao, 693–711, North-Holland.

Shanken, J. and G. Zhou, 2006. Estimating and Testing Beta Pricing Models: Alternative Methods and Their Performance in Simulations. *Journal of Financial Economics*, forthcoming.

Shao, J. and D. Tu, 1995. *The Jacknife and Bootstrap*. Springer Verlag, New York.

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, North-Holland.

Zhou, G., 1993. Asset pricing tests under alternative distributions. *Journal of Finance* **48**, 1925–1942.

Zhou, G., 1994. Analytical GMM Tests: Asset pricing with time-varying risk premiums. *Review of Financial Studies* **7**, 687–709.