Multivariate Tests of Asset Pricing: The Comparative Power of Alternative Statistics

John Affleck-Graves and Bill McDonald*

Abstract

This paper examines estimation issues associated with multivariate tests of asset pricing. Two issues are considered: (1) the constraint that the sample size (N) must be less than the time series (T), and (2) the relative effect on power of using the multivariate statistic versus a univariate counterpart. We find that an alternative statistic that allows for large N does not dominate the usual portfolio tests. More notably, we find that the power of a simple diagonal statistic usually dominates the multivariate statistic for cases considered in this study.

I. Introduction

Recently tests of the CAPM have focused on multivariate approaches rather than the traditional tests such as those performed by Fama and MacBeth (1973) and Black, Jensen, and Scholes (1972).[1] Although the multivariate approach overcomes many of the limitations of previous methods, there are still a number of unique considerations associated with the empirical power of these tests. In this study, we consider a multivariate test recently proposed by Gibbons, Ross, and Shanken (1989). Two issues concerning the estimation and power of the test are addressed. First, a major constraint of the multivariate test is the restriction that the number of assets tested (N) must be smaller than the time series of observations ($T-2$, with 2 being an adjustment for degrees of freedom). Our results suggest that a maximum entropy statistic that insures the nonsingularity of the residual covariance matrix is a viable alternative in the case where $N \geqslant T-2$; however, the results also indicate that there may be little or no gain in disaggre-

* Both authors, College of Business, University of Notre Dame, Notre Dame, IN 46556. This research was funded in part by the Transamerica Fund and the Business Partners Fund. Data were provided by the Jesse H. Jones Research Data Base. The authors are grateful to Jay Shanken, an anonymous *JFQA* referee, and workshop participants from the University of Illinois and the University of Notre Dame for helpful comments.

[1] The term "multivariate" is not clearly defined in the statistical literature and is used somewhat indiscriminately in the asset pricing literature. In our application, we are referring specifically to those asset pricing tests that consider N return vectors of T observations in relation to a single vector of T observations representing the market surrogate. In most cases, these multivariate tests attempt to incorporate interdependencies between the corresponding regression error vectors. Although the sequential procedure of Fama-MacBeth can be considered "multivariate," it is not analogous to the previously described methods frequently considered in the realm of multivariate statistical analysis. Our usage is consistent with previous applications, specifically beginning with Gibbons (1982).

gating the sample from a small number of portfolios to a large number of individual securities. Second, it is not clear from an empirical standpoint that the power of the multivariate statistic will dominate a simple aggregate of univariate statistics due to the substantial number of parameters that must be estimated from a fixed information set for the multivariate statistic, and the precision with which they can be estimated. We find that for the cases examined here, a test that ignores the off-diagonal information in the residual covariance matrix generally dominates the exact multivariate statistic.

A number of multivariate methods have been proposed for testing the mean-variance efficiency of a given portfolio. Gibbons (1982) suggests a nonlinear multivariate regression model that increases the precision with which the risk premiums can be estimated. Subsequent studies have questioned whether the asymptotic results used by Gibbons were adequate, given the small sample sizes. For example, Stambaugh (1982) presents simulation results indicating that use of the likelihood ratio test results in too frequent rejection of the null hypothesis, especially when the number of assets is large relative to the time series of observations. Using a cross-sectional regression test, which he shows to have an asymptotic Hotelling's T^2 distribution, Shanken (1985) rejects the efficiency of the CRSP equal-weighted index at the 0.01 level.

Most of the above tests are based on asymptotic theory. Gibbons, Ross, and Shanken (GRS) (1989), assuming a given risk-free rate in each period and excess returns that are multivariate normal, propose a test statistic for which the exact small sample distribution is known. This statistic provides an unambiguous test of the Sharpe-Lintner CAPM although, as GRS illustrate, the manner in which the securities are grouped into portfolios can affect the magnitude of the test statistic and the conclusions of the test. While GRS's test statistic has the advantage of a known distribution for finite sample sizes, MacKinlay (1987) shows that it may have low power in some cases.[2]

The power of the GRS test could be improved by increasing the number of right-hand side assets (N). Since the test statistic requires an estimate of the covariance matrix of the residuals (Σ) from the excess returns market model, increasing the number of right-hand side assets presents a number of estimation problems. The test assumes stationarity of excess returns and thus limits the number of time periods (T) that can be considered. Clearly, to obtain a nonsingu-

[2] As the focus of this paper is on the comparative power of several alternative statistics, some preliminary comments on our terminology are in order. The power of a given test of a null hypothesis is the probability that the null will be rejected, which is a function of the sample size, the significance level (or size) of the test, and the specified distribution from which the sample is drawn (i.e., the true state of nature). While changes in the sample size and the significance level of the test will affect the absolute level of the power of the various statistics examined, such changes are unlikely to affect their relative power. Thus, we arbitrarily fix the sample size (at 60 months) and the significance level (at 5 percent). The specification of the alternative hypothesis will also affect the absolute value of the power. Consistent with MacKinlay (1987), we specify an alternative where the annualized measurement error in the risk-free rate is 10 percent. Once again, while changes in the magnitude of the measurement error will affect the absolute magnitude of power, they should not affect the relative power (see MacKinlay). When discussing our results below, we should always refer to the power for a sample size of 60 observations at a significance level of 5 percent given the specified alternative hypothesis of a 10 percent per annum measurement error in the risk-free rate. However, as the results are unlikely to be affected by these particular choices, and for the sake of brevity, we will merely refer to the power of the statistic.

lar estimate of Σ, N must be less than $T-2$. Indeed, it is preferable that N be considerably smaller than T, and this restricts the number of right-hand side assets that can be examined. Even if the number of observations for an asset (T) could be dramatically increased, the need to invert the $N \times N$ covariance matrix of residuals would effectively limit the size of N.[3]

To overcome the above problems, most tests have combined individual assets into portfolios. A typical example would be $N = 20$ and $T = 60$ months. Under this condition, the power of the test can be low because of the relatively few number of assets included. In addition, the asymptotic results are not applicable since T remains small. Most importantly, because portfolios are used, it is possible that grouping may further lower the power (see Roll (1977)). It is important to examine the comparative power of a test that will enable a large number of individual securities to be simultaneously examined without having to increase the time period (T) and, thereby, violate the stationarity assumption.

Two possible solutions to the above problem are to use either an alternative estimation procedure that results in a nonsingular estimate of Σ even when $N \geqslant T-2$, or to assume some simplified structure for the covariance matrix of residuals. In this paper, both alternatives are examined. In the first case, the maximum entropy (ME) approach (see Theil and Laitinen (1980)) is used to obtain an estimate of the covariance matrix that is nonsingular even when $N \geqslant T-2$. This overcomes the nonsingularity problem, but not necessarily the inversion problem if N is exceptionally large. In the second case, the covariance matrix is assumed to be diagonal, overcoming both the estimation and the inversion problems.

In addition to the issue of singularity, the diagonal covariance approach allows a second issue to be examined concerning the informational contribution of the covariance structure to the test statistic. Even for the case where the multivariate statistic is well-defined, it is not clear that the estimated covariance structure contributes to the power of the statistic. This is attributable to the relatively large number of parameters that must be estimated for any given sample size. An approximate statistic (i.e., assuming a diagonal covariance matrix) may result in a more powerful test than the exact test. This issue is examined over a variety of covariance structures empirically observed during the 1931–1985 period.

The paper is arranged as follows. In the next section, the basic GRS methodology is summarized. Section III presents a discusson of the methodologies used in both the ME and diagonal estimation procedures. The fourth section describes the simulation design used to compare the empirical power of the tests. The methods are illustrated by examining the ex ante efficiency of the CRSP

[3] Our tests focus on the relative power of the alternative statistics including the effect of changing N. Another dimension to consider is shortening the measurement interval to increase T (without expanding the underlying time interval). Note, however, that the increase in the number of observations resulting from the use of a shorter measurement interval is offset by the decrease in the measurement error (γ in Equation (13)) necessary to maintain the same "economic" deviation from the null hypothesis. Theoretically, such a change will not alter the value of the noncentrality parameter. However, there will be a small increase in power due to the increased degrees of freedom for the F statistic. Additional problems associated with this approach are that, as the measurement interval is shortened, we know less about the independence and distributional properties of the returns. Also, as the interval decreases, we face the problem of nonsynchronous data. This issue is not a predominant concern in our study, since we are primarily concerned with the relative power of certain tests for a given T.

equal weighted index over six nonoverlapping five-year periods corresponding to the sample intervals in MacKinlay (1987). Section V examines the statistical power of the proposed tests. The power of the proposed tests is compared to that of the GRS statistic, and simulation evidence is presented to verify that the results are not attributable to the modified bootstrap method employed in the study. Concluding comments are provided in the final section.

II. The GRS Test

In deriving an exact test for the ex ante efficiency of a given portfolio, GRS assume the existence of a riskless rate of interest and that excess asset returns are independently and identically multivariate normally distributed through time. This is sufficient to imply that returns can be described by the excess return market model

$$(1) \qquad \tilde{r}_{it} = \alpha_{ip} + \beta_{ip}\tilde{r}_{pt} + \tilde{e}_{it}, \qquad i = 1,2,\ldots,N, \quad t = 1,2,\ldots,T,$$

where $\tilde{r}_{it}$ = the excess return on asset i in period t,
$\tilde{r}_{pt}$ = the excess return on portfolio p, whose efficiency is being tested,
N = the number of assets being tested, and
$\tilde{e}_{it}$ = the disturbance term for asset i in period t.

The assumptions also imply that

$$E(e_t) = 0, \text{ and}$$

$$E(e_t, e_s) = \begin{cases} \Sigma & t = s \\ 0 & t \neq s, \end{cases}$$

where Σ is the $N \times N$ disturbance covariance matrix.

Using standard multivariate statistical theory, GRS demonstrate that testing a particular portfolio for mean-variance efficiency is equivalent to testing the null hypothesis

$$H_0 : \alpha_{ip} = 0, \qquad i = 1,2,\ldots,N.$$

Clearly, if ordinary least squares is used to estimate Equation (1) for each asset, then

$$(2) \qquad \hat{\alpha}_p \sim \text{MVN}\left[\alpha_p ; T^{-1}\left(1 + \hat{\theta}_p^2\right)\Sigma\right],$$

where $\hat{\alpha}_p' = (\hat{\alpha}_{1p}, \hat{\alpha}_{2p}, \ldots, \hat{\alpha}_{Np})$,
$\hat{\theta}_p = \bar{r}_p / s_p$,
$\bar{r}_p$ = the sample mean of $\tilde{r}_{pt}$, and
s_p^2 = the maximum likelihood estimate of the variance of $\tilde{r}_{pt}$.

In addition,

$$T\left(1 + \hat{\theta}_p^2\right)^{-1} \hat{\alpha}_p' \Sigma^{-1} \hat{\alpha}_p \sim \chi_N'^{\,2},$$

where χ'^2_N is a noncentral χ^2 distribution with N degrees of freedom. In general, Σ is not known but, conditional on r_{pt} (for $t = 1, 2, \ldots, T$), GRS show that

$$\Gamma_G = [T/(T-2)]\,[(T-N-1)/N]\,W \sim F'_{N;T-N-1}, \tag{3}$$

where $W = (1+\hat{\theta}_p^2)^{-1}\hat{\alpha}_p'\hat{\Sigma}^{-1}\hat{\alpha}_p$,

$\hat{\Sigma}$ = the unbiased sample covariance matrix of the residuals,

F' = the noncentral F distribution with noncentrality parameter λ, and

$\lambda = T(1+\hat{\theta}_p^2)^{-1}\alpha_p'\Sigma^{-1}\alpha_p$.

Under the null hypothesis λ is zero, Γ_G has a central F distribution, and the test of the null hypothesis is the uniformly most powerful invariant test and is the likelihood ratio test (see Muirhead (1982), p. 212). In addition, because the noncentrality parameter is zero under the null hypothesis, independent of the market return, the central F is also the unconditional distribution of Γ_G (see GRS). Note that a uniformly most powerful invariant test is most powerful among all the tests that preserve the original parameter space of the class of models under consideration. If knowledge or assumptions reduce the parameter space, then the uniformly most powerful test in the larger parameter space often will be less powerful than tests that work within the smaller parameter space. In our case, the GRS test is uniformly most powerful in the class of invariant tests that do not restrict the parameter space with respect to Σ. If certain of the elements in Σ are known (or assumed known), then the GRS test need not be more powerful as the parameter space of the problem has been reduced.[4] The tradeoff between power and the imposition of restrictions on Σ is the focus of this paper.

We consider the Sharpe-Lintner model as the null hypothesis being tested against an alternative where the risk-free rate is measured with some prespecified level of error. For this particular alternative hypothesis, the test focuses on the extent to which the market portfolio (which is on the efficient frontier) deviates from the tangency portfolio. It is important to stress that we do not examine the power to reject mean-variance efficiency against the myriad other reasonable alternatives possible; however, the magnitude of the difference in power obtained suggests the importance of considering this issue when testing other alternative hypotheses.

III. Alternatives to the GRS Statistic

While the GRS test provides an exact test of the ex ante efficiency of a given portfolio, its use is limited because of the stationarity assumption implicit in Equation (1). This effectively limits T to approximately five years, which in turn limits N, as Equation (3) requires N to be less than $T-2$. In an attempt to resolve this problem, GRS resort to using portfolios rather than individual securities, forcing N to be substantially less than T. Unfortunately, the power of the statistic

[4] A simple example of this tradeoff is the univariate test that the mean (μ) of a normal population is zero against the alternative that $\mu > 0$. If the variance is unknown and must be estimated, then the traditional t-statistic is the uniformly most powerful unbiased test. However, if the parameter space is reduced because the variance is known, then the z-statistic is uniformly most powerful (i.e., z is more powerful than t because the variance is known or assumed known).

could be substantially compromised if the portfolio formation criterion averages out systematic deviations from the null hypothesis observable in individual securities. There is the additional concern that the statistic might lack power due to the overparameterization associated with estimating each element of the residual covariance matrix.

In the following two sections, we introduce a method that allows for estimation when $N \geqslant T-2$, and then derive a second statistic that ignores off-diagonal terms and compare it to its multivariate counterpart.

A. A Maximum Entropy Statistic

One alternative to the standard estimation approach is to use the maximum entropy method (see Theil and Laitinen (1980)) in estimating Σ, which results in a positive definite estimated covariance matrix even when $N \geqslant T-2$. The maximum entropy method estimates Σ subject to two constraints, a mass-preserving constraint and a mean-preserving constraint (i.e., a constraint on the moments of order zero and one, respectively). It is thus a constrained maximum-entropy estimator of Σ (the method is briefly described in the Appendix). Essentially, the method assigns each observation to an interval in a manner that increases the dimensionality of the problem such that the residual covariance matrix is no longer singular.

Replacing $\hat{\Sigma}$ in (3) by $\tilde{\Sigma}$, the maximum-entropy estimator, an analogous statistic to the GRS statistic can be computed as

$$\Gamma_{\mathrm{ME}} = [T/(T-2)]\,[(T-N-1)/N]\left(1+\hat{\theta}_p^2\right)^{-1}\hat{\alpha}_p'\tilde{\Sigma}^{-1}\hat{\alpha}_p\,. \tag{4}$$

Unfortunately, the distribution of Γ_{ME} under the null hypothesis is unknown. In fact, even if we were willing to assume that the distribution of Γ_{ME} is the same as Γ_{G}, the denominator degrees of freedom of the F statistic $(T-N-1)$ would be negative when N is greater than $T-1$, precluding us from using the test. As subsequently detailed, the lack of a known underlying distribution restricts the tests of statistical power to nonparametric simulation techniques.

B. A Constrained Residual Covariance Matrix Approach

By constraining the off-diagonal elements of the residual covariance matrix, we can examine the effects of trading-off information in order to substantially reduce the dimensionality of the parameter estimation problem. A constrained residual covariance matrix is another means of avoiding the singularity of Σ when $N \geqslant T-2$. Unfortunately, the residual correlations reported in GRS do not suggest a pattern or structure (e.g., a constant correlation or band matrix approximation) that would simplify the computational procedure. Thus, we use the limiting case of structuring the covariance matrix by assuming that the off-diagonal elements are zero—that is, Σ is a diagonal matrix. Define the statistic

$$\Gamma_D = k\hat{\alpha}_p'\hat{\Sigma}_D^{-1}\hat{\alpha}_p\,, \tag{5}$$

where $\hat{\alpha}_p$ = the OLS estimates of the α_{ip}'s,

$\hat{\Sigma}_D$ = the estimated covariance matrix of the residuals—with zeros in the off-diagonal and $\hat{\sigma}_i^2$ on the diagonal,

$\hat{\sigma}_i^2$ = the OLS estimate of the variance of e_i,

and k is a constant discussed in the next section that equates the expected value of Γ_D with the expected value of Γ_G. Except for a constant, Equation (5) is equivalent to

$$N^{-1}\sum_{i=1}^{N} t_i^2 , \tag{6}$$

where t_i is the t-statistic computed when testing H_0: $\alpha_i = 0$. Asymptotically, (i.e., as $T \rightarrow \infty$), each t_i will be a standardized normal variate and thus $N \cdot \Gamma_D$ will be asymptotically chi-squared with N degrees of freedom.

This is based on the assumption that the true Σ is in fact diagonal. If Σ is not diagonal, then we need to find the distribution of Γ_D, i.e., the statistic calculated using a diagonal estimate of a nondiagonal covariance matrix. In this case, the above statistical reasoning is clearly invalid because Γ_D is now the sum of nonindependent t^2 statistics. The distribution of Γ_D is unknown and dependent on the population Σ structure. Once again, this lack of an underlying distribution will constrain us, in our subsequent tests, to simulation-based methods.

The diagonal statistic is somewhat counter to the spirit of the multivariate test since it does not take into account any interrelationship between securities other than their relation through the chosen portfolio. The critical question addressed by comparing the diagonal statistic to the GRS statistic is whether there is sufficient information in any reasonable sample of N and T to estimate the substantially larger number of parameters required in the GRS test, such that the estimates of these parameters are sufficiently precise to increase the power of the test.[5]

C. The Relationship between Γ_D and Γ_G

The GRS statistic, while providing an exact test of H_0: $\alpha_p = 0$, suffers from the problem of requiring estimation of the full covariance matrix Σ. Consequently, the sampling variation of the GRS statistic may be high, resulting in a test of relatively low power. The approach proposed in the previous section of overcoming this problem is to assume specific values for the off-diagonal elements of Σ. A simple structure such as the above is to assume that all off-diagonal elements are zero. This reduces the estimation problem to only considering the N diagonal elements of Σ. Thus, we have the statistic defined in Equation (5) where k is a constant chosen such that under the null hypothesis[6]

$$E(\Gamma_D) = E(\Gamma_G) . \tag{7}$$

[5] For example, where $N = 20$ and $T = 60$, the diagonal statistic estimates only 60 parameters from 1,200 pairs of observations, while the GRS statistic requires estimates of 252 parameters.

[6] Since our bootstrap methodology estimates the empirical distribution under the null hypothesis in order to determine the critical (5 percent) point, we make no assumptions about the distribution of the statistic Γ_D nor do we use any classical statistical theory. Consequently, the location of the distri-

It is straightforward to show that for this condition to be met,

$$k = [(T-4)/(T-N-3)]\,[T/(T-2)]\,[(T-N-1)/N]\left(1+\hat{\theta}_p^2\right)^{-1}.$$

To provide some insight as to the possible benefits of Γ_D, we examine the relative efficiency of the two statistics Γ_D and Γ_G. To do this, it is necessary to compute their variances. In the case of Γ_G, this is straightforward as Γ_G is known to have a central F distribution under the null hypothesis. Thus,

$$\text{(8)}\quad \mathrm{Var}\left(\Gamma_G\right) = \left[2(T-N-1)^2(T-3)\right]/\left[N(T-N-3)^2(T-N-5)\right].$$

The variance of Γ_D is more difficult to compute since we do not know the distribution of Γ_D under the null. Proceeding from first principles,

$$\text{(9)}\quad \begin{aligned}\mathrm{Var}\left(\Gamma_D\right) &= \mathrm{Var}\left[k\hat{\alpha}_p'\hat{\Sigma}_D^{-1}\hat{\alpha}_p\right]\\ &= k^2\left[N\,\mathrm{Var}\left(z_i\right) + 2\sum_{i<j}\sum \mathrm{Cov}\left(z_i,z_j\right)\right],\end{aligned}$$

where $z_i = (\hat{\alpha}_{ip}^2/\hat{\sigma}_{ip}^2)$.

The second term in Equation (9) requires calculation of the covariance between two $F_{1;T-2}$ random variables. To obtain an indication of the extent to which $\mathrm{Var}(\Gamma_D)$ may differ from $\mathrm{Var}(\Gamma_G)$, consider the case where the population Σ is in fact diagonal so that the estimates of α_{ip} are independent. Then, the second term in Equation (9) will be zero and

$$\mathrm{Var}\left(\Gamma_D\right) = k^2 N\,\mathrm{Var}\left(z_i\right).$$

Since $T(1+\hat{\theta}_p^2)^{-1}z_i \sim F_{1;T-2}$, it follows that

$$\mathrm{Var}\left[T\left(1+\hat{\theta}_p^2\right)^{-1} z_i\right] = \left[2(T-2)^2(T-3)\right]/\left[(T-4)^2(T-6)\right].$$

Therefore,

$$\text{(10)}\quad \mathrm{Var}\left(z_i\right) = \left[\left(1+\hat{\theta}_p^2\right)^2 \Big/ T^2\right]\left[2(T-2)^2(T-3)\right]/\left[(T-4)^2(T-6)\right].$$

Substituting Equation (10) and the value of k into Equation (9) we obtain

$$\text{(11)}\quad \mathrm{Var}\left(\Gamma_D\right) = \left[2(T-N-1)^2(T-3)\right]/\left[N(T-N-3)^2(T-6)\right].$$

Notice that $\mathrm{Var}(\Gamma_D)$ is defined for all combinations of N and T, whereas $\mathrm{Var}(\Gamma_G)$ is defined only when $N < T-5$.

---

bution is not of importance (i.e., the constant k can neither help nor hinder the statistic's performance). For comparative purposes, we locate the distribution of Γ_D at the same mean as that of Γ_G. (The bootstrap method is subsequently discussed in more detail.)

Under the null hypothesis, the efficiency of Γ_D relative to Γ_G is

$$\text{(12)} \qquad \mathrm{Var}(\Gamma_D)/\,\mathrm{Var}(\Gamma_G) = (T-N-5)/(T-6)\ .$$

Clearly, if the population Σ is diagonal, then Γ_D has lower variance than Γ_G and is more efficient. This suggests that Γ_D may be a more powerful test statistic (under these conditions) than Γ_G (although the power clearly depends on higher order moments as well).[7] Note also that in the range of T and N used empirically, this difference in efficiency is not small. For example, if $N = 20$ and $T = 60$, then Equation (12) shows that the variance of Γ_G will be 1.54 times the variance of Γ_D.

If the population Σ is not diagonal, the variance of Γ_D will increase as the second term of Equation (9) increases.[8] However, the variance of Γ_G will be unaffected and the relative efficiency of Γ_D will decline as the population Σ is increasingly different from diagonal. Thus, the choice between Γ_D and Γ_G rests on the assumptions one is willing to make about how close the population Σ is to diagonal. On the one hand, Γ_G provides an exact test but with fairly high sampling variation. On the other hand, Γ_D can provide a test for which the sampling distribution is unknown but which has lower variability in some cases. Whether or not the net result is a more or less powerful test depends on the extent to which our assumed diagonal structure violates the true population structure and the number of sample observations available. In our test, we examine this tradeoff between the efficiency and precision of the test statistic using covariance structures typically present in market model residual data.

IV. The Simulation Design

A. The Parametric and Nonparametric Simulations

From the previous discussion, we are interested in comparing the alternative test statistics in both the standard case where $N < T-2$ and for the extended sample where $N \geqslant T-2$. We focus on what is equivalent to Case I in MacKinlay (1987), where the risk-free rate is assumed to be measured with error (γ). Given

$$\text{(13)} \qquad \alpha_p = \gamma(\iota - \beta_p)\ ,$$

where ι is a $(N \times 1)$ vector of ones, the null hypothesis is that $\gamma = 0$. From MacKinlay's results, we know that the power of the GRS statistic will be relatively low, thus we set the annualized measurement error equal to 10 percent in

[7] Of course, since the two statistics do not have identical distributions, the statistic with lower variance will not necessarily have higher power. The power will depend on the higher order moments (in addition to the first and second moments), and on the mapping of the original sample points to the test statistic. However, since this mapping appears similar in the case of Γ_G and Γ_D, we present the analysis of relative efficiency to demonstrate a possible way in which Γ_D might yield a more powerful test than Γ_G. (Again, this is similar to the univariate z-statistic, assuming the variance is known, having smaller variance than the univariate t-statistic and providing a more powerful test.)

[8] Evaluation of the second term ($\mathrm{Cov}(z_i,z_j)$) requires working with the bivariate t distribution of Siddiqui (1967). Following Johnson and Kotz ((1972), pp. 133–135), it can be shown that, under reasonable conditions, the covariance term will be nonnegative regardless of the sign of the correlation between $\hat{\alpha}_i$ and $\hat{\alpha}_j$. This is analogous to the result that $E(X_2^2, X_1^2) = 1 + 2\rho^2$, if $(X_1;X_2)$, has a bivariate standardized normal distribution (see Johnson and Kotz (1972), p. 91).

all tests of the simulated alternative (i.e., $\gamma \approx 0.008$, which is one of the values used in MacKinlay). We also fix the significance level at 5 percent throughout and calculate the power of the tests for this level of significance.[9]

Recall that we do not know the underlying distribution for the Γ_{ME} and Γ_D statistics. For cases where $N < T-2$, we can, by exploiting the multivariate normal assumption, use a standard simulation approach and generate multivariate normal errors with a desired covariance structure. By generating the empirical distribution of Γ_{ME} and Γ_D under the null ($\gamma = 0$) and a selected alternative ($\gamma = c$, where c is a constant), the power of the tests can be empirically evaluated.[10]

For cases where $N \geqslant T-2$, however, the singularity of the estimated residual covariance matrix once again presents difficulties, as both conditional distribution and decomposition methods of generating a multivariate normal sample of residuals from a given covariance matrix require the matrix to be nonsingular. Thus, standard parametric simulation techniques cannot be used to examine the case where $N \geqslant T-2$.

To avoid this problem, we use a modified bootstrap method. This procedure can be used to obtain the empirical distribution of the test statistics for any sample of N assets in any time period. In the multivariate case, we are faced with two alternative methods of bootstrap sampling. In one case, we could randomly sample securities from the data in the typical manner of bootstrapping. Note, however, that since bootstrap sampling is done with replacement there is a substantial probability that the same security will be selected more than once. In this case, the covariance matrix is obviously singular. To avoid this problem, we adopt an alternative approach that relies on the assumed temporal independence of the errors. In our bootstrap method, we randomly select, across all securities, single time periods with replacement until a sample with T observations is obtained. This process retains the covariance structure of the population and eliminates the singularity problem. Another benefit is that the sample of securities remains the same. Since we do not have theoretical results, we cannot assume that the distribution of the statistics will be the same for all N, T, and for any covariance structure. Consequently, this procedure must be executed for each specific case examined.

The bootstrap approach applied in this study is briefly outlined below.

(1) For each of the N securities, Equation (1) is individually estimated using ordinary least squares and the T observations available. This produces the OLS estimates $\hat{\alpha}_{ip}$, $\hat{\beta}_{ip}$, and a $T \times N$ matrix of estimated residuals, E.

(2) Define G^* as the discrete probability distribution that places probability $1/T$ on each row of E computed in step 1.

(3) A random sample of size T is drawn from G^*—i.e., the rows of E are resampled independently with replacement. Denote this sample by e_{it}^*; $i = 1, 2, \ldots, N$, and $t = 1, 2, \ldots, T$.

[9] The choice of 5 percent is purely conventional and is not an implied recommendation of the significance level appropriate in portfolio efficiency tests. Shanken (1988) provides an interesting discussion of the appropriate choice of significance level in such tests.

[10] Clearly, we cannot assume that Γ_{ME} and Γ_D have the same distribution (and degrees of freedom) as Γ_G and use the critical value of the central F distribution in hypothesis testing, as this would bias the size of the test and the resulting p value.

(4) Set $\alpha_{ip} = \gamma(1-\hat{\beta}_{ip})$ and $\beta_{ip} = \hat{\beta}_{ip}$ and construct a "pseudo-sample" of T returns for each security as follows

$$r_{it}^{*} = \alpha_{ip} + \beta_{ip} r_{pt} + e_{it}^{*} ,$$

where the r_{pt}'s are the same as in the original estimation sample.

(5) Equation (1) is now reestimated by OLS using the pseudo data (r_{pt}, r_{it}^{*}) to obtain estimates $\alpha_{ip}^{*}, \beta_{ip}^{*}$, and Σ^{*}.

(6) Compute the test statistics Γ_{ME} and Γ_{D} from Equations (4) and (5), respectively.

(7) Steps 3 to 6 are repeated a large number of times to obtain a bootstrap distribution of the statistics under the null ($\gamma = 0$) and alternative hypotheses ($\gamma = c$).[11] The null distribution can be used to determine critical values for a given significance level, with the alternative distribution indicating the empirical power of the statistic.[12]

Notice that the bootstrap method is a nonparametric simulation technique. The obvious disadvantage of using a nonparametric procedure is the potential loss of power in comparison to a parametric procedure (in the case where the parametric assumptions are appropriate). Our proposed statistics do not have a known distribution, thus this tradeoff is not a matter of choice. A positive artifact is that bootstrapping avoids the assumption of multivariate normality. Also, since the critical value is obtained from the bootstrap distribution under the null hypothesis, the size of the test is equal to the significance level by construction. Consequently, the power comparisons are more meaningful since differences cannot be caused by distributional problems or misspecification of the size of the test.

B. Data and Simulation Parameters

The alternative statistics for testing the efficiency of a given portfolio are illustrated using test periods that parallel those used in MacKinlay (1987). The tests are conducted using five years of monthly data from the six consecutive periods occurring from 1954 to 1983. In cases where only portfolios are being examined, all securities with complete data are assigned to 20 beta-sorted portfolios. For cases where the singularity issue is being addressed, 240 securities are randomly sampled without replacement from the population of securities with complete data. Although using the entire CRSP sample for any of the periods still

[11] For bootstrap estimates of the null distribution, the process is repeated 2,000 times. Subsequent tests of the alternative hypothesis are based on 1,000 repetitions. The structure of our experiments is dictated somewhat by the extraordinary computational expense of examining a large sample of individual securities (specifically, inverting the maximum entropy covariance matrix 3,000 times for each test in each period).

[12] Bickel and Freedman (1981) provide some asymptotic theory for the bootstrap method. In particular, they show that (asymptotically) the bootstrap works for the familiar t-statistic, including extensions to multidimensional data. However, our results are based on finite sample sizes and finite bootstrap sampling. Hence, the results must be subject to the caveat that the underlying sample is representative of the population studied.

The bootstrap methodology assumes that the variables are i.i.d. By fixing all securities and bootstrapping across time periods, we only require temporal independence of residual returns to satisfy the i.i.d. assumption.

provides a computationally invertible covariance matrix, the problem is too expensive for the large number of iterations necessary in the simulation. Thus, the sample is fixed at 240 randomly selected securities in each period. The same random sample of 240 securities is used for a given period throughout the sequence of tests. As in MacKinlay, the tested portfolio p is the CRSP equal weighted index. Portfolios are formed by ranking securities on the beta calculated from the 60 months prior to the sample period.[13] The riskless rate is measured using the U.S. Treasury Bill rate reported on the CRSP Indices file.

V. Results

The results in Table 1 provide a general characterization of the samples and a benchmark comparison with the data used by MacKinlay (1987). The number of securities with complete data in each period ranges from 739 to 1,008. The results in Table 1 for the standard application of the Γ_G statistic to portfolios are generally consistent with MacKinlay's; however, there are minor differences. Our samples differ for two reasons: (1) we use a newer version of the CRSP tape that is continually revised to correct data errors, and (2) we use only prior period betas—MacKinlay includes a security if either a prior or post period beta is available. As will be shown, the power results for the GRS statistic are sample sensitive and so the minor differences in Table 1 are not surprising.

The theoretical power of the GRS statistic ranges from 51 percent in the 1954–1958 period to less than 10 percent in the 1974–1978 period.[14] As in MacKinlay, the efficiency of the equal weighted index, using an unspecified alternative, can be rejected at the 0.05 level only in the first sample period.

A. Preliminary Simulation Results

In this section, results that provide some benchmarks for the simulation environment are presented. In Table 2, the basic GRS statistic applied to portfolios formed from all securities with complete data and aggregated across time periods, is used to examine the initial simulations. The descriptive statistics (first four sample moments) associated with the theoretical null distribution and the simulated null distributions using simulated normal errors are virtually identical, which provides some indication of internal consistency for the methodology used in this study. As expected, the mean of the statistic under the alternative hypothesis is much larger than the mean under the null.

---

[13] For comparison with MacKinlay (1987) and Fama and MacBeth (1973), the equal weighted index is the portfolio whose efficiency is tested throughout the paper. Note, however, that if the value weighted index is to be tested, the comparisons between the GRS statistic and the diagonal model are potentially biased in favor of the multivariate test. Frequently, in testing the value weighted index, the corresponding weighting vector for the asset returns is ignored. To the extent that it is, much of this information will be captured in the off-diagonal elements of the residual matrix, thus increasing the relative power of the multivariate test.

For cases where the total sample of securities (N) is not evenly divisible by 20, the first k portfolios contain one additional security, with $k = N - \text{INT}(N/20)*20$.

[14] Theoretical power for the GRS statistic is determined by estimating the noncentrality parameter for a given γ (see Equation 12 in MacKinlay). The power is then calculated as the area of the noncentral F above the critical value of a central F, with probability of Type I error equal to 5 percent.

TABLE 1

Sample Parameter Estimates

Date	# of Individual Securities Used to Form 20 Portfolios	Market Mean Excess Return	Std. Dev. of Market Excess Return	Noncentrality Parameter	Power	Estimated Γ_G	p-Level
1/54–12/58	831	1.61%	3.51%	17.71	51.5%	1.99	0.033
		(1.6)	(3.5)	(23.2)	(67.0)	(2.11)	(0.023)
1/59–12/63	825	0.57	4.10	7.02	19.5	1.03	0.454
		(0.6)	(4.1)			(1.65)	(0.089)
1/64–12/68	739	1.43	3.96	6.60	18.4	1.30	0.236
		(1 5)	(4.0)	(8.54)	(24.0)	(0.75)	(0.750)
1/69–12/73	831	−0.85	6.00	9.74	27 2	1.03	0.452
		(−1.0)	(6.2)			(1.14)	(0.350)
1/74–12/78	929	1.01	6.86	2.86	9.8	1.61	0.101
		(0.6)	(6.7)	(4.95)	(14.0)	(0.68)	(0.820)
1/79–12/83	1,008	1.23	5.17	4.20	12.6	0.73	0.774
		(1.3)	(5.2)			(0.66)	(0.840)

Note: The descriptive statistics are based on 20 beta-sorted portfolios formed from all securities with complete data in each period. For purposes of comparison, values from MacKinlay's (1987) Table 1 and Table 6 are reported parenthetically where available. The market portfolio is measured using the CRSP equal weighted index. Power estimates for the noncentral F are based on a program by Narula and Weistroffer (1986). The program was benchmarked using values reported in Tiku (1965). The noncentrality parameter is calculated as specified in the definition of terms in Equation (3).

TABLE 2

Descriptive Results for the Γ_G Statistic

	Normal Errors				Bootstrap Raw Data	
	Null (theoretical)	Null (simulated)	Alternative (theoretical)	Alternative (simulated)	Null	Alternative
Mean	1.054	1.059	1.477	1.475	1 331	1.796
Std. Dev.	0 425	0.429	0.577	0.611	0.734	1.057
Skewness	1.156	1 180	1.127	1.105	2 185	3.377
Kurtosis	5.522	5.624	5.408	5.055	9.292	27.976

Note: The descriptive statistics are based on 20 beta-sorted portfolios formed from all securities with complete data in each period. Statistics are aggregates over the six sample periods. Simulation and bootstrap values are based on 2,000 and 1,000 replications respectively for the null and alternative hypothesis for each period. Thus, N = 12,000 for the null cases and N = 6,000 for the alternative cases. The theoretical values under the null are for an $F_{20,39}$ (see Johnson and Kotz (1977), p. 78). The theoretical values under the alternative are only approximate as they represent the average of the theoretical moments over the six noncentrality parameters corresponding to the six periods examined.

The descriptive statistics for the bootstrapped null distribution reveal a notable difference from the theoretical distribution. This could result either because the underlying stock data are not normally distributed or because of the bootstrap procedure. In either case, the effect is that use of standard F-tables to compute the critical value of the test could result in misspecification of the size of the test—that is, the actual probability of rejection, given the null hypothesis is true, may differ from the specified significance level of 0.05. The problem is overcome in this study by using the empirical bootstrap distribution under the null hypothesis to compute the critical value of the test, thereby forcing the size of the test to be equal to the specified significance level.

As previously discussed, the nature of our experiments dictates the use of a

subset of securities from the total sampling population.[15] We selected a sample size of 240 securities as a number that would be representative of the population, provide a sufficient number of securities for portfolio formation, and retain the computational feasibility of the study. Before proceeding to our comparative tests, it is important to consider the sensitivity of the tests to the sampling procedure. We examine this by performing ten experiments in the first sample period. Only the Γ_G statistic is considered and the seed for the random number generator that determines the initial sample subset is changed for each of ten experiments. The 240 securities selected in each experiment are allocated to 20 beta-sorted portfolios and the theoretical power of the test is determined. For the ten cases reported in Table 3, there is substantial variability, with the power ranging anywhere from approximately 20 to 63 percent, where the power using the full sample to form portfolios (as reported in Table 1) is 51 percent. To the extent that the basket of assets included in typical tests of asset pricing is not well-defined (and usually defaults to available data), these results suggest that the power of the test is not a relatively stable parameter. Also, since T is relatively small (60) the stochastic behavior of the estimated variance-covariance matrix almost certainly results in large variability from sample to sample. Shanken ((1985), p. 333) discusses this aspect in terms of his cross-sectional regression test and also points out that the expected value of the inverse is biased (Press (1970)). This problem is addressed in subsequent results by making comparisons over more than one test period and, for cases where only portfolios are being compared, using all securities with complete data in a given period.

TABLE 3

Sample Sensitivity of Γ_G

Test	Noncentrality Parameter	Power	Test	Noncentrality Parameter	Power
1	7.19	19.9%	6	17.08	49.6%
2	14.08	40.4	7	19.05	55.4
3	16.13	50.3	8	21.69	62.7
4	14.07	40.4	9	12.48	35.5
5	9.88	27.6	10	13.14	37.5

Note. The theoretical power is based on 20 beta-sorted portfolios formed from 240 randomly selected securities using only the first period (1/54–12/58). In each test, the seed for the random number generator used to select the sample was changed. The noncentrality parameter is calculated as specified in the definition of terms of Equation (3).

Given that the essence of the multivariate test is its ability to incorporate the correlations of the market model residuals, it is of interest to examine the magnitude of these correlations for the samples to be used in subsequent tests. Table 4 reports the mean correlation and mean absolute correlation for the samples of 240 securities and the 20 portfolios formed from these securities in each period. Reflecting the distribution of positive and negative correlations, the mean correlation is, in all cases, less than 0.04 in magnitude and slightly positive in all but the first test period. The mean absolute level of correlation is approximately 0.12

[15] Note that this limitation is attributable to the simulation expense. The sample magnitudes in this study do not computationally preclude the inversion of the covariance matrix for tests using individual securities.

(with a range of 0.01 across periods) for individual securities, and approximately 0.16 (with a range of 0.05) for portfolios. The observed range of correlations for the portfolio residuals was typically from about −0.50 to 0.75 across the six periods. As will be seen in subsequent results, the level of these values does not closely correlate with the corresponding power of the test statistics in each period.

TABLE 4

Residual Correlations Based on Individual Securities versus Portfolios

	Residual Correlations			
	$N = 240$		$NP = 20$	
Date	Mean Correlation	Mean Absolute Correlation	Mean Correlation	Mean Absolute Correlation
1/54–12/58	−0.000	0.117	−0.018	0.136
1/59–12/63	0.004	0.117	0.014	0.159
1/64–12/68	0 008	0.124	0.012	0.183
1/69–12/73	0.009	0.122	0.034	0.142
1/74–12/78	0.008	0.133	0.036	0.218
1/79–12/83	0.008	0.125	0.025	0.151

Note: The table reports the mean correlation and the mean absolute correlation across securities and portfolios for the market model residuals of Equation (1). For the individual securities, 240 are randomly selected from those with complete data. These 240 securities are then combined into 20 beta-sorted portfolios.

Before examining the comparative power of the statistics, we can gain some insight into the maximum entropy method by considering its performance relative to the GRS statistic in the case where the estimate of Σ is nonsingular (i.e., the portfolio case). Table 5 reports the rejection frequencies for the Γ_{ME} and Γ_G statistics in each period for the 20 beta-sorted portfolios formed from the same subset of 240 securities used in subsequent tests. The rejection frequencies are established by bootstrapping the null distribution to determine the critical value and then bootstrapping 1,000 cases of the alternative and tabulating the number of rejections. In addition, the table reports for each period the correlation between the 2,000 pairs of statistics calculated under the null and 1,000 pairs of statistics calculated under the alternative. Clearly, the ME statistic, under the condition where $N < T-2$, closely approximates the exact statistic. The rejection frequencies for Γ_{ME} versus Γ_G are virtually identical, and the correlation of the bootstrapped statistics is, in all cases, greater than 0.99.[16]

B. The Comparative Power of the Alternative Statistics

From the statistics developed in Section IV, there are a number of comparisons of interest vis-à-vis the Γ_G statistic. First, we consider the effect of using the ME statistic on a large sample of securities, where the ME statistic is, in essence, very similar to the standard GRS statistic. Second, the Γ_D statistic can be exam-

[16] It must be emphasized that the comparison is presented merely to demonstrate that the ME approach provides a "reasonable" alternative to the GRS test. It is not intended to argue that the ME approach should be used when $N < T-2$.

TABLE 5

A Comparison of Γ_{ME} versus Γ_G—Where $\hat{\Sigma}$ is Nonsingular

Date	Rejection Frequency Γ_{ME}	Rejection Frequency Γ_G	Correlation between Γ_{ME} and Γ_G	
			Null	Alternative
1/54–12/58	32.9%	32.0%	0.996	0.996
1/59–12/63	10.9	10 3	0.997	0.997
1/64–12/68	8.8	8.7	0.995	0.996
1/69–12/73	10.8	10.2	0.996	0.997
1/74–12/78	10.4	9.1	0.992	0.994
1/79–12/83	7.4	7.6	0.997	0.996

Note. The figures are based on the 20 portfolios formed from 240 randomly selected securities in each period. Rejection frequencies are the percent of cases where the null hypothesis that $\alpha_p = 0$ is rejected when the annualized constant of measurement error in the risk-free rate is equal to 10 percent and the probability of Type I error is equal to 5 percent Correlations are for the bootstrap replications of the statistic (2,000 for the null and 1,000 for the alternative).

ined to determine the effect of ignoring off-diagonal information and, since the statistic is not constrained with respect to N, we can also observe the effect of using a large number of individual securities versus a small number of corresponding portfolios. The basis of comparison in all cases is the theoretical power of the exact statistic applied to the 20 beta-sorted portfolios.

Table 6 reports the rejection frequencies for the various statistics in each period for the case where the annualized error in measuring the risk-free rate is 10 percent and the probability of Type I error is 5 percent. Bootstrap results are reported for the maximum entropy statistic (Γ_{ME}, column 1) applied to 240 individual securities, and the diagonal statistic (Γ_D) applied to 240 individual securities (column 2) and to 20 portfolios (column 3). As before, portfolios are created from the sample of 240 individual securities. The last column (Γ_{GT}, column 4) reports the theoretical power based on the standard GRS statistic applied to the same 20 portfolios.

The results for the Γ_{ME} statistic are somewhat ambiguous. In four of the six test periods, the statistic is roughly half as powerful as the portfolio-based exact statistic (relative power is reported parenthetically in the table). In the fourth test period, the two statistics are essentially equal in power, and in the third period the Γ_{ME} statistic actually dominates. Although the Γ_{ME} statistic is generally less powerful, its advantage is that it avoids the portfolio formation problem. On the other hand, it is more difficult to implement, as it requires bootstrapping. The importance of the former (avoiding portfolio formation) should not be minimized. In the simulations above, we have examined the effect of changes in γ. Since $\alpha = (1-\beta)\gamma$, it is reasonable to assume that portfolio grouping on β favors the GRS test; however, against other alternatives, the grouping should be done on the basis of the alternative being investigated. Grouping on the incorrect alternative could mask deviations from the theory and reduce the power of the test. Thus, in cases where the alternative hypothesis is unspecified, the results presented in Table 6 suggest that the ME approach using individual securities provides a reasonable alternative to the portfolio approach. Clearly, given knowledge of the appropriate grouping criterion, the GRS test is more powerful,

TABLE 6

A Comparison of the Power of Alternative Test Statistics for Individual Securities and Portfolios

	Individual Securities (N = 240)		Portfolios (NP = 20)	
Date	(1) Γ_{ME}	(2) Γ_D	(3) Γ_D	(4) Γ_{GT}
1/54–12/58	31.2% (0.67)	26.9% (0.58)	52.4% (1 12)	46.7%
1/59–12/63	12.1 (0.56)	17.1 (0.79)	27.5 (1.27)	21.7
1/64–12/68	15.8 (1.17)	24.3 (1.80)	32.1 (2.38)	13.5
1/69–12/73	15.1 (1.02)	14.9 (1.01)	32.4 (2.19)	14.8
1/74–12/78	7.8 (0.51)	10.7 (0.70)	22.9 (1.51)	15.2
1/79–12/83	10.4 (0.44)	10.3 (0.44)	20.5 (0.87)	23.5

Note: The table reports the percent of cases where the null hypothesis that $\alpha_p = 0$ is rejected when the annualized constant of measurement error is equal to 10 percent and the probability of Type I error is equal to 5 percent. For the individual securities, 240 are randomly selected from those with complete data. These 240 securities are then combined into 20 beta-sorted portfolios. Empirical power relative to corresponding theoretical power of the GRS test (Γ_{GT}, column 4) is reported parenthetically.

but in the absence of such knowledge the ME approach provides an alternative that does not require arbitrary grouping.

The results for the Γ_D statistic applied to individual securities are very similar to those of the Γ_{ME} statistic. It is a reasonable alternative, but does not dominate the exact statistic applied to portfolios when the appropriate grouping criterion is known.

The results in Table 6 also allow a comparison between the Γ_D and Γ_G statistics in the portfolio case (column 3 versus column 4). The GRS statistic has a higher theoretical rejection frequency only in the last period (23.5 percent versus 20.5 percent), and this difference is not statistically significant. The Γ_D statistic clearly dominates the Γ_G statistic in all remaining periods. In the third period, the diagonal test is 2.3 times more powerful than the Γ_G statistic. This result indicates that even when the appropriate grouping criterion is used, the simple diagonal statistic usually dominates the GRS statistic. This implies that the precision with which the additional parameters are estimated in the exact test is not sufficient to create a net increase in power. Given the results of Table 6, we will now focus on additional comparisons of the Γ_D and Γ_G statistics for the portfolio case.

An interesting question arising from the previous results is whether the bootstrap method applied in the Γ_D case has any notable impact on the resultant power of the tests (i.e., whether the reported results are simply an artifact of the modified bootstrap methodology). Since the estimated Σ is nonsingular in the portfolio case, a parametric simulation can be performed under the assumption of multivariate normality. Table 7 examines the effect of the bootstrap method on the Γ_G and Γ_D statistics vis-à-vis a comparison with a parametric simulation. For the Γ_G statistic, the bootstrap method is compared to a simulation that assumes normal errors for the cases where the critical value is the theoretical (tabled) value (column 2), and where the critical value is chosen from the empirical null

distribution (column 3). The theoretical power is reported in column 1 and differs from the corresponding information in column 4 of Table 6 since, in this case, all securities with complete data are included in the sample portfolios (versus randomly selecting 240 securities for each period). The first three columns serve primarily to validate the process, with the standard simulations conforming very closely to their theoretical expectation. The bootstrap method applied to the Γ_G statistic, however, appears to impose a substantial loss in power. This difference associated with the bootstrap results could be attributable to two potential sources. It is possible that the substantial differences are caused by the assumption of normality where it is not appropriate. To the extent that the underlying distribution is not normal, the theoretical power of Γ_G could be misstated due to misspecification of the size of the test. The bootstrap results are nonparametric and, thus, will not be affected. Consistent with the results of previous studies (e.g., Fama (1976) or Affleck-Graves and McDonald (1989)), we assume that the error distributions are reasonably normal and, therefore, it is unlikely that minor deviations would have such a substantial impact. It is more likely, after considering the results in Table 7 for Γ_D, that the loss of power is attributable to the bootstrap method (i.e., the loss of power in moving from a parametric to a nonparametric test in the finite sample case). Looking at the rejection frequencies for the Γ_D statistic for both the parametric and nonparametric cases in Table 7 (columns 5 and 6), it does not appear that there is any loss in power associated with the use of the bootstrap approach for the diagonal statistic.[17]

However, our previous conclusion on the appropriateness of Γ_D versus Γ_G is based on the comparisons of columns (1) and (6). Using simulated normal data, this is equivalent to a comparison of columns (1) and (5). As can be seen, the comparisons are almost identical—Γ_D dominates Γ_G, whether we use simulated data or bootstrapped data. This provides strong evidence that our results are not simply an artifact of the bootstrap method employed.

C. Additional Tests of the Sharpe-Lintner CAPM

In this section, the GRS and diagonal statistics are used to test the Sharpe-Lintner specification of the CAPM over the 1931–1985 period.[18] The 5-year subperiods that are tested allow some comparisons with Fama and MacBeth (1973). In each period, all securities with complete data are aggregated into 20 beta-sorted portfolios to test the efficiency of the CRSP equal weighted index. Table 8 presents the results of these tests. The levels of residual correlations reported in the table are similar to those previously observed, with the mean correlation consistently less than 0.10 and the mean absolute correlation ranging from 0.157 to 0.330. Thus, these results provide additional comparisons between the Γ_D and Γ_G statistics over a wide range of covariance structures present in market model

[17] Although not reported here, the effect of bootstrapping on the ME statistic was essentially identical to the reduction in power observed when bootstrapping the GRS statistic (based on portfolio tests). This suggests that the ME method applied to individual securities could be far more competitive in terms of power if the distribution of the statistic when $N \geq T-2$ was known.

[18] Strictly, only the mean-variance efficiency of the CRSP equal weighted index is tested. However, this is frequently referred to as a test of the CAPM.

TABLE 7

The Effect of Bootstrapping on Power Tests of the Γ_G and Γ_D Statistics

	Γ_G				Γ_D	
Test Period	(1) Theoretical Power	(2) Simulated Normal Errors (tabled value)	(3) Simulated Normal Errors (empirical dist.)	(4) Bootstrap Raw Data (empirical dist.)	(5) Simulated Normal Errors (empirical dist.)	(6) Bootstrap Raw Data (empirical dist.)
1	51.5%	50.0%	50.0%	27.6%	77.0%	80.3%
2	19.5	20.7	21.6	12.8	21.9	26.1
3	18.4	19.8	20.6	11.0	45.2	43.7
4	27.2	25.8	24.6	13.6	40.8	39.5
5	9.8	10.2	9.6	6.5	22.8	22.2
6	12.6	13.4	13.6	9.1	29.9	25.0

Note. Proportion of rejections where the annualized constant measurement error (γ) is equal to 10 percent and the probability of Type I error is equal to 5 percent. The data is based on 20 beta-sorted portfolios formed from all securities with complete data in each period. The Γ_G statistic is based on Equation (3). The Γ_D statistic assumes that the off-diagonal elements of the covariance matrix are all equal to zero and is estimated using Equation (5). The six test periods are 1/54–12/58, 1/59–12/63, 1/64–12/68, 1/69–12/73, 1/74–12/78, and 1/79–12/83. The source of the critical value for each test is reported in parentheses. The tabled value is an $F_{20,39}$ = 1.846. The 95th percentile from the simulated normal or bootstrapped null distributions is used in cases labeled "empirical dist."

residuals.[19] For the 11 subperiods, the power of the simple diagonal statistic unilaterally dominates the rejection frequency of the GRS statistic (although this difference is not statistically significant in the first period). The diagonal statistic rejects the Sharpe-Lintner CAPM at the 0.10 level in 6 of the 11 periods, the GRS statistic in only 3 of the 11 periods, with Fama-MacBeth, using a different method, rejecting the null in 2 of their 6 subperiods. The results, using the different approaches, are generally consistent—one exception being the 1/61–12/65 period where only the Γ_G statistic rejects the null.

The reported p-levels for the Γ_D statistic in Table 8 are based on bootstrap distributions. Again, to verify the efficacy of this technique, we also generated the p-levels using a parametric simulation and found the levels to be essentially identical.

Gibbons and Shanken (1986) document a substantial increase in power provided by aggregating tests from sample subperiods. Using their test to aggregate our results, the Sharpe-Lintner CAPM can be rejected at a p-level of 0.001 for the 1931–1985 period using either the Γ_G or Γ_D statistics (the test statistics are, respectively, 42.753 and 62.483, which are compared to a chi-square with 11 degrees of freedom). Using the Γ_D statistic, there is no contiguous 20-year interval in which the Sharpe-Lintner model cannot be rejected at the 0.05 level. Although these tests address selected issues of estimation and power, as tests of the CAPM they are still subject to the usual caveats concerning survival bias, selection of the market index, and stationarity assumptions.

[19] Once again, there is no apparent relationship between the reported correlations and the power of the tests. In addition, the complete residual correlation matrices were examined, and there were no obvious patterns that correspond to the theoretical power of the GRS test for a given period.

TABLE 8

Comparative Tests of the Sharpe-Lintner CAPM over the 1931–1985 Period

Date	# of Individual Securities Used to Form 20 Portfolios	ρ and $\lvert\rho\rvert$	Γ_G	Γ_G Power	Γ_D	Γ_D Power	Fama-MacBeth t-Statistic for H_0: $\gamma_0 = R_F$
1/31–12/35	322	0.093 |0.214|	1.122 (0.368)	6.6%	1.454 (0.445)	8.6%	—
1/36–12/40	540	−0.011 |0.330|	0.965 (0.519)	11.5	0.810 (0.829)	26.4	0.31 (0.756)
1/41–12/45	603	0.009 |0.274|	1.400 (0.180)	16.8	2.441 (0.069)*	35.8	1.22 (0.223)
1/46–12/50	728	−0.012 |0.201|	0.881 (0.610)	32.6	2.966 (0.048)**	50.5	1.10 (0.272)
1/51–12/55	781	−0.031 |0.157|	1.453 (0.156)	71.5	3.759 (0.002)***	80.3	4.56 (0.001)***
1/56–12/60	847	−0.034 |0.228|	3.032 (0.001)***	55.9	11.341 (0.001)***	71.5	4.89 (0.001)***
1/61–12/65	812	−0.003 |0.219|	2.163 (0.019)**	38.2	1.747 (0.275)	43.7	−0.80 (0.576)
1/66–12/70	763	0.031 |0.309|	0.626 (0.868)	12.8	0.566 (0.958)	23.8	—
1/71–12/75	878	0.065 |0.308|	0.767 (0.734)	11.7	1.269 (0.525)	17.6	—
1/76–12/80	979	0.072 |0.233|	1.928 (0.039)**	20.6	4.413 (0.005)***	36.8	—
1/81–12/85	951	0.016 |0.230|	1.159 (0.337)	20.9	3.871 (0.016)**	31.1	—

Note: The Γ_G and Γ_D statistics are estimated using 20 beta-sorted portfolios formed from all securities with complete data in each period. Power for the Γ_G statistic is based on the theoretical power. Power for the Γ_D statistic is based on bootstrap replications of the statistic for the null and alternative. In both cases, the power is measured for an annualized constant of measurement error in the risk-free rate equal to 10 percent and the probability of Type I error equal to 5 percent. The correlations reported in the third column are the mean correlation and mean absolute correlation between the market model residuals in each portfolio. The Fama-MacBeth results are taken from the Sharpe-Lintner tests reported in Panel A, Table 3 of Fama and MacBeth (1973). The final period in their paper is actually 1/61–6/68, so it is not strictly comparable to our tests of the 1/61–12/65 period. The asterisks (*, **, ***) denote significance at the 0.10, 0.05, and 0.01 levels, respectively. p-levels are reported parenthetically.

VI. Conclusions

In this study, we examine estimation issues associated with the GRS multivariate test of the efficiency of a given portfolio. Specifically, we apply alternative tests that allow a large number of independent assets to be examined and the contribution of off-diagonal information to the power of the test to be evaluated. In the first case, the use of individual securities in a maximum entropy statistic is shown to be feasible, thereby avoiding the necessity to form portfolios and the attendant grouping dilemma. Although, for the case examined in this paper, the ME statistic is not dominant over the standard GRS statistic applied to corresponding portfolios, the grouping procedure used to form the portfolios is likely to bias the results in favor of the GRS statistic. Thus, the ME statistic is likely to perform even better in cases where the appropriate grouping procedure is unknown. The major disadvantage of the ME statistic is that the distribution of the test statistic is unknown, which means that the bootstrap distribution must be used in hypothesis testing. This results in a loss of power and is computationally

expensive when N is large. However, the bootstrap method has the additional advantage of not requiring the multivariate normality assumption and of ensuring that the size of the test is fixed at the assumed significance level.

In the second case, a simple aggregate of univariate portfolio statistics is found to be more powerful than the GRS statistic. This suggests that the benefits of incorporating the interrelationship between the residuals is more than offset by the lack of precision with which the relationships can be estimated. This result was confirmed using a parametric simulation. As with the ME statistic, the disadvantage of the diagonal statistic is that any hypothesis testing must resort to some form of bootstrap or parametric simulation to characterize the distribution of the statistic under the null hypothesis. The advantage is that the method is generally more powerful than the GRS statistic in standard applications.

We test the null hypothesis of $\alpha_p = 0$ against an unspecified alternative using five years of monthly data. Clearly, similar tests would have to be conducted for different alternative hypotheses and different measurement intervals. The fundamental structure of the tests and the magnitude of the differences in power suggest that these results are not unique to the experimental design. In fact, Bernard (1987) finds that including cross-sectional information in event tests becomes less important as the measurement interval is shortened. This result implies that the dominance of the diagonal statistic could be greater using weekly or daily data.

It must be emphasized that the dominance of the diagonal statistic is important primarily from an empirical/estimation standpoint. Substantive insights still can be gained from the structure of the GRS statistic—e.g., the intuitive interpretations developed in GRS of the test, and extensions such as those of Shanken (1987) and Kandel and Stambaugh (1987), which use the GRS statistic as a conceptual framework to delineate the robustness of inferences concerning mean-variance efficiency.

Appendix: The Maximum Entropy Statistic

The maximum entropy statistic of Theil and Laitinen (1980) is applicable in situations where the covariance matrix is singular. Begin with a univariate case where the random variable X takes on values $x_1, \ldots, x_T$. If $f(\ )$ is the estimated density, the entropy is defined as

$$H = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx \,. \tag{A.1}$$

The observations are arranged in ascending order, denoted as $x^1, \ldots, x^T$, and assigned to intervals I_1 to I_T. The mass-preserving constraint implies that

$$\int_{I_t} f(x)\, dx = 1/T, \qquad t = 1, \ldots, T. \tag{A.2}$$

The mean-preserving constraint requires that the mean of each interval is a

homogeneous linear function of the associated x^t's, and that the overall mean is preserved. Thus, Equation (A.1) is maximized subject to Equation (A.2). The means of the intervals are then given by

$$
\text{(A.3)} \qquad
\begin{aligned}
E\left(X \mid X \in I_t\right) &= 0.75x^1 + 0.25x^2, && \text{if } t=1, \\
&= 0.25x^{t-1} + 0.5x^t + 0.25x^{t+1}, && \text{if } t=2,\ldots,T-1, \\
&= 0.25x^{t-1} + 0.75x^T, && \text{if } t=T.
\end{aligned}
$$

The variance is given by

$$
\text{(A.4)} \qquad \mathrm{Var}(X) = s^2 - (1/4T)\sum_{t=1}^{T-1}\left(x^{t+1} - x^t\right)^2 - (1/24T)\sum_{t=2}^{T-1}\left(x^{t+1} - x^{t-1}\right)^2,
$$

where s^2 is the mean square of $x^1 - x_m, \ldots, x^T - x_m$, with x_m = sample mean.

Extending the statistic to the multivariate case, Theil and Laitinen show that, given

$$
\text{(A.5)} \qquad x_k^* = E\left(X \mid X \in I_t\right), \qquad \text{if } x^t = x_k,
$$

and y_k^* similarly defined, then, for a simple bivariate case with the random variables X and Y,

$$
\text{(A.6)} \qquad \mathrm{Cov}(X,Y) = (1/T)\sum_k \left(x_k^* - x_m\right)\left(y_k^* - y_m\right).
$$

This statistical adaptation based on the entropy concept will always provide a nonsingular covariance matrix because the mass of each observation is now spread out over a cell that has N dimensions.

References

Affleck-Graves, J., and B. McDonald. "Nonnormalities and Tests of Asset Pricing Theories." *Journal of Finance,* 44 (Sept. 1989), 889–908.

Bernard, V. "Cross-sectional Dependence and Problems in Inference in Market-based Accounting Research." *Journal of Accounting Research,* 25 (Spring 1987), 1–48.

Bickel, P. J., and D. A. Freedman. "Some Asymptotic Theory for the Bootstrap." *Annals of Statistics,* 9 (Nov. 1981), 1196–1217.

Black, F. M.; M. Jensen, and M. Scholes. "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets,* M. Jensen, ed. New York, NY: Praeger (1972).

Fama, E. *Foundations of Finance,* New York, NY: Basic Books, Inc. (1976).

Fama, E., and J. MacBeth. "Risk, Return and Equilibrium: Empirical Tests." *Journal of Political Economy,* 81 (May/June 1973), 607–636.

Gibbons, M. "Multivariate Tests of Financial Models: A New Approach." *Journal of Financial Economics,* 10 (March 1982), 3–27.

Gibbons, M.; S. Ross; and J. Shanken. "A Test of the Efficiency of a Given Portfolio." *Econometrica,* 57 (Sept. 1989), 1121–1152.

Gibbons, M., and J. Shanken. "Subperiod Aggregation and the Power of Multivariate Tests of Portfolio Efficiency." *Journal of Financial Economics,* 19 (Dec. 1987), 389–394.

Kandel, S., and R. Stambaugh. "On Correlations and Inferences about Mean-variance Efficiency." *Journal of Financial Economics,* 18 (March 1987), 61–90.

Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions.* New York, NY: John Wiley & Sons, Inc. (1972).

———. *Distributions in Statistics: Continuous Univariate Distributions* – 2. Boston, MA: Houghton Mifflin Co. (1977).

MacKinlay, A. "On Multivariate Tests of the CAPM." *Journal of Financial Economics,* 18 (June 1987), 341–371.

Muirhead, R. *Aspects of Multivariate Statistical Theory.* New York, NY: John Wiley & Sons (1982).

Narula, S., and H. Weistroffer. "Computation of Probability and Non-centrality Parameter of a Non-central F-distribution." *Communications in Statistics—Simulation,* 15 (Sept. 1986), 871–878.

Roll, R. "A Critique of the Asset Pricing Theory's Tests—Part 1: On Past and Potential Testability of the Theory." *Journal of Financial Economics,* 4 (March 1977), 129–176.

Shanken, J. "Multivariate Tests of the Zero-beta CAPM." *Journal of Financial Economics,* 14 (Sept. 1985), 327–348.

———. "A Bayesian Approach to Testing Portfolio Efficiency." *Journal of Financial Economics,* 19 (Dec. 1987), 195–215.

———. "Multivariate Proxies and Asset Pricing Relations: Living with the Roll Critique." *Journal of Financial Economics,* 18 (March 1987), 91–110.

Siddiqui, M. "A Bivariate t Distribution." *Annals of Mathematical Statistics,* 38 (Feb. 1967), 162–166.

Stambaugh, R. "On the Exclusion of Assets from Tests of the Two-parameter Model: A Sensitivity Analysis." *Journal of Financial Economics,* 10 (Nov. 1982), 237–268.

Theil, H., and K. Laitinen. "Singular Moment Matrices in Applied Econometrics." In *Multivariate Analysis,* P. R. Krishnaiah, ed. Amsterdam: North-Holland Publ. (1980).

Tiku, M. "Laguerre Series Form of Non-central χ^2 and F Distributions." *Biometrika,* 52 (Dec. 1965), 415–427.