Ranking Portfolio Performance by a Joint Means and Variances Equality Test

by

Joel Owen*

and

Ramon Rabinovitch**

February 1998

* Professor of Statistics, Stern School of Business, New York University, 44 West Fourth Street, New York, N.Y. 10012. Tel. (212) 998 0446.
* EMAIL: JOWEN@STERN.NYU.EDU

** Professor of Finance, College of Business Administration, University of Houston, Houston, TX 77204-6282. Tel. (713) 743 4782. EMAIL: RAMON@UH.EDU

1. Introduction

In the last four decades, numerous authors have suggested methods to evaluate portfolio performance. Treynor (1965), Sharpe (1966), and Jensen (1968), proposed performance measures which produce a score for every portfolio being evaluated. These scores are used to compare the performance of any two portfolios or rank the performance of all portfolios in a given set. Later, Hendriksson and Merton (1981), Lehmann and Modest (1987), Moses, Chaney, and Veit (1987), Grinblatt and Titman (1989a), Okunev (1990) and others, developed new ways to evaluate portfolio performance. In their methods, management market timing and selection abilities received increasing attention. The general idea common to most of these works is that the method assigns to every portfolio a numerical score, often called excess return. This score is derived by taking the difference between the average portfolio return and the expected portfolio return predicted by some model. Usually, the score assigned to every portfolio depends on the risk-free rate and a benchmark (market) portfolio. For example, denote the Jensen score by J and use the CAPM, then using the prior population parameters, $J = \mu - R_f - \beta(\mu_m - R_f)$, where μ_m is the mean return of an efficient portfolio whose return is R_m, μ is the mean return of the portfolio being evaluated, whose return is R, $\beta = cov(R,R_m)/\sigma_m^2$, where cov is the covariance operator and σ_m^2 is the variance of R_m. J depends on the risk-free rate, R_f, and on the portfolio m, which needs to be an efficient portfolio. Roll (1977, 1978) and Green (1986) pointed out, however, that the concepts of efficiency and excess return need not be consistent. Moreover, Dybvig and Ross (1985a) showed that even if an efficient portfolio is used as a benchmark, both superior as well as inferior portfolios could produce positive J values, thus casting doubt on the usefulness of this approach. In fact, Dybvig and Ross (1985b) showed that an uninformed observer may calculate a positive or a negative J score even when evaluating a manager with superior information. They proved that if one uses this method of evaluation, a manager with superior information might appear to be performing in a suboptimal fashion. Grinblatt and Titman (1989a and 1989b) offered further criticisms when models other than the CAPM were the basis of the performance evaluation.

The criticism mentioned so far represents theoretical arguments based on population parameters. In practice, additional problems arise. For example, Lehmann and Modest (1987) conclude that the method of estimation used in the evaluation substantially changes the J score. Complicating this estimation problem is the uncertainty as to whether the benchmark portfolio being estimated is itself efficient. If it is not efficient, does it still matter, which benchmark portfolio is used? On the last point a study by Grindblatt and Titman (1991a) concludes that it does matter and Baily (1992) reaches the same conclusion. Although, Grindblatt and Titman (1989a) attempt to deal with some of these issues, no particular method of evaluation has found general acceptance. Moreover, while there is a growing body of literature on the estimation and testing the properties of efficient portfolios, see e.g., Kandel, McCulloch and Stambauch (1995) and the references therein, there is very little on the sampling errors of performance measures. In the case of the Jensen measure, other than tests to determine whether a particular score is statistically different from zero, no attempt has been made to deal with sampling errors when ranking portfolios. That is, two portfolios, each of whose score was deemed significant, were then ranked by these scores without any regard as to whether they are significantly different from one another or not. An exception is Jobson and Korkie (1981), who proposed statistical tests of equality of the Sharpe and (separately) the Treynor measures, based on their sample estimates.

In this paper, we develop a new portfolio performance evaluation method that addresses the issue of sampling errors directly, without the need to use an efficient portfolio or any other particular benchmark in order to determine the rankings.

The 1991 annual report of TIAA-CREF - one of the largest funds in the world - stated that the goal for CREF's stock fund was "...for CREF participants to earn stock market-type investment returns, but with less risk (fluctuations of returns) than the market." This stated goal for the fund indicated the wish to manage the fund such that it will dominate other investments in some risk-return sense. Thus, the first stage of the method we propose consists of a new statistical method that directly compares the performance of any two portfolios based on the mean-variance dominance (MV) criterion. Given a set of N portfolios, we directly compare every portfolio against each of the remaining N - 1 portfolios, thereby producing an NxN matrix containing the outcomes of all possible pairwise comparisons. In the second stage of the method we employ a ranking function which maps the elements of the comparison matrix into a complete ranking of the portfolio set. In general, our approach leads to a very different ranking of portfolios than those given by the Treynor, the Jensen and the Sharpe measures. We discuss, however, the conditions under which both approaches produce similar rankings.

The method we propose is different from other performance measures in several regards:
(i) The methods mentioned above assign an independent score to every portfolio and then, use these scores to compare and rank the portfolios. We take the reverse approach. First, our method produces direct pairwise comparisons among the portfolios and only then we rank them based on these comparisons. (ii) The basis for the other methods' ranking are scores that depend on a univariate measure concerning returns. That is, the rank is based on a conditional risk-adjusted average return, or a suitably standardized average return. We propose a ranking based on a bivariate criterion related to mean-variance dominance. (iii) These methods do not account for sampling errors. The method we propose is based on a statistical argument. Hence our procedure is implementable and justifiable even in a short time period and for fixed sample sizes. (iv) Our method is based on pairwise comparisons of all the portfolios thus, it is independent of a particular benchmark. As such, it does not allow for a manager to manipulate the rankings by gaming. (v) Our method does not depend on the existence of a risk-free rate. We show, however, how to appropriately adjust the procedure when a risk-free rate does exist.

We proceed as follows: In Section 2 we present a two-stage statistical procedure of ordering and ranking portfolio performance and illustrate this procedure with an example, using eight portfolios. In Section 3, we illustrate the use of the method by applying it to a set of 133 mutual funds. The first 130 funds in this set were first studied by Lehmann and Modest (1987) and again by Connor and Korajczyk (1990). To this set we added the S&P500 index and the equal and value weighted CRSP indexes. We produce our ranking of these 133 funds over three different time periods, 5, 10, and 15 years. In Section 4 we compare the ranking we obtain with the ranking calculated by the Treynor, the Sharpe, and the Jensen measures. In Section 5 we investigate several theoretical connections between our method and the Treynor, the Sharpe, and the Jensen measures. We offer some additional remarks in Section 6.

2. A Procedure of Ranking Portfolio Performance

In this section we develop a statistical method of ordering and ranking portfolios following the concept of Mean-Variance (MV) Dominance.

2.1 The Theoretical Foundation of the Test of Dominance

We follow the usual definition of MV dominance :

Definition 1. Let R_j denote the return on portfolio j with mean μ_j and variance σ_j^2, j = 1,...,N. For any two portfolios in the set, j and k, we say that portfolio j dominates portfolio k if $\mu_j \geq \mu_k$ and $\sigma_j^2 \leq \sigma_k^2$, with at least one inequality being strict. This type of dominance is denoted by R_jDR_k. If portfolio j is dominated by portfolio k we write $R_j \overline{D} R_k$. If the means are equal and the variances are equal too, we say that the two portfolios are equal and denote this by R_jER_k. If neither dominance nor equality determines the relationship between the two portfolios, we call the portfolios noncomparable and denote it by R_jNCR_k .//

The following three theorems establish a new method of determining MV dominance between any two portfolios. We begin by recalling a basic result in regression theory.

Lemma A. Let $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ represent a simple linear regression model, i = 1,...,N, with ε_i being independent and normally distributed with mean zero and variance σ^2. Then the simultaneous Null hypothesis H_o: $\beta_0 = \beta_1 = 0$ may be tested against the general alternative, using the uncorrected F statistic $UF = \frac{N-2}{2}[\Sigma(Y_i^2 - \Sigma(Y_i - \hat{Y}_i)^2] / \Sigma(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the least-squares fitted model. Under H_o, UF has a central F distribution with 2 and N - 2 degrees of freedom. Small values of UF do not allow H_o to be rejected.

Proof: See any text on testing the general linear hypothesis.

This result leads to a statistical method for testing the simultaneous equality of the mean returns and the return variances of any two portfolios:

Theorem A. Let R_{it} be the random return on portfolio i, i=1,...,N at time t=1,..., T. Assume that the vectors (R_{it}, R_{jt}) are a random sample from a bivariate normal distribution, and run the regression:

$$\mathbf{Y_t = \beta_0 + \beta_1 (X_t - \overline{X}) + \varepsilon_t} \qquad t = 1,...,T,$$

where, $Y_t = R_{jt} - R_{it}$, $X_t = R_{jt} + R_{it}$ and $\mathbf{\overline{X} = \frac{1}{T}\Sigma X_t}$; i, j = 1,...,N.

Then, the corresponding UF-statistic given in Lemma A, can be used to test the simultaneous hypothesis: H_o: $\mu_j = \mu_i$ and $\sigma_j^2 = \sigma_i^2$ versus H_a: $\mu_j \neq \mu_i$ or $\sigma_j^2 \neq \sigma_i^2$. This UF-statistic follows an $F_{2,T-2}$ distribution under H_o.

Proof: See Bradley and Blackwood (1989).

Two remarks are in place:

Remark 1: Notice that the UF-statistic is not the usual F statistic, in that it is testing whether both the intercept and slope of the regression line are simultaneously zeros. The essence of the proof of Theorem A involves the recognition that the least square estimators of the intercept and slope are estimating $EY_t = \mu_1 - \mu_2$, and

$$\mathbf{\frac{cov(Y,X)}{V(X)} = \frac{E[(R_j - \mu_j) - (R_i - \mu_i)][(R_j - \mu_j) + (R_i - \mu_i)]}{V(X)} = \frac{\sigma_j^2 - \sigma_i^2}{V(X)}}, \text{ respectively.}$$

Thus, the two coefficients will be zero if and only if $\mu_j = \mu_i$ and $\sigma_j^2 = \sigma_i^2$.

Remark 2: The regressions in Theorem A (and in Theorem B below,) do not in any way, constitute return generating models. These regressions are only used as a tool (a trick, if the reader wishes to call it this way) that enables us to obtain the UF-statistic needed for the test of the above simultaneous Null hypothesis. This is achieved by defining Y_t to be the difference between the returns on any two portfolios and X_t to be the sum of the returns on the same two portfolios. These definitions do not describe any economic model between the returns on the portfolios or between the dependent and independent variables and neither portfolio need be a benchmark portfolio.

The next theorem generalizes Theorem A to the case when the vectors (R_{jt}, R_{it}) are a sample from any elliptically contoured distribution.

Lemma B. Let $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ represent a simple linear regression model, i=1,...,N, with ε_i being jointly elliptically distributed with zero means. Then, the simultaneous null hypothesis H_o: $\beta_o = \beta_1 = 0$ may be tested against the general alternative using the UF-statistic of Lemma A. Furthermore, the distribution of UF under H_o is the same as under normality: $F_{2,N-2}$.

Proof: Theorem 6 in Anderson and Fang (1990) establishes that when the error term is elliptical, the likelihood ratio test criterion for testing H_o: $\beta_o = \beta_1 = 0$ is the same as under the normal case. Here this criterion is: $[EY_i^2 / E(Y_i - \star)^2]^{n/2}$ which is a monotonic function of UF of Lemma A.

We now generalize Theorem A to the case of elliptical distributions.

Theorem B. Let (R_{jt}, R_{it}), t = 1,...T, be a sample from a bivariate elliptical distribution. Then, the UF-statistic of Lemma A can be used to test the simultaneous hypothesis H_o: $\mu_j = \mu_i$ and $\sigma^2{}_j = \sigma^2{}_i$ against the general alternative. Under the Null hypothesis, the distribution of UF is $F_{2,T-2}$.

Proof: {R_{jt}, R_{it}} are bivariate elliptical, then (Y_t, X_t) in Theorem A are elliptical and Y_t is conditionally linear in X_t; Y_t may be written as $Y_t = \beta_0 + \beta_1 (X_t - \overline{X}) + \varepsilon_t$, where ε_t are elliptically distributed. See Owen and Rabinovitch (1983). Lemmas A and B then lead to the same conclusion as in Theorem A.

So far, theorems A and B establish a test for the simultaneous equality of the means and the variances of elliptically distributed bivariates. If the Null hypothesis cannot be rejected based on the UF-statistic, we conclude that $R_j E R_i$. Theorem C below, establishes the implications of the rejection of H_o using the statistical properties of the individual regression coefficients.

Theorem C. For i = 0, 1, let t_i be the t-statistic associated with the least squares estimator $\hat{\beta}_i$ that is used to test H_o: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$, respectively. Given that the UF-statistic is significant, the t_o-value may be used to test H_o: $\mu_j = \mu_i$ versus H_a: $\mu_j \neq \mu_i$ and the t_1-value may be used to test H_o: $\sigma_j^2 = \sigma_i^2$ versus H_a: $\sigma_j^2 \neq \sigma_i^2$.

Proof: Testing $\beta_o = 0$ is equivalent to testing $\mu_j = \mu_i$. Testing $\beta_1 = 0$ is equivalent to testing $\frac{\sigma_j^2 - \sigma_i^2}{V(X)} = 0$ or $\sigma_j^2 = \sigma_i^2$.

2.2 The Statistical Test of Dominance Without a Risk-Free Rate

We summarize the first stage of our two-stage procedure of ranking portfolio performance, using theorems A, B and C as follows: Let {R_{it}} be a random sample of N portfolio returns, i = 1,...,N during a sample period t, t = 1,...,T. Assume that R_i and R_j follow a bivariate elliptical distribution which is stationary over time. In order to run the regression in Theorem A in a systematic manner, fix j = 1 and run N - 1 regressions, where $Y_t = R_{1t} - R_{it}$ and $X_t = R_{1t} + R_{it}$ for portfolios i:i = 2,3,...,N. For each of these N - 1 regressions, test the simultaneous hypothesis of Theorem A. If the UF-statistic is not significant, report that $R_1 E R_i$. If, on the other hand, the UF-statistic is significant, reject H_o, i.e., conclude that the simultaneous equality of the means and variances does not hold. In the latter case, use the regression's t-values to test the separate hypotheses on the equality of the means and the equality of the variances. Table 1 specifies all the possible results of these tests, where t_k^* (t_k), k = 0,1, denotes a significant (insignificant) t-value at the desired level of significance. Note that if, following a significant F test, neither t_o nor t_1 is significant, we record the result as a noncomparable result. We show how to resolve the cases of noncomparable portfolios in Section 2.3 below. To continue, notice that the result of the N - 1 regression runs of R_1 on R_i, i = 2,3,...,N, and the above tests for each regression, is a vector of MV comparisons between portfolio j = 1 and the rest of the portfolios in the set. Every entry in this comparison vector is: R_1DR_i, or R_1ER_i, or $R_1 \hat{D} R_i$ or R_1NCR_i. For simplicity, we also run R_1 on itself thus, the first entry in the first comparison vector, (j=1, i=1), is R_1ER_1, trivially. The next step is to compare portfolio j = 2 to all the other portfolios. Thus, fix j = 2 and repeat the previous iteration of regression runs and tests of hypotheses for all i, i = 1,...,N. A second vector

Table 1

The Results of the Separate Tests on the Means and the Variances

		t_1 – test		
		$t_1^* > 0$ ($\sigma_j^2 > \sigma_i^2$)	$t_1^* < 0$ ($\sigma_j^2 < \sigma_i^2$)	t_1 ($\sigma_j^2 = \sigma_i^2$)
t_o - test	$t_o^* > 0$ ($\mu_j > \mu_i$)	R_jNCR_i	R_jDR_i	R_jDR_i
	$t_o^* < 0$ ($\mu_j < \mu_i$)	$R_i \hat{D} R_i$	R_jNCR_i	$R_i \hat{D} R_i$
	t_o ($\mu_j = \mu_i$)	$R_i \hat{D} R_i$	R_jDR_i	R_jNCR_i

will result, yielding information about the MV ordering of portfolio 2 relative to all the portfolios in the set. Again, running R_2 on itself yields R_2ER_2 trivially, in the second entry of this vector. Continue these iterative steps for j = 3, j = 4, and so on up to j = N. Every iteration produces another comparison vector so that after N iterations, we have created an NxN matrix of comparisons. To simplify the discussion, we transform the information in each entry of the comparison vectors into a numerical value using a comparison function:

Definition 2. For i,j = 1,...,N, a comparison function, COMP(i,j), is defined as follows:

$$COMP(i,j) = \begin{cases} 1 & \text{if } R_jDR_i, \\ 0 & \text{if } R_jER_i, \\ -1 & \text{if } R_j \hat{D} R_i, \\ 4 & \text{if } R_jNCR_i.// \end{cases}$$

Definition 3. A comparison matrix C is a matrix whose elements are COMP(i,j); j,i, = 1,...,N.//

From definitions 2 and 3, it follows that the elements of the matrix C are the real numbers 1,0,-1 and 4, depending on the type of dominance found in the hypothesis testing that followed the regressions of Theorem A.

$^{*}_{i}$

The first stage of the procedure could end at this point. Further refinement is possible, however, when a risk-free rate exists. In the next section we show how to employ the risk-free rate in order to resolve all the cases of noncomparable portfolios.

2.3 An Adjustment in Case That There Exists a Risk-Free Rate, R_f.

Suppose that a risk-free rate, R_f, exists in the economy. First, observe that subtracting R_f from all the portfolios' returns and performing stage one of the procedure with "access returns", leaves the above results unchanged because of the way Y_t, X_t and the regression of Y_t on X_t are defined. We use R_f, however, in order to resolve the cases of noncomparable portfolios. To do this, we extend the notion of dominance in the following way:

Definition 4. Let portfolio i and j be a pair of noncomparable portfolios. For portfolio i compute a new return, which is the linear combination of its original return and the risk-free rate:

$$R_i^* = (1 - \delta)R_f + \delta R_i \ .$$

Then, portfolio j is said to dominate (be dominated by) portfolio i if there exists a δ_{ij}, such that employing R_i^* and R_j leads to $R_j D R_i^*$ ($R_j \overline{D} R_i^*$) for that δ_{ij}. If portfolio j does not dominate or is not being dominated by portfolio i, we say that they are equal.//

Theorem D below establishes that all the cases of noncomparable portfolios are resolved under definition 4 of extended dominance. For simplicity we drop the subscripts ij in the sequel.

Theorem D. Let i and j be a pair of noncomparable portfolios and assume that a risk-free rate R_f exists. Then, there exists a δ with the corresponding R_i^* such that $R_j D R_i^*$, $R_j E R_i^*$, or $R_j \overline{D} R_i^*$

Proof: Let $\overline{R}_i$ be the sample mean return of portfolio i. If $\overline{R}_i - R_f \neq 0$, then, a possible choice is: $\delta = (\overline{R}_j - R_f) / (\overline{R}_i - R_f)$.

Adjust R_i using this δ and run the regression of Theorem B with Y_t and X_t redefined with R_j band R_i^* . It can be shown that $\hat{b}_0 = 0$ for this choice of δ, which implies that $\mu_i^* = \mu_j$ Thus, if the UF-statistic is not significant, it follows that $\sigma_i^* = \sigma_j$ as well, i.e, $R_j E R_i^*$. If, on the other hand, the UF-statistic is significant, it follows that the variances are different. This fact, coupled with the fact that the means are equal, implies that dominance must exist between the two portfolios. If $\mu_i = R_f$ but $\mu_j \neq R_f$, the same conclusion follows by reversing the roles of the portfolios. If $\overline{R}_i = \overline{R}_j = R_f$ then, again, $\hat{b}_0 = 0$ and the conclusion follows. This completes the proof.

For the complete procedure that we propose below, we first construct the matrix C without the use of R_f. We then introduce the risk-free rate and resolve all the cases of the noncomparable portfolios by setting δ equal to the value in Theorem D and comparing the resulting R_i^* to R_j. Note that we could have resolved the comparison between noncomparable portfolios by a complete search over δ values. Instead, we chose a simpler form of the algorithm. As Theorem D asserts, the result of this comparison resolves all the cases of noncomparabe portfolios. We use the notation COMP(i^*,j) to refer to the comparison of R_i^* with R_j, with δ chosen as in Theorem D. Therefore, for these cases we now have COMP(i^*,j) = -1, 0, or 1, where we previously had COMP(i,j) = 4. In this way the matrix C is transformed into a risk-free rate adjusted matrix that contains no noncomparable portfolios in it.

Definition 5. The risk-free rate adjusted comparison matrix C_f is the matrix with the elements $COMP_f(i,j)$: $COMP_f(i,j)$ = COMP(i,j), for all the cases in which COMP(i,j) = -1, 0, 1.

$COMP_f(i,j) = COMP(i^*,j)$, for all the cases in which $COMP(i,j) = 4$.//

This concludes stage one of our procedure in which we create comparison matrices containing the results of the pairwise comparisons of all the portfolios in the set. In stage two the procedure is completed by using a ranking function that maps the information in the comparison matrix in order to rank the portfolios, i.e., converting the C_f, (or the C) matrix into a complete ranking of the portfolios.

2.4 The Ranking Function

The j-th column (j = 1,...,N) of the risk-free rate adjusted comparison matrix C_f, contains the results of comparing the j-th portfolio against each and every portfolio in the set, expressed by $COMP_f(i,j)$. In a more general context, $COMP_f(i,j)$ might be some measured difference between portfolios j and i. In our analysis these results are expressed as real numbers $COMP_f(i,j)$ satisfying:

$$COMP_f(i,j) + COMP_f(j,i) = 0,$$

and $$COMP_f(i,j) = 1, 0, -1, \qquad i,j = 1,...,N.$$

We propose to rank the portfolios in a descending order of the row-sum as follows:

Definition 6. Let us use the matrix C_f in which the entries are: COMP(i,j) = 1, 0, or -1. A ranking function is defined by the row-sum function: $s_j = \Sigma_i^N COMP_f(i,j)$, j = 1,...,N,
and the portfolios are ranked in a descending order of s_j, breaking ties at random, or arbitrarily.//

By Definition 6, the "best" portfolio, ranked number 1, is the portfolio with the highest value of s_j, the "second best" portfolio, ranked number 2, is the portfolio with the second highest s_j score, and so on, down to the " worst" portfolio, ranked number N, which is the portfolio with the lowest value of s_j.

This ranking function follows from Huber's (1963) analysis of pairwise comparisons and ranking. Theorem 1 in Huber (1963) demonstrates that under some very general conditions, the ranking in descending order of s_j, breaking ties arbitrarily or at random, uniformly minimizes the risk among all (invariant) ranking procedures that depend on $COMP_f(i,j)$ only through s_j, for all “reasonable” risk functions. Moreover, Huber shows that minimum risk is achieved in cases in which the $COMP_f(i,j)$s are not necessarily independent random variables, as well in the case in which the $COMP_f(i,j)$s are independent random variables. (The form of the probability distributions of $COMP_f(i,j)$ and s_j and their properties are beyond the scope of this work and is currently under investigation.) Finally, we point out that the row-sum ranking function s_j applies to the comparison matrix C as well, with the slight modification of redefining COMP(i,j) = 4 to be zero, since the corresponding portfolios are noncomparable.

In conclusion, the two-stage ranking procedure described in the above sections may be summarized as follows: <u>in the first stage</u>, a statistical test of MV dominance determines the order of every portfolio with respect to all other portfolios. This information is presented in the form of the risk-free rate adjusted comparison matrix C_f (and/or C). <u>In the second stage</u>, the row-sum ranking function s_j is employed to rank the portfolios from the "best", ranked 1, to the "worst", ranked N. We refer to this procedure as the O-R procedure.

2.5 An Eight-Funds Portfolio Illustrative Example of the O-R Procedure

In Section 3 below, we apply the O-R procedure to a set of 133 portfolios. Much of the analysis and the tables of the 133 funds are too lengthy to be included in the paper. Therefore, we end this section with an example using a subset of eight portfolios. With this small subset of portfolios, the example exhibits the O-R procedure in its entirety thus, enabling the reader to verify every step of the procedure by inspection.

The data are the portfolios' monthly returns over a period of 15 years. The results for the eight portfolios chosen for the illustration are presented in tables 2 and 3. The upper part of Table 2 displays the UF-statistic and the statistics t_o, t_1 from the regressions described in Theorem A, using the level α=.005 for the F test and α=.01 for the t-tests. The critical UF and t values at these levels are 5.3 and 2.6, respectively. (Other reasonable values of significance levels produced similar results.) For example, the regression between portfolios 114 and 100, in which $Y_t = R_{114,t} - R_{100,t}$ and $X_t = R_{114,t} + R_{100,t}$, resulted in a significant UF = 6.1 > 5.3 and therefore, the rejection of H_0. Then, the t_1-value, t_1 = - 0.8 < 2.6, implied the equality of these portfolios' variances, but t_0 = 3.4 > 2.6 implied that $\mu_{114} > \mu_{110}$. Thus, $R_{114}DR_{100}$ and COMP(1,2) = 1 in Matrix C. The comparison matrix C is displayed in the lower-left part of the table. Similar explanation applies to all the 64 pairwise comparisons among the eight portfolios, including the eight trivial comparisons that lead to zeros along the main diagonal of the lower-left side of the table. The ranking function s_j and the O-R ranking based on the matrix C are shown in the lower-right part of the table. For the eight portfolios in the example, the rankings may be arrived at by inspection. It is seen that portfolio 114, (j = 2), is the "best" with s_2 = 6, portfolio 131 - the S&P500 Index with dividends, (j = 6), is the "second best" with s_6 = 2, breaking the tie between portfolio 131 and portfolio 120 (whose s_4 = 2, as well), arbitrarily. The ranking function continues in a descending order of s_j to rank portfolios 100, 130, 133 - the CRSP value-weighted index with dividends, 132 - the CRSP equal-weighted index with dividends and finally, portfolio 115, with s_3 = - 6, is ranked number 8, the "worst."

In Table 2 we find that portfolios 115 and 130 are noncomparable, COMP(5,3) = COMP(3,5) = 4, and so are portfolios 130 and 132, COMP(7,5) = COMP(5,7) = 4. To convert matrix C to matrix C_f, we employed the risk-free rate, which was taken to be the average monthly yield on T-bills over the sample period of 15 years. Table 3 displays the risk-free rate adjusted comparisons matrix C_f and the

Table 2

The O-R Procedure: An eight-funds example. Sample period: 1968 - 1982

Fund		100	114	115	120	130	131	132	133
100	t_0		3.4	1.6	0.8	-1.2	-0.1	1.2	0
	t_1		-0.8	11.5	0.7	-1	0.7	7.8	1.7
	F		6.1	66.8	0.6	1.2	0.3	31	1.4
114	t_0	-3.4		-0.4	-2.3	-3.9	-3.5	-1.3	-3.4
	t_1	0.8		13.8	1.6	-0.4	1.6	10.1	2.8
	F	6.1		95.6	3.9	7.7	7.6	51.5	9.8
115	t_0	-1.6	0.4		-1.4	-2.8	-2.5	-0.6	-2.5
	t_1	-11.5	-13.8		-14	-13.7	-15.5	-5.2	-15.1
	F	66.8	95.6		99.4	97.6	123.6	13.9	116.6

120	t_0	-0.8	2.3	1.4		-2.5	-1.5	1.1	-1.5
	t_1	-0.7	-1.6	14		-2.1	0	13.7	1.6
	F	0.6	3.9	99.4		5.3	1.1	94.2	2.3
130	t_0	1.2	3.9	2.8	2.5		1.6	2.7	1.7
	t_1	1	0.4	13.7	2.1		2.3	10.3	3.6
	F	1.2	7.7	97.6	5.3		4	56.6	7.9
131	t_0	0.1	3.5	2.5	1.5	-1.6		1.9	0.8
	t_1	-0.7	-1.6	15.5	0	-2.3		9.9	6.6
	F	0.3	7.6	123.6	1.1	4		51	21.8
132	t_0	-1.2	1.3	0.6	-1.1	-2.7	-1.9		-2
	t_1	-7.8	-10.1	5.2	-13.7	-10.3	-9.9		-9.8
	F	31	51.5	13.9	94.2	56.6	51		50.6
133	t_0	0	3.4	2.5	1.5	-1.7	-0.8	2	
	t_1	-1.7	-2.8	15.1	-1.8	-3.6	-6.6	9.8	
	F	1.4	9.8	116.6	2.3	7.9	21.8	50.6	

Fund	**100**	**114**	**115**	**120**	**130**	**131**	**132**	**133**	**Fund**	**s_j**	**O-R Ranking**
100	0	1	-1	0	0	0	-1	0	**114**	6	1
114	-1	0	-1	0	-1	-1	-1	-1	**131**	2	2
115	1	1	0	1	4	1	1	1	**120**	2	3
120	0	0	-1	0	0	0	-1	0	**100**	1	4
130	0	1	4	0	0	0	4	-1	**130**	0	5
131	0	1	-1	0	0	0	-1	-1	**133**	-2	6
132	1	1	-1	1	4	1	0	1	**132**	-4	7
133	0	1	-1	0	1	1	-1	0	**115**	-6	8
s_j	**1**	**6**	**-6**	**2**	**0**	**2**	**-4**	**-2**			

Table 3

An eight-funds portfolio example adjusted for the risk-free rate. Sample period: 1968 - 1982

Fund		100	114	115	120	130	131	132	133
100	to		3.4	1.6	0.8	-1.2	-0.1	1.2	0
	t1		-0.8	11.5	0.7	-1	0.7	7.8	1.7
	F		6.1	66.8	0.6	1.2	0.3	31	1.4

114	to	-3.4		-0.4	-2.3	-3.9	-3.5	-1.3	-3.4
	to	0.8		13.8	1.6	-0.4	1.6	10.1	2.8
	F	6.1		95.6	3.9	7.7	7.6	51.5	9.8
115	to	-1.6	0.4		-1.4	0	-2.5	-0.6	-2.5
	t1	-11.5	-13.8		-14	10.1	-15.5	-5.2	-15.1
	F	66.8	95.6		99.4	51.1	123.6	13.9	116.6
120	to	-0.8	2.3	1.4		-2.5	-1.5	1.1	-1.5
	t1	-0.7	-1.6	14		-2.1	0	13.7	1.6
	F	0.6	3.9	99.4		5.3	1.1	94.2	2.3
130	to	1.2	3.9	0	2.5		1.6	0	1.7
	t1	1	0.4	-9.6	2.1		2.3	-8.6	3.6
	F	1.2	7.7	45.9	5.3		4	37	7.9
131	to	0.1	3.5	2.5	1.5	-1.6		1.9	0.8
	t1	-0.7	-1.6	15.5	0	-2.3		9.9	6.6
	F	0.3	7.6	123.6	1.1	4		51	21.8
132	to	-1.2	1.3	0.6	-1.1	0	-1.9		-2
	t1	-7.8	-10.1	5.2	-13.7	9.3	-9.9		-9.8
	F	31	51.5	13.9	94.2	43.3	51		50.6
133	to	0	3.4	2.5	1.5	-1.7	-0.8	2	
	t1	-1.7	-2.8	15.1	-1.8	-3.6	-6.6	9.8	
	F	1.4	9.8	116.6	2.3	7.9	21.8	50.6	

Fund	**100**	**115**	**130**	**120**	**131**	**131**	**132**	**133**	Fund	s_j	O-R Ranking
100	0	-1	0	0	0	0	-1	0	**114**	6	1
114	-1	-1	-1	0	-1	-1	-1	-1	**131**	2	2
115	1	0	-1	1	1	1	1	1	**120**	2	3
120	0	-1	0	0	0	0	-1	0	**100**	1	4
130	0	1	0	0	0	0	1	-1	**133**	-1	5
131	0	-1	0	0	0	0	-1	-1	**130**	-2	6
132	1	-1	-1	1	1	1	0	1	**132**	-3	7
133	0	-1	1	0	1	1	-1	0	**115**	-5	8
s_j	**1**	**6**	**-5**	**2**	**-2**	**2**	**-3**	**-1**			

O-R ranking after the adjustment. The two cases of noncomparable portfolios are resolved using the δ from Theorem D. We found that portfolio 130 was dominated by both portfolio 115 and portfolio 132. The ranking itself was almost, but not exactly, the same as before, as portfolios 130 and 133 exchanged their former ranking.

3. A Study of 130 Mutual Funds

In this section we construct the C_f matrix and apply our ranking procedure to a set of 133 portfolios. Except for three portfolios, this is the same data set analyzed by Lehmann and Modest (1987) and Connor and Korajczyk (1990). The data set consists of monthly returns of the funds R_1 through R_{130} with dividends reinvested, over a fifteen year period from January 1968 through December 1982. To this data set we added three more funds with returns:

R_{131} = The Standard & Poors 500 Index (SP500 Index) with dividends,

R_{132} = The CRSP equal - weighted Index with dividends, and

R_{133} = The CRSP value - weighted Index with dividends.

We applied the O-R ranking procedure to the full data set and generated the UF, t_o and t_1 values and the 133x133 comparison matrix C. The critical F value, F=5.3, corresponds to α=.005. This α value was chosen this small to help control for the over-all error rate (a Bonferroni adjustment). The t-tests were performed at α=.01 with critical t-value of 2.6. We then, estimated the risk-free rate and transformed C into C_f, resolving all the cases of noncomparable portfolios. For the risk-free rate we used the average monthly yield on T-bills during the sample period. We performed separate analyses for three overlapping sample periods of 5 years: 1982 - 1978, 10 years: 1982 - 1973 and 15 years: 1982 - 1968. Because of space limitations we cannot present the results of the F and the t tests as we did for

Table 4
The O-R ranking of 133 funds for the period of 1968-1982

Rank	15yrs	10yrs	5yrs	Rank	15yrs	10yrs	5yrs	Rank	15yrs	10yrs	5yrs
1	73	73	13	46	120	40	83	91	16	112	25
2	105	105	73	47	6	29	74	92	66	93	126
3	51	92	112	48	63	5	37	93	46	90	65
4	92	51	4	49	91	107	109	94	59	113	7
5	114	68	116	50	41	100	90	95	69	52	84
6	71	116	115	51	111	129	47	96	113	46	41
7	116	88	99	52	104	48	5	97	75	11	81
8	125	125	88	53	107	43	117	98	86	108	36
9	88	71	63	54	2	25	129	99	36	69	8
10	44	62	51	55	7	3	24	100	109	9	80
11	62	4	59	56	77	20	127	101	110	80	38
12	23	128	44	57	38	**131**	100	102	65	94	124
13	15	115	125	58	25	6	29	103	52	16	31
14	128	18	18	59	20	14	32	104	11	66	67
15	18	15	105	60	14	22	120	105	112	121	97
16	106	13	128	61	22	77	6	106	119	8	61
17	4	44	62	62	85	17	**132**	107	81	119	56
18	99	87	106	63	40	45	14	108	42	31	69
19	89	99	71	64	96	37	19	109	30	65	11
20	39	106	12	65	45	117	76	110	70	110	86
21	87	33	113	66	53	85	**131**	111	28	78	121
22	60	23	130	67	72	111	43	112	97	81	16
23	21	39	104	68	10	35	46	113	101	49	119
24	74	12	33	69	**133**	123	20	114	78	82	75
25	68	89	92	70	123	38	17	115	**132**	97	108
26	33	60	23	71	50	19	52	116	8	61	93
27	13	64	111	72	35	98	35	117	49	75	101
28	64	83	57	73	57	**133**	123	118	31	86	110
29	12	74	21	74	3	79	98	119	82	28	1
30	103	57	91	75	126	2	122	120	94	30	28
31	83	118	26	76	67	7	2	121	90	26	27
32	130	21	87	77	124	96	107	122	9	42	94
33	29	114	89	78	19	1	79	123	102	109	70
34	118	120	60	79	17	10	10	124	26	95	95
35	76	76	45	80	93	72	22	125	34	102	58
36	5	63	9	81	11	124	49	126	95	101	30
37	127	41	68	82	79	67	48	127	6	70	42
38	48	127	118	83	1	50	103	128	54	34	34
39	84	32	77	84	98	53	114	129	121	58	102
40	**131**	103	39	85	37	56	85	130	27	27	78
41	32	130	40	86	24	59	53	131	58	122	82

42	43	104	15	87	115	24	96	132	55	54	54
43	129	91	64	88	80	36	66	133	122	55	55
44	47	47	50	89	108	**132**	**133**				
45	100	84	3	90	56	126	72				

the eight fund example. Nor can we present the entire 133x133 C_f matrix. Therefore, we only present the O-R rankings of the 133 funds. The results of the full ranking over the 15, 10 and 5 year periods appear in Table 4.

It is interesting to note that the three indexes we added to the original data set did not fare too well. For example, for the 15 year period between 1968 and 1982 the SP500 Index (fund number 131) is ranked 40. The CRSP equal-weighted index (fund number 132) is ranked 115 while the CRSP value-weighted index (fund number 133) is ranked 69.

As mentioned above, the full 133x133 C and C_f matrices were too large to include in this paper. In order to illustrate the information contained in the C_f matrix, however, Table 5 presents the tests results that lead to the creation of one vector in this comparison matrix namely, the column corresponding to the SP500 Index. We chose to present this particular column because the SP500 Index is commonly employed as the benchmark portfolio in empirical analyses done in the financial literature as well as by practitioners in fund management evaluations. Table 5 is divided up into three parts. All the parts display the fund number and the number of observations in the regression of Theorem A, with the SP500 Index as portfolio j (j = 131) and the indicated fund as portfolio i, i.e, $Y_t = R_{131,t} - R_{i,t}$ and $X_t = R_{131,t} + R_{i,t}$, i = 1,...,133. The UF, t_0 and t_1 values are also given. The left part contains all the funds that were dominated by the SP500 Index during the 15 years sample period and hence, the 131st column of the matrix C_f will have the entries $COMP_f(i,131) = 1$ for the corresponding row of fund i. The central part contains all the funds with UF-statistic values less than the critical value of 5.3. These are the funds that were equal to the SP500 Index and therefore, $COMP_f(i,131) = 0$ for these funds. The right part of the table shows the portfolios that dominated the SP500 Index during these 15 years, implying that $COMP_f(i,131) =- 1$. Table 5 shows that during the fifteen years between

Table 5

Comparing 132 Portfolios Against The SP500 Index During The Period 1968 - 1982

The SP500I Dominates Fund i					The SP500I Equals Fund i			Fund i Dominates The SP500I				
Fund	**NOB**	**F**	**t_0**	**t_1**	**Fund**	**NOB**	**F**	**Fund**	**NOB**	**F**	**t_0**	**t_1**
37	180	5.33	-0.07	3.26	7	180	0.03	128	180	78.04	-0.91	-12.46
79	180	5.51	1.01	3.16	47	180	0.04	23	180	77.26	-0.84	-12.4
24	180	6.61	-2.97	2.1	48	180	0.12	73	180	76.91	-12.72	-12.4
41	180	7.12	2.24	3.04	100	180	0.26	105	180	74.82	0.31	-12.23
98	180	8.3	-0.68	4.02	32	180	0.31	71	180	58.87	0.06	-10.85
93	180	8.96	1.59	3.92	107	156	0.38	51	180	57.66	-0.33	-10.73
117	180	9.81	-0.54	4.4	14	180	0.39	62	180	43.01	-0.01	-9.27
80	180	11.71	1.35	4.65	6	180	0.52	125	180	42.13	0.49	-9.17
56	180	12	0.22	4.89	63	180	0.72	21	180	35.79	-0.25	-8.46
69	178	12.03	-1.3	4.73	25	180	0.76	44	180	35.63	-0.77	-8.41
36	180	12.69	-0.67	4.99	20	180	0.84	88	180	32.95	1.37	-8
81	180	15.27	-0.36	5.52	91	180	0.9	92	180	26	-0.48	-7.2
52	180	15.28	-1.27	5.38	45	180	0.96	12	180	24.48	-1.56	-6.82
42	180	15.52	-1.18	5.45	76	180	-1.02	4	180	22.46	-0.63	-6.67
8	180	17.07	0.85	5.78	22	180	1.05	116	180	22.01	0.57	-6.61
46	180	17.35	-1.18	5.77	120	180	1.14	106	180	20.4	1.42	-6.23
16	180	17.71	-1.16	5.84	2	178	1.15	15	180	14.87	0.27	-5.45
66	180	19.9	-0.47	6.29	29	180	1.52	99	180	14.79	-1.17	-5.31
97	161	20.67	-0.91	6.36	83	180	1.56	89	180	14.07	0.49	-5.28
133	180	21.82	0.76	6.56	35	180	1.65	60	180	14.03	-0.33	-5.29
75	180	21.87	-2.21	6.23	77	180	1.69	18	180	13.22	1.76	-4.83
11	180	22.56	-1.48	6.55	72	156	1.77	114	180	7.56	3.53	-1.64
65	180	23.38	-0.66	6.81	126	180	1.77					
113	180	23.61	1.11	6.78	43	180	1.84					
112	180	24.43	1.68	6.78	53	180	1.88					
90	180	24.74	-0.21	7.03	3	144	1.89					
31	180	24.95	0.06	7.06	127	180	1.94					
86	180	25.02	-1.62	6.88	85	174	1.97					
101	176	25.5	-0.45	7.13	5	180	2.06					
94	180	26.71	-1.8	7.08	40	180	2.16					
110	180	28.32	0.53	7.51	118	180	2.2					
78	180	28.81	-0.62	7.57	111	180	2.23					
82	180	28.99	0.69	7.58	96	180	2.35					
26	177	29.91	1.01	7.67	103	180	2.57					
119	180	32.46	-1.51	7.92	84	180	2.65					
49	180	34.71	-0.96	8.28	50	180	2.67					
9	150	35.71	-2.32	8.13	123	153	2.81					
108	180	40.03	1.96	8.73	19	180	2.91					
95	177	41.19	-3.12	8.52	10	180	2.98					
28	144	41.96	-3.23	8.57	68	180	3.11					
54	180	42.63	0.21	9.23	38	180	3.14					

30	180	43.16	-1.6	9.15	67	180	3.59
59	180	43.22	2.16	9.04	17	180	3.81
61	180	44.24	-4.21	8.41	130	180	3.95
109	180	46.27	2.2	9.37	33	180	4.15
58	166	50.91	1	10.04	1	180	4.2
132	180	51.03	1.92	9.92	13	180	4.42
102	180	51.48	0.44	10.14	64	180	4.5
70	180	51.72	0.47	10.16	129	180	4.51
34	180	53.13	-0.55	10.29	87	180	4.62
121	180	54.25	1.06	10.36	57	177	4.76
27	180	70.68	-0.3	11.89	39	180	5.01
122	180	110.59	0	14.87	74	180	5.06
115	180	123.64	2.51	15.52	124	180	5.23
55	180	124.92	-0.04	15.81	104	180	5.79

1968 and 1982, the SP500 Index dominated 55 of the funds, was equal to 55 of the funds, and was dominated by 22 funds. Notice that out of these 22 funds, fund number 114 showed a statistically higher mean return than the mean return of the SP500 Index. The other 21 funds, dominated the SP500 Index based on lower variances. More importantly, perhaps, the table shows that no fund was noncomparable to the SP500 Index during the sample period. This is somewhat puzzling. If the SP500 index was an efficient portfolio during the sample period, we would expect it to dominate or be noncomparable to all other funds. The results in Table 5 cast some doubt as to the MV efficiency of SP500 Index. We found that the same was true of the two CRSP indices. Shanken (1985), Green (1986) and others reported similar results.

We can now illustrate the difference between ordering with respect to one benchmark portfolio such as the SP500 Index and our method of ordering based on all the pairwise comparisons. Consider two funds that "look like" the SP500 Index. For example, funds 10 and 13, are statistically equivalent to the SP500 Index (see the mid-section of Table 5). Using one benchmark, both portfolios 10 and 13 would rank the same. Table 4, however, reveals that using the O-R Method the SP500 index is ranked 40, while fund 13 has rank 27 - better than the SP500 Index, and fund 10 has rank 68 - worse than the SP500 Index. This example provides evidence to the claim that ranking relative to an inefficient fund is not consistent with ranking based on dominance. This result mirrors a result of Dybvig and Ross (1985a, p.388), although they used population parameters for their demonstration.

4. An Empirical Comparison of The Rankings Robustness

Given a sample of portfolio returns, R_{it}, i = 1,...,N, t = 1,...,T, the returns on a proxy for the market portfolio, R_{mt} and the risk-free rate R_f, the Treynor, T_i, the Sharpe, S_i, and the Jensen, J_i, portfolio performance measures for portfolio i are:

$$T_i = \frac{\bar{R}_i - R_f}{b_i} \; ; \qquad S_i = \frac{\bar{R}_i - R_f}{\hat{\sigma}_i^2} \; ; \qquad J_i = \bar{R}_i - R_f - b_i \left(\bar{R}_m - R_f \right),$$

where bars denote the sample averages, $\hat{\sigma}_i^2$ and b_i denote sample estimates of the i-th portfolio's return variance and the portfolio beta, respectively. We estimated these measures with the data of the 133 funds, using (separately) the three market indexes as proxies for R_m and the average yield on one month T-bills for the risk-free rate. We then ranked the portfolios using each one of these measures for each of the periods and for each of the market indexes. In order to investigate the stability of these rankings over time, we again, divided the 15 year period into 3 overlapping periods: 5 years (1978-1982), 10 years (1973-1982) and 15 years (1968-1982). Each of these periods was treated separately and our ranking and those of the other three methods were computed. The rank correlations were then computed both within periods and between periods for all methods and for different levels of significance. Again, the three different benchmark portfolios, the SP500 index, and the two CRSP indexes were used in the computations of T_i, S_i, and J_i. The results were very similar so Table 6 exhibits the rank correlations obtained with the SP500 Index as the proxy for R_m. The rank correlations among the four methods within the same time period appear in 4x4 portions along the main diagonal of the table. These portions are for 15 years at the upper-top left, for 10 years in the middle and for 5 years at the lower-bottom right of the table. The off-main diagonal portions of the table show the between-time periods rank correlations. The correlation value is accompanied with its p-value below it in order to enable the reader to assess the validity of the correlation. The table reveals that within each time period the rankings of the Treynor, the Sharpe and the Jensen measures were virtually the same, with rank correlation values ranging from

.92 to .99. There seems to be little or no similarity between their rankings and the O-R rankings with rank correlation values between .06 and .31. Furthermore, Table 6

Table 6

Rank correlations Among the Portfolio Performance Measures*

	15 Years				**10 Years**				**5 Years**			
	Sharpe	**Jensen**	**Treynor**	**O-R**	**Sharpe**	**Jensen**	**Treynor**	**O-R**	**Sharpe**	**Jensen**	**Treynor**	**O-R**
15 Years												
Sharpe	1.0000 0.0000											
Jensen	0.9861 0.0001	1.0000 0.0000										
Treynor	0.9971 0.0001	0.9869 0.0001	1.0000 0.0000									
O-R	0.3130 0.0002	0.2961 0.0005	0.3047 0.0004	1.0000 0.0000								
10 Years												

Sharpe	0.8710	0.8498	0.8641	0.1667	1.0000						
	0.0001	0.0001	0.0001	0.0551	0.0000						
Jensen	0.8495	0.8483	0.8448	0.1140	0.9845	1.0000					
	0.0001	0.0001	0.0001	0.1915	0.0001	0.0000					
Treynor	0.8677	0.8434	0.8603	0.1564	0.9980	0.9810	1.0000				
	0.0001	0.0001	0.0001	0.0722	0.0001	0.0001	0.0000				
O-R	0.2099	0.3017	0.3052	0.9416	0.2078	0.1615	0.1975	1.0000			
	0.0003	0.0004	0.0004	0.0001	0.0164	0.0632	0.0227	0.0000			

<u>5 Years</u>

Sharpe	0.4692	0.4517	0.4569	-0.2530	0.6545	0.6617	0.6586	0.1581	1.0000		
	0.0001	0.0001	0.0001	0.0033	0.0001	0.0001	0.0001	0.0691	0.0000		
Jensen	0.4802	0.4932	0.4765	0.3495	0.6532	0.6946	0.6558	-0.2444	0.9241	1.0000	
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0046	0.0001	0.0000	
Treynor	0.4932	0.4830	0.4901	0.0000	0.6634	0.6751	0.6681	-0.1467	0.9759	0.9452	1.0000
	0.0	0.000	0.000	0.0003	0.0001	0.0001	0.0001	0.0921	0.0001	0.0001	0.0000

	001	1	1									
O-R	0.2052	0.2112	0.2058	0.7114	0.1309	0.1075	0.1236	0.8011	-0.0643	-0.1649	-0.0621	1.0000
	0.0178	0.0147	0.0175	0.0001	0.1332	0.2182	0.1564	0.0001	0.4620	0.0587	0.4773	0.0000

* The SP500 Index is the benchmark portfolio for the Sharpe, the Jensen and the Treynor Measures.

The p-value appears below the rank correlation.

shows that the between-periods rank correlations of the Treynor, the Sharpe and the Jensen measures drop relatively quickly in spite of the periods being overlapping. These values drop from .99 within each period to around .85, between 15 and 10 years, .66, between 10 and 5 years, and .47, between 15 and 5 years. For the same time periods, the O-R ranking has rank correlation values of .9416, between 15 and 10 years, .8011, between 10 and 5 years, and .7114, between 15 and 5 years. These results indicate that the O-R rankings show greater stability over time than the ranking of the other three methods.

In the next section we provide a theoretical explanation of this greater level of stability over time of the O-R ranking, as well as an explanation for the stark dissimilarity between our ranking and the others.

5. Some Theoretical Connections Among the Ranking Procedures

We saw that while the Treynor, the Sharpe and the Jensen measures ranked the funds virtually the same, our method had statistically no (or negative) correlation with their rankings. In this section we argue that these results are to be expected. We focus first on the similarity of the Sharpe, the Treynor and the Jensen measures.

Let $\overline{R}_i$, $\hat{\sigma}_i^2$, b_i, $\overline{R}_m$ and V_m represent the sample values over the period under discussion, of the average return of fund i, its variance, its regression slope against the reference portfolio m, the average return of the reference portfolio and its variance. T_i, J_i and S_i denote the i-th fund score of Treynor, Jensen and Sharpe when using sample estimates.

Definition 7. A portfolio whose return variance is completely explained by the return variance of a reference portfolio m, is said to be a well diversified portfolio (WDP) with respect to the reference portfolio m. Notationally, a portfolio is a WDP if $\hat{\sigma}^2 = b^2 V_m$.//

We now show that for WDPs, the rankings of the three methods are the same:

Theorem E. Let R_f exist and $\overline{R}_m > R_f$. Let R_i and R_j be WDP's. Unless portfolios i and j are noncomparable, the ranking of these portfolios will be the same using the Sharpe, the Treynor or the Jensen scores.

Proof: For WDPs the Sharpe score is proportional to the Treynor score and any inequality involving the variances is equivalent to one involving the slopes. So we need only consider Treynor's measure. To complete the proof we show that the ranking of Treynor and Jensen are the same when the portfolios are not noncomparable. The case $\overline{R}_i = \overline{R}_j$ and $b_i = b_j$ is obvious since $T_i - T_j = 0 = J_i - J_j$. Next, consider the case $\overline{R}_i \geq \overline{R}_j$ and $b_i \leq b_j$ with at least one strict inequality. By computation, $T_i > T_j$. Since $J_i - J_j = \overline{R}_i - \overline{R}_j - (b_i - b_j)(\overline{R}_m - R_f)$, at least one term on the right is positive and none is negative. Therefore, $J_i > J_j$. Lastly, suppose that $b_i \geq b_j$ and $\overline{R}_i \leq \overline{R}_j$, with at least one strict inequality. Since this is exactly the same situation as the previous one with j replacing i, the same conclusion follows.

The next theorem shows that the ranks corresponding to the Jensen and the Treynor scores are the same even for some of the noncomparable portfolios and therefore, all three rankings are the same.

Lemma F. $T_i - T_j = J_i/b_i - J_j/b_j$

Proof: Using the definition of T_i:

$$T_i = \frac{\overline{R}_i - R_f}{b_i} = \frac{\overline{R}_i - R_f - b_i(\overline{R}_m - R_f)}{b_i} + (\overline{R}_m - R_f) = \frac{J_i}{b_i} + (\overline{R}_m - R_f).$$

It then follows that $T_i - T_j = J_i/b_i - J_j/b_j$

Theorem F. If $\overline{R}_i < \overline{R}_j$ and $0 < b_i < b_j$ and if in addition $T_i - T_j < 0$, then $J_i - J_j < 0$.

Proof: Since $T_i < T_j$ if follows from lemma F that $J_i/b_i - J_j/b_j < 0$ or that $J_i < \frac{b_i}{b_j} J_j$.
Since $b_i < b_j$ the result follows.

Theorem G. If $\overline{R}_i > \overline{R}_j$ and $b_i > b_j > 0$, and if in addition $T_i - T_j > 0$, then $J_i - J_j > 0$.

Proof: As in the last proof $T_i - T_j > 0$ implies that $J_i > \frac{b_i}{b_j} J_j$ and the result follows.

If we assume that mutual funds are WDPs then, theorems E, F and G imply that the Sharpe, the Treynor and the Jensen performance measures should yield the same rankings except for some noncomparable funds. In the previous section, Table 5 demonstrated that no fund was noncomparable to the SP500 Index in the sample period. Thus, the explanation for the high correlation among these rankings reflects the small number of noncomparable cases present in our sample. Of course, when the risk-free rate is assumed to exist, the O-R procedure employs the matrix C_f and resolves all the cases of noncomparable portfolios. The low correlations with our ranking also reflects the large number of funds that were found by our method to be equal to the SP500 Index. This last comment underscores the main statistical problem with these methods. They treat the sample estimates as if they were populations parameters and therefore, they rank few or no portfolios as equal. Theorem H below shows that if the O-R method also measured dominance using these estimates as if they were population parameters, all four methods would yield the same ranking. Clearly, we must assume the existence of a risk-free rate in order to compare our method to the other three measures . Thus, as in Theorem D above, comparing R_j against R_i should lead to the same conclusion as comparing R_j against $R_i^* = \delta R_i + (1-\delta) R_f$ with $\delta = (\overline{R}_j - R_f)/(\overline{R}_i - R_f)$. Then, all the parameters and the parameters estimates with a super * correspond to R_i^* .

Theorem H. Let a risk-free asset be available such that each investor may borrow or lend at the risk-free rate, R_f and assume that $\overline{R}_m > R_f$. Then,

a) $S_i > S_j$ iff there exists R_i^* : $\overline{R}_i^* \geq \overline{R}_j$ and $\hat{\sigma}_i^* \leq \hat{\sigma}_j$ with at least one strict inequality.

b) $T_i > T_j$ iff there exists R_i^* : $\overline{R}_i^* \geq \overline{R}_j$ and $b_i^* \leq b_j$ with at least one strict inequality.

c) $J_i > J_j$ iff there exists R_i^* : $\overline{R}_i^* \geq \overline{R}_j$ and $b_i^* \leq b_j$ with at least one strict inequality.

Proof: a) Let $S_i > S_j$ and choose δ such that $\overline{R}_i^* = \delta \overline{R}_i + (1-\delta)\overline{R}_f = \overline{R}_j$. Then, $S_i > S_j$ implies that $\delta\hat{\sigma}_i < \hat{\sigma}_j$. But $\delta\hat{\sigma}_i$ is the standard deviation of R. The converse is straightforward.

b) The proof for part b is the same as in a) with b replacing $\hat{\sigma}$.

c) Let $J_i > J_j$. Then, $\overline{R}_i - R_f - b_i(\overline{R}_m - R_f) > \overline{R}_j - R_f - b_j(\overline{R}_m - R_f)$. Choosing δ as in part a) and substituting in the last inequality yields $\delta b_i < b_j$. But δb_i is the sample regression slope of R_i^* . The converse follows by substitution.

Suppose the mean returns, the returns variances and covariances are estimated. And suppose that instead of using our statistical tests, the MV dominance criterion is employed using the sample estimates to decide on MV dominance. Then, Theorem H demonstrates that when a risk-free rate exists, the MV dominance criterion and the Treynor, the Sharpe and the Jensen performance measures yield the same rankings.

This result sheds some light on the reasons for the lack of correlation between these ranking and the O-R ranking. First, the O-R method does not take the sample statistics at face value, but rather uses them to test the hypothesis of dominance based on the regression of Theorem A. Therefore, this method leads to many ties, i.e., portfolios that are statistically too close to warrant ranking them differently. All such portfolios are deemed to be statistically equal to one another by our method. For example, Table 5 revealed that 55 out of the 132 funds in our study were judged to be equal to the SP500 Index. Analyzing the same data and using the sample statistics as if they were the true parameters, the Treynor, the Sharpe and the Jensen measures gave different rankings to these 55 funds relative to the SP500 Index. Secondly, Sharpe's measure ranks funds based on statistics computed from each fund separately, while the Treynor and the Jensen measures rank funds based on statistics computed for each fund that are corrected by a third fund, namely, the reference portfolio. We already saw, however, that two funds, 10 and 13, which appeared to be equal when compared only to one fund, the SP500 index, received very different ranking when judged relative to all the funds in the set. This example highlights the fact that the O-R procedure may deem portfolios equal to one another in the MV sense, yet rank them differently. This is so, because the O-R ranking function is defined over a much larger information set. This set contains not only the MV dominance relationship between some benchmark portfolio and the rest of the portfolios, but the MV relationship of every portfolio relative to all other portfolios in the set. In fact, the more funds under consideration the less one would expect that ranking relative to one fund would sustain itself when comparisons are done relative to all funds.

6. Discussion and Conclusions

The purpose of our paper is to present a method to evaluate the performance of portfolios when the underlying parameters of the distribution are unknown. To accomplish this we introduced a sequence of hypothesis tests concerning portfolios' returns parameters and used the results of these tests to create an ordering of performance. There are several advantages of this method of ordering. First, each test, and therefore the collection of tests, speaks to and about the true parameter values of the portfolios. Suppose, for example, a test that correctly concluded that the true parameter values of portfolios i and j were in the relationship $\mu_i > \mu_j$ and $\sigma^2_i < \sigma^2_j$. Then, Theorem F implies that the ranking using any of the three other methods, using population parameters, would place fund i over fund j. Conversely, if the sample values were used as if they were true in the Sharpe, the Treynor or the Jensen scores and ranked fund i over j, then the ranking based on the data would not imply the same rankings as if population values had been used. If the population values equated the funds, the ranking based on the data could be just the result of noise. A second advantage lies in the fact that the O-R method uses all the NxN (minus the N trivial) pairwise comparisons, rendering gaming possibilities virtually impossible even when the identity of a benchmark portfolio is known to the manager. Evaluating a fund's performance against one reference fund leads to easier gaming possibilities. Third, the ordinal nature of our procedure of ordering, may itself be advantageous. As mentioned at the outset, the numbers and cardinality of ranks implied by some of the other methods have come under severe criticism by Roll (1977, 1978) and Dybvig and Ross (1985a, 1985b). The ordinality of our method survives some of the criticism. For example, the theorems of the last section connecting the various ranking alternatives are true no matter which reference portfolio is used or whether the reference portfolio is efficient or not. This result should be compared to the discussion by Dybvig and Ross (1985a) on a conjecture of Roll; see Section III of this reference. Fourth, the only assumption needed for the O-R method is that the returns on portfolios are bivariate elliptically distributed with stationary distributions. We do not require as others do, an efficient benchmark portfolio and a linear model specifying the relationship of each portfolio to this benchmark portfolio. We emphasize, however, that should such models exist and an efficient portfolio be used, then it would be detected by our method because such a portfolio could never be dominated. Furthermore, in equilibrium, all other efficient portfolios would be noncomparable to the benchmark portfolio. In this case, all the portfolios would have no portfolios dominating them, and they all would tie for rank number one. Fifth, the O-R method allows investors to evaluate portfolios over relatively short time periods. To elaborate this point, we might ask what is implied about the management of the funds if a fund ranks higher than another by our method? The data set here does not contain the information that the funds had the same managers over the sample period and there may have been different managers within the period for some funds. So the test cannot refer to a particular manager's performance. If a particular manager at a particular period could time the market, it is not known whether that is temporary or permanent. If it is temporary, the process would be nonstationary and so the test need not necessarily refer to timing. Similarly, since we do not know if at some specific time some manager had specific information about a specific stock, the test cannot refer to this possibility. Yet, the evaluation of a superior manager seems to be important to investors. This means that investors may wish to be able to rank portfolios as frequently as every year or every six months, or even every quarter. Our method is implementable even in the latter case because with daily returns,e.g., the O-R procedure has enough observations for the regressions and the hypothesis tests of Theorem A every quarter. The robustness of the O-R procedure to the use of daily vis-a-vis weekly or monthly data is currently under investigation. Notice that the O-R procedure may be used to test the objective of the CREF fund quoted at the outset. Here, all managers of that fund presumably shared the goal of seeking a portfolio which dominated a particular reference fund. Our method permits an evaluation of this claim. On the other hand, were we to know that a particular manager was in place over the observed period, then the test results could be credited to that manager. Or if a fund claims to have the ability to consistently ferret out specific information about specific stocks and to be able to use this information to its advantage, these tests could be used to test this claim.

The literature has addressed two types of information that a manager might use to the advantage of the fund: timing information on the reference fund and specific information of the mean of a particular fund (the expected value of the error term of a linear model being unequal to zero). Of course, information could be of a more complicated form involving the distributions of the reference portfolio or of the error. The position we have taken here is that the information, if it is available, must lead to an improved mean-variance position of that fund relative to funds that do not have this information. Since information could be available to more than one fund, it is only in the multiple comparisons of performance that a fund can be evaluated. Our procedure is constructed to be consistent with this viewpoint.

Finally, it is important to realize that our ranking procedure is not intended to be used as a portfolio selection method. Nor is it a guidance only for investors who will invest their entire capital in one portfolio vis-a-vis investors who will invest in some combination of several portfolios. At the end of the investment horizon, both types of investors will need a performance measure. The O-R procedure is designed to rank the ex post performance of any combination of portfolios.

REFERENCES

Anderson, T.W. and Fang, K.T.,(1990), "Inference in Multivariate Elliptically Contoured Distributions based on Maximum Likelihood,"*Statistical Inference in Elliptically Contoured and Related Distributions*, edited by Fang, K.T and Anderson, T.W, Allerton Press.

Bailey, V.J., (1992), "Evaluating Benchmark Quality," Financial Analysts Journal, 48, 33-39.

Bradley, E.L. and Blackwood,.L.G., (1989), "Comparing Paired Data: A Simultaneous Test for Means and Variances," *The American Statistician*, 43, 234-235.

Blattberg, R.C. and Gonedes, N.J., (1974), "A Comparison of the Stable and Student Distributions As Statistical Models of Stock Prices," *The Journal of Business*, 47, 244-280.

Colson, G. and Zeleny, M., (1980), "Uncertain Prospects Ranking and Portfolio Analysis Under Condition of Partial Information," in *Mathematical Systems in Economics 44*.

Connor, G. and Korajczyk, R.A., (1990), "The Attributes, Behavior and Performance of US Mutual Funds," *Working Paper No. 30R, Northwestern University*. Ross, S.A., (1986a),

Dybvig, P.H. and Ross, S.A., (1986a) "Differential Information and Performance Measurement Using Security Market Line," *The Journal of Finance*, 40, 388-399.

Dybvig, P.H. and Ross, S.A., (1985b), "The Analytics of Performance Measurement Using a Security Market Line," *The Journal of Finance*, 40, 401-416.

Green, R.C. (1986), "Benchmark Portfolio Inefficiency and Deviations from the Security Market Line," *The Journal of Finance*, 61, 295-312.

Grinblatt, M. and Titman, S., (1989a), "Portfolio Performance Evaluation: Old Issues and new Insights," *The Review of Financial Studies*, 2, 393-421.

Grinblat, M. and Titman, S., (1989b), "Mutual Funds Performance: An Analysis of Quarterly, Portfolio Holdings," *The Journal of Business*, 62, 393-416.

Grinblatt, M. and Titman, S., (1991a), "Do Benchmarks Matter? Do Measures Matter? A Study of Monthly Mutual Returns," *working paper, UCLA.*

Grinblatt, M. and Titman, S., (1991b), "The Persistence of Mutual Fund Performance," *working paper UCLA.*

Grinblatt, M. and Titman, S., (1993), "Performance Measurement Without Benchmarks: An Examination of Mutual Fund Returns," *The Journal of Business*, 66, 47-68.

Grinblatt, M. and Titman, S., (1994), "A Study of Monthly Mutual Fund Returns and Performance Evaluation Techniques," *Journal of Financial and Quantitative Analysis*, 29, 419-444.

Hendriksson, R. (1984), "Market Timing and Mutual Fund Performance: An Empirical Investigation," *The Journal of Business*, 57, 73-96.

Hendrikssom, R. and Merton, R., (1981), "On Market Timing and Investment Performance, II. Statistical Procedures for Evaluating Forecasting Skills," *The Journal of Business*, 54, 513-533.

Huber, J. P., (1963), "Pairwise Comparison and Ranking: Optimum Properties of the Row Sum Procedure," *The Annals of Mathematical Statistics*, 34, 511-520.

Jagannathan, R. and Korajczyk, R.A,. (1986), "Assessing the Market Timing Performance of Managed Portfolios," *The Journal of Business*, 59, 217-235.

Jensen, M., (1968), "The Performance of Mutual Funds in the Period 1945-64," *The Journal of Finance*, 23, 389-416.

Jobson, D.J. and Korkie, B.M., (1981), "Performance Hypothesis Testing With the Sharpe and Treynor Measures," *The Journal of Finance*, 36, 889-908.

Kandel, S.,McCulloch, R. and Stambauch, R.F., (1995), "Bayesian Inference and Portfolio Efficiency" *The review of Financial Studies*, 8, 1-53.

Lehmann, B.N. and Modest, D.M., (1987), "Mutual Fund Performance: A Comparison of Benchmarks and Benchmark Comparison," *The Journal of Finance*, 42, 233-265.
Merton, R., (1981), "On Market Timing and Investment Performance, I. An Equilibrium Theory of Value for Market Forecasts," *The Journal of Business*, 54, 363-406.
Moses, E.A., Cheyney, J.M. and Veit, E.T., (1987), "A New and More Complete Performance Measure," *The Journal of Portfolio Management*, 24-33.
Okunev, J., (1990), "An Alternative measure of Mutual Fund Performance," *Journal of Business Finance and Accounting*, 17, 247--264.
Owen, J. and Rabinovitch, R., (1983), "On the Class of Elliptical Distributions and their Applications to the Theory of Portfolio Choice," *The Journal of Finance*, 38, 745-752.
Owen, J. and Rabinovitch, R., (1988), “Ranking Portfolio Performance: An Application of A Joint Means and Variances Equality Test and Risk-Minimizing Pairwise Comparisons,” *The University of Houston CBA Working Paper*.
Rennie, E.P. and Cowhey, T.J., (1990), "The Successful Use of Benchmark Portfolios: A Case Study," *Financial Analysts Journal*, 46, 18-26.
Roll, R., (1977), "A Critique of the Capital Asset Pricing Theory Test," *Journal of Financial Economics*, 4, 129-176.
Roll, R., (1978), "Ambiguity When Performance is Measured by the Security Market Line," *The Journal of Finance*, 33, 1051-1069.
Sharpe, W., (1966), "Mutual Fund Performance," *The Journal of Business*, 39, 119-138.
Treynor, J.L., (1965), "How to Rate Management of Investment Funds," *Harvard Business Review*, 43, 63-75.
Treynor, J.L. and Mazuy, K.K., (1966), "Can Mutual Funds Outguess the Market?" *Harvard Business Review*, 44, 131-136.

Footnotes

The second-named author acknowledges support of the Dean of the CBA at the University of Houston research grant. We are grateful for the criticism we have received from the participants of the finance seminars at Rice University, The University of Houston and New York University. In particular, we thank, for their comments, suggestions and support, Yakov Amihud, Amir Barnea, Sarabjeet Seth, Ron Singer and Oldrich Vasicek. Special thanks are extended to Robert Korajczeyk for providing us with the data set. Errors remain ours.

ABSTRACT

We propose a new procedure to rank portfolio performance. Given a set of N portfolios, we use statistical tests of dominance which produce direct mean-variance comparisons between any two portfolios in the set. These tests yield an NxN matrix of pairwise comparisons. A ranking function maps the elements of the comparison matrix into a numerical ranking. To illustrate the procedure we use a set of 133 mutual funds, including the S&P500 index and the CRSP equal and value weighted indexes. We explore the empirical and theoretical relationships between our ranking procedure and the Treynor, Sharpe and Jensen performance measures. In general, the new procedure's ranking is relatively robust, does not allow for gaming and can be performed with small samples.