Generalized Tests of Investment Fund Performance*

Márcio Poletti Laurini**

Rogério da Costa Monteiro***

Antônio Zoratto Sanvicente****

Abstract

The paper discusses the use of statistical methods in the comparison of investment fund performance indicators. The analysis is based on the robust statistics proposed by Ledoit and Wolf (2008), involving a pairwise comparison of funds and two generalizations for sets of multiple investment funds. The multiple investment fund tests use the Wald and Distance Metric statistics, based on an estimation by the Generalized Method of Moments (GMM) using HAC matrices. The test distributions are obtained through block-bootstrap procedures, in order to correct for the size limitations of the GMM estimation in the case of a large number of moment conditions. For the period from July 2006 to July 2008, we applied the proposed procedures to daily return data for the largest 97 actively managed equity funds in the Brazilian market. The results indicate that there are no significant differences in the performances of the 97 funds in the sample, both in pairwise and joint comparisons, thus providing what is believed to be the first Brazilian market evidence for the so-called herding hypothesis.

Keywords: Sharpe Ratio, GMM, Investment Analysis.

JEL Codes: G11, G14.

*Submitted in December 2012. Revised in April 2013.

**Departament of Economics – FEA-RP USP and Associated Researcher at CNPq. Av. Bandeirantes 3900, 14040-905. Ribeirão Preto – SP, Brazil. Phone 55 16 3329 0867. E-mail: mplaurini@gmail.com

***Insper Institute of Education and Research.

****Insper Institute of Education and Research.

1. Introduction

The evaluation of investment fund performance is an important issue in finance, since both investor returns and manager compensation from those funds is based on the fund's performance. Normally, the fund's performance is assessed with reference to possible benchmarks such as market indices, or through a comparison to the performance of other funds in the same asset class. Fund performance evaluation is done by comparing of performance measures, such as the Sharpe ratio. The Sharpe ratio is defined as the ratio between average return and the standard deviation of returns during the period under analysis. Usually, performance evaluation conducted through the use of point estimates only. However, point estimates may be problematic, since true parameter values are not observed and must be estimated on the basis of historical data. As noted by Ledoit and Wolf (2008): "Since the true quantities are not observable, the Sharpe ratios have to be estimated from historical return data and the comparison has to be based on statistical inference, such as hypothesis tests or confidence intervals." Therefore, this requires the use of statistical tools in order to compare fund performances, such as the use of tests of hypotheses. In the present paper, we discuss the use of testing hypotheses in the relative performance comparison of two funds (pairwise comparisons), and we also present a generalization for joint performance comparison in a set of many funds.

The finance literature has already characterized, a so-called herding behavior on the part of managed portfolios, which includes investment funds. This type of behavior is defined as the tendency, on the part of funds, to buy and sell the same stocks at the same time. This behavior was studied by Grinblatt et al. (1995), who claim to have found relatively weak evidence for this type of tendency.

Their evidence was obtained by using a statistic proposed by Lakonishok et al. (1992), involving differences in the individual stock proportions of the portfolios of various stock mutual funds. While Grinblatt et al. (1995) provide information on which differences are significant or not, the authors do not discuss the assumptions under which they make use of t statistics for reaching their conclusions.

The present paper, not only proposes a more robust statistical methodology, but also provides the first kind of evidence for fund herding behaviour in the Brazilian market that we know of, without using portfolio proportions. Instead, our analysis directly employs the traditional Sharpe ratio measure of fund performance.

2. Hypothesis Testing Using the Sharpe Ratio

In order to compare the performances of two funds, the most traditional test is based on Jobson and Korkie (1981). This test assumes that returns are IID processes from a bivariate normal distribution. Considering that the mean vector

is represented by $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and the variance-covariance matrix for funds i and n is given by $\Sigma = \begin{pmatrix} \sigma_i^2 & \sigma_{ik} \\ \sigma_{ik} & \sigma_k^2 \end{pmatrix}$, the difference between the Sharpe ratios of two funds is given by $\Delta = \frac{\mu_i}{\sigma_i} - \frac{\mu_k}{\sigma_k}$, and the estimator of that difference is $\hat{\Delta} = \frac{\hat{\mu}_i}{\hat{\sigma}_i} - \frac{\hat{\mu}_k}{\hat{\sigma}_k}$, with $\hat{\mu}_j$ and $\hat{\sigma}_j$ as sample estimators for the mean and the variance of a given fund's returns.

In order to perform the hypothesis test, it is crucial to obtain the standard deviation of the estimator $\hat{\Delta}$. This estimator is obtained through the Delta method using asymptotic normality, thus being assumed that $\sqrt{T}(\hat{\mu} - \mu) \xrightarrow{d} N(0, V)$, where the V matrix is given as:

$$V = \begin{pmatrix} \sigma_i^2 & \sigma_{ik} & 0 & 0 \\ \sigma_{ik} & \sigma_k^2 & 0 & 0 \\ 0 & 0 & 2\sigma_i^4 & 2\sigma_{ik}^2 \\ 0 & 0 & 2\sigma_{ik}^2 & 2\sigma_k^4 \end{pmatrix} \quad (1)$$

Please note that this formulation is valid only if the returns are independent Gaussian processes. Thus, violations of that assumption, such as fat tails, nonlinearities and serial correlation/dependency in returns, affect the performance of tests using this estimator, as discussed in Ledoit and Wolf (2008) and Kosowski et al. (2006).

It is possible to modify the previous estimation of matrix V to generate a robust version of this test by using a long-run variance matrix that is robust in relation to autocorrelation and heteroscedasticity, such as an HAC matrix (Andrews, 1991). We would then have $\sqrt{T}(\hat{\mu} - \mu) \xrightarrow{d} N(0, V_{HAC})$. This HAC matrix is constructed by Ledoit and Wolf (2008) as:

$$\hat{V}_{HAC} = \frac{T}{T-4} \sum_{j=-T+1}^{T-1} \kappa_{j,S_t} \left(\frac{j}{S_t} \right) \hat{\Gamma}(j) \quad (2)$$

where:

$$\hat{\Gamma}(j) = \begin{cases} \frac{1}{T} \sum_{t=j+1}^T \hat{y}_t \hat{y}_{t-j}' & \text{if } j \geq 0 \\ \frac{1}{T} \sum_{t=-j+1}^T \hat{y}_{t+j} \hat{y}_t' & \text{if } j < 0 \end{cases} \quad (3)$$

and $\hat{y}_{t-j}' = (r_{it} - \hat{\mu}_1, r_{kt} - \hat{\mu}_k, (r_{it} - \hat{\mu}_i) - \sigma_i^2, (r_t - \hat{\mu}_k) - \sigma_k^2)$.

In this function, κ_{j,S_t} represents the chosen kernel function (Bartlett, Quadratic Spectral, etc.), and S_t is the bandwidth chosen for this function (e.g. Andrews, 1991, Anatolyev and Gospodinov, 2011).

Although the use of these estimators corrected for heteroscedasticity and autocorrelation provides only a partial solution to the problems in the tests of performance measure comparisons, their finite sample performance may not be satisfactory. This affects the size and power of the tests of hypotheses using the corrected

matrices, as was discussed in Ledoit and Wolf (2008). This problem results from the finite sample properties of HAC estimators (e.g. Anatolyev and Gospodinov, 2011). To get around this problem, Ledoit and Wolf (2008) propose the use of a studentized bootstrap method (e.g. Efron, 1982, Efron and Tibshirani, 1993) to correct the distributions obtained with HAC estimators in finite samples. In this type of procedure, the test for the equality of the performances of two funds is obtained with the inversion of the confidence intervals generated by the bootstrap procedure. In this case, the test is defined by the standardization of the difference between the performance measures:

$$\mathcal{L}\left(\frac{\hat{\Delta} - \Delta}{S(\hat{\Delta})}\right) = \mathcal{L}^*\left(\frac{\hat{\Delta}^* - \Delta}{S(\hat{\Delta}^*)}\right) \quad (4)$$

with $\hat{\Delta}^*$ and $S(\hat{\Delta}^*)$ denoting the estimator of Δ obtained using the bootstrap procedure, $S(\hat{\Delta}^*)$ indicating the standard deviation estimator obtained from the bootstrap method, and $\mathcal{L}^*$ being the empirical distribution obtained by the bootstrap. The $1-\alpha$ confidence interval for Δ is obtained as $\hat{\Delta} \pm z_{[\cdot], 1-\alpha}^* S(\hat{\Delta}^*)$, with $z_{[\cdot], \lambda}^*$ denoting the λ quantile of the bootstrap distribution of $\mathcal{L}^*\left(\hat{\Delta}^* - \Delta/S(\hat{\Delta}^*)\right)$.

An important worth highlighting is the performance of the bootstrap procedure. In order to deal with the temporal dependency problems in the series of returns, Ledoit and Wolf (2008) use a circular block-bootstrap procedure (Politis and Romano, 1992). In such a procedure, samples are taken in blocks of consecutive observations, facilitating the reproduction of the temporal dependency characteristics of the sample. With a detailed simulation analysis, Ledoit and Wolf (2008) show that the block-bootstrap procedure generates good properties for both IID data and data with various forms of temporal dependency, with superior performance *vis-à-vis* those of the Jobson and Korkie (1981) procedures and the correction using only HAC matrices (Andrews, 1991).

A fundamental issue is how to compare, in a robust way, the joint performance of sets of investment funds. A simple alternative would seem to be the aggregation of the results obtained in the Ledoit and Wolf (2008) test, through a simultaneous inference procedure, for example, using the p -values estimated in each individual comparison. This class of procedures is known as a set of multiple comparisons (e.g. Lehmann and Romano, 2005). The problem, however, when performing multiple comparisons with the use of a fixed significance level α , it is difficult to control for the so-called family-wise error rate (FWER), given by the increasing number of false rejections in multiple comparison procedures. The usual correction procedures, like the Bonferroni and Holm method (Lehmann and Romano, 2005), are applicable only to independent tests. This is not the case for the hypothesis of fund performance equality, since, for example, we may accept that $\frac{\mu_1}{\sigma_1} = \frac{\mu_2}{\sigma_2}$ and $\frac{\mu_2}{\sigma_2} = \frac{\mu_3}{\sigma_3}$, while rejecting $\frac{\mu_1}{\sigma_1} = \frac{\mu_3}{\sigma_3}$, as discussed in Lehmann and Romano (2005), subsection 9.3. Due to such problems in the aggregation of dependent hypotheses,

the recommended alternative is to aggregate all of these procedures to a single joint test, as is discussed in Lehmann and Romano (2005).

Other studies using bootstrap methods to analyze the performance of funds include Fama and French (2010), Kosowski et al. (2006) and, most notably, Barras et al. (2010), who presents a method to control the False Discovery Rate. Our approach differs from these studies by analyzing the Sharpe ratio, whereas the objective in these studies is the α obtained in multi-factor models. These studies are especially concerned with the direct control of the dependence structure between the returns of the funds, by correcting the GMM simultaneous estimation using HAC matrices for the array of returns.

3. Generalized Tests for Fund Performance

An issue worth noting, similar to the estimation procedure using sample moments, is that it is possible to define the estimation procedures for the mean and variance returns with a GMM estimator (Hansen, 1982), by using a specific vector as moment conditions. This vector is defined as:

$$\varphi(R_t, \theta) = \begin{bmatrix} R_t - \mu \\ (R_t - \mu)^2 - \sigma^2 \end{bmatrix} \quad (5)$$

The GMM estimator is obtained by setting $\frac{1}{T} \sum_{t=1}^T \varphi(R_t, \theta) = 0$. In this case, one may use the HAC matrix in exactly the same manner as proposed by Ledoit and Wolf (2008) for producing the variance-covariance matrix of the estimated parameters. This matrix is used in the GMM estimation procedures, in order to construct the optimal weighting matrix in the case of estimations with more moment conditions than parameters (overidentified system), producing GMM moments through the minimization of the quadratic form $\frac{1}{T} \sum_{t=1}^T \varphi(R_t, \theta)' \hat{V}_{HAC}^{-1} \varphi(R_t, \theta)$. As demonstrated by Hansen (1982), the weighting of this quadratic form by the long-run matrix generates asymptotically efficient estimators when returns are correlated and/or heteroscedastic. Through GMM estimation, we are able to use this HAC matrix to perform tests of the difference between the Sharpe ratios of two funds in the form proposed by Ledoit and Wolf (2008), or, for instance, for instance, we can perform Wald and GMM-Distance type tests (e.g. Anatolyev and Gospodinov, 2011) to analyze joint hypotheses about fund performance.

The tests proposed in the literature for comparing fund performances (Jobson and Korkie, 1981, Ledoit and Wolf, 2008) only involve pairs of Sharpe ratios. These studies did not have a procedure for simultaneously checking the performance of many funds. Testing joint performance is important for determining whether funds exist with superior performance *vis-à-vis* the others. This type of analysis is particularly important, since the performances of investment funds may be highly correlated, which leads all funds to exhibit basically the same performance, differing only by virtue of idiosyncratic shocks. This possibility is associated with the

validity of the efficient market hypothesis (e.g. Leroy, 1989), which conjectures that investment funds or active investment strategies systematically producing higher returns than those generated by passive management cannot exist. In such a situation, the systematic performance of all actively managed funds should be the same, and eventual performance differences in a given period would be caused by non-systematic events. In the finance literature, a test for the equality of the performance of sets of many funds was proposed by Leung and Wong (2010), based on a Hotelling T^2 statistic assuming IID Gaussian returns. As we will show in the empirical analysis section of this paper, this statistic is not robust in reference to those assumptions, and the finite sample distributions are quite distinct from the asymptotic distribution. This is similar to the problems involved in the pairwise comparisons of funds conducted by Jobson and Korkie (1981).

In order to perform the joint test, we have used the previously defined GMM context, and specified a vector of moment conditions for the k sample assets. The vector is comprised of $2k$ moment conditions necessary for the mean and variance parameter estimation of those assets and is defined as:

$$\varphi(R, \theta) = \begin{bmatrix} R_{1t} - \mu_1 \\ (R_{1t} - \mu_1)^2 - \sigma_1^2 \\ \vdots \\ R_{kt} - \mu_k \\ (R_{kt} - \mu_k)^2 - \sigma_k^2 \end{bmatrix} \quad (6)$$

where R_{it} represents the return on fund i in period t , and μ_i and σ_i^2 denote the mean and variance of fund i 's returns. In this unrestricted specification, we have an exactly identified non-linear system. However, under some restriction of fund performance equality the system becomes overidentified, since there are $2k$ moment conditions for estimating a smaller number of parameters under a non-linear equality restriction.

Using this moment condition vector, we perform a GMM estimation, obtaining estimates for the mean and variance parameters, as well as the $\hat{V}_{HAC}$ variance-covariance matrix for those parameters, as was previously defined. In this case, should be noted that matrix does not assume the usual restrictions of absence of correlation among assets, as in the form used by Jobson and Korkie (1981), and that the estimated matrix $\hat{V}_{HAC}$ is not block-diagonal. Based on this estimation, we can test the null hypothesis of equal performance for k funds: $H_0: \frac{\hat{\mu}_1}{\hat{\sigma}_1} = \frac{\hat{\mu}_2}{\hat{\sigma}_2} = \dots = \frac{\hat{\mu}_k}{\hat{\sigma}_k}$, against an alternative H_a hypothesis of the existence of at least one fund with a different performance, controlling for the possible dependency of performance among all funds analyzed.

In order to test this hypothesis, we have used two tests: one based on a Wald test and the other based on a GMM-Distance test, analogous to a likelihood ratio procedure.

The Wald test analyzes the non-linear hypothesis of the joint equality of Sharpe ratios with the following formulation:

$$W = ng(R_t, \theta)' \left(\frac{\partial g(R_t, \theta)}{\partial \theta} \hat{V}^{-1} \frac{\partial g(R_t, \theta)'}{\partial \theta} \right) g(\theta) \quad (7)$$

where $g(\theta)$ is specified as the restriction for the null hypothesis of performance equality among all funds, as previously defined, and $\hat{V}$ is the estimated variance-covariance matrix, with the aforementioned HAC correction. In this case, the model is estimated only for H_a . In other words, the model is estimated without requiring the validity of the performance equality restrictions. However, the Wald test is not invariant to reparametrizations, thus, it may be affected by how the null hypothesis is specified, as is discussed by Gregory and Veall (1985). Hansen (2006), another essential reference, shows that the use of HAC matrices in the construction of the Wald statistic virtually eliminates the Wald test's size problems.

Leung and Wong (2010) propose a test of performance equality based on the Sharpe ratio, predicated on independent, normally distributed returns. Defining the vector of Sharpe ratios estimated with sample estimators (plug-in estimation) as:

$$\hat{S} = \left(\frac{\hat{\mu}_1}{\hat{\sigma}_1}, \frac{\hat{\mu}_2}{\hat{\sigma}_2}, \dots, \frac{\hat{\mu}_k}{\hat{\sigma}_k} \right) \quad (8)$$

Leung and Wong (2010) show that, with the multivariate Delta method, it is possible to demonstrate that:

$$\sqrt{T}(\hat{S} - S) \xrightarrow{D} N(0, \Omega) \quad (9)$$

with Ω representing the asymptotic variance matrix for the Sharpe ratio estimator given by:

$$\Omega = \frac{1}{2} \begin{bmatrix} 2 + S_1^2 & 2\rho_{12} + S_1 S_2 \rho^2_{12} & \cdots & 2\rho_{1k} + S_1 S_k \rho^2_{1k} \\ 2\rho_{12} + S_1 S_2 \rho^2_{12} & 2 + S_2^2 & \cdots & 2\rho_{2k} + S_2 S_k \rho^2_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 2\rho_{k1} + S_1 S_k \rho^2_{1k} & 2\rho_{k2} + S_2 S_k \rho^2_{2k} & \cdots & 2 + S_k^2 \end{bmatrix} \quad (10)$$

The test for the equality of Sharpe ratios proposed by Leung and Wong (2010) is formulated with a null hypothesis given by $H_0 : CS = 0$ against an alternative $H_a : CS \neq 0$, where C is a matrix given as:

$$C = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & -1 \end{bmatrix} \quad (11)$$

Leung and Wong (2010), starting from the assumption of normal IID returns, propose a test based on a Hotelling T^2 statistic, given by:

$$T^2 = t^2 (CS)' (C\hat{\Omega}C) (CS) \quad (12)$$

and a test at level α will reject the null if:

$$T^2 > \frac{(t-1)(k-1)}{(t-k-1)} F_{k-1, t-k-1}(\alpha) \quad (13)$$

where $F_{k-1, t-k-1}(\alpha)$ denotes the 100α percentile of an F distribution with $k-1, t-k-1$ degrees of freedom. It should be noted that the use of this F statistic is based on the normality of returns. Also worth highlighting, is the fact that a version assuming just the asymptotic normality of returns may be developed, leading to the following test statistic:

$$W = t (CS)' (C\hat{\Omega}C) (CS) \quad (14)$$

corresponding to the usual form of the Wald test for linear restrictions, assuming IID returns. In this case, the distribution once more is $\chi^2(q)$, with q denoting the number of restrictions. A version assuming symmetric non-Gaussian distributions was proposed by Galea and Vilca (2010).

An alternative test that is invariant to reparametrizations is the test based on distance metrics, originally formulated by Newey and West (1987a). The test (GMM Distance Metric) is based on the distance between the restricted and unrestricted estimations of the model by GMM, where restricted estimation imposes that the performance of all funds is the same, whereas the unrestricted versions presupposes a different performance for each fund. The GMM-Distance statistic, using the HAC estimator for $\hat{V}$, is constructed in the following manner:

$$D = n(\varphi(R, \theta_{res})' \hat{V}_{HAC}^{-1} \varphi(R, \theta_{res}) - \varphi(R, \theta_{unres})' \hat{V}_{HAC}^{-1} \varphi(R, \theta_{unres})) \quad (15)$$

where θ_{res} denotes the parameter vector, with the performance equality restriction for all funds, and θ_{unres} is the parameter vector estimated in the unrestricted form; thus, the test is based on the weighted quadratic distance between the restricted and unrestricted estimators. The Wald and Distance tests have the same $\chi^2(q)$ asymptotic distribution, where q indicates the number of restrictions in the null hypothesis (e.g. Anatolyev and Gospodinov, 2011). It must be noted that the GMM Distance test is equivalent to a Lagrange Multiplier test, plus a $O_p(1)$ term. Thus, it should inherit the good computational properties of the Lagrange Multiplier test, as discussed in Anatolyev and Gospodinov (2011).

When testing joint fund performance, the GMM Distance statistic can be simplified to:

$$D = n(\varphi(R, \theta_{res})' \hat{V}_{HAC}^{-1} \varphi(R, \theta_{res})) \quad (16)$$

since, under unrestricted estimation θ_{unres} , the objective function $\varphi(R, \theta_{res})' \hat{V}_{HAC}^{-1} \varphi(R, \theta_{res})$ is equal to zero, because the model is exactly identified. Hence, the test is equivalent to the J test for the validity of overidentification conditions, and the number of restrictions is given by the number of overidentification conditions. In this case, we only need to estimate the model under the alternative hypothesis. This formulation is especially important because in this particular problem, the estimation under the alternative hypothesis is computationally simpler. However, in order to perform a test of performance equality using the GMM Distance, we need to modify the moment conditions used, so that a test with the same number of Wald test restrictions is obtained. In this case, we may specify the moment conditions for the Sharpe ratio in the unrestricted model as follows:

$$\varphi(R, \theta_{unres}) = \begin{bmatrix} R_{1t}/(R_{1t}^2 - E(R_{1t})^2) - S_1 \\ \dots \\ R_{kt}/(R_{kt}^2 - E(R_{kt})^2) - S_k \end{bmatrix} \quad (17)$$

and the model is given by the following moment conditions:

$$\varphi(R, \theta_{res}) = \begin{bmatrix} R_{1t}/\sqrt{(R_{1t}^2 - E(R_{1t})^2)} - S_{res} \\ \dots \\ R_{kt}/\sqrt{(R_{kt}^2 - E(R_{kt})^2)} - S_{res} \end{bmatrix} \quad (18)$$

where S_i denotes fund i 's Sharpe ratio. In the present case, we have $k - 1$ overidentification conditions, corresponding to the same number of restrictions in the Wald test. Note that it is possible to execute a more restricted form of an Sharpe ratio equality test, assuming that all assets have identical means and variances through out the estimation of a restricted model, given by:

$$\varphi(R, \theta_{res}) = \begin{bmatrix} R_{1t} - \mu_{res} \\ (R_{1t} - \mu_{res})^2 - \sigma_{res}^2 \\ \dots \\ R_{kt} - \mu_{res} \\ (R_{kt} - \mu_{res})^2 - \sigma_{res}^2 \end{bmatrix} \quad (19)$$

In this case, the number of restrictions is equal to $2k - 2$. This procedure is a semiparametric version of the joint test of equality of means and variances proposed by Galea and Vilca (2010), where a Wald test is suggested for this same hypothesis.

Similar to the test procedure using the HAC matrix, GMM estimation contains a large bias problem in finite samples, which is a detriment to the properties of the resulting tests of hypotheses. However, in the particular case of a test of joint

performance involving multiple funds, the test's bias problem is amplified by the large number of moment conditions used, as is discussed by Hansen et al. (1996) and Anatolyev and Gospodinov (2011). Thus, the finite sample distribution may differ substantially from the asymptotic distribution. As discussed by Hansen (2006), the Distance Metric-based test is robust to heteroscedasticity and serial correlation problems when the test statistic is constructed with the use of an HAC matrix.

Another relevant problem is the number of moment conditions and restrictions used in this procedure. As discussed by Anatolyev (2012), in testing procedures in which the number of restrictions is a significant fraction of the sample size, the classical asymptotic tests (Wald, Likelihood Ratio and Lagrange Multiplier) are not consistent. This type of problem had already been reported in the literature. For example, Berndt and Savin (1977) discuss significant differences between these tests when the number of restrictions is comparable to the number of observations. Rothenberg (2006) shows that the asymptotic χ^2 distribution is a poor approximation to the Wald test in this same situation. The work of Burnside and Eichenbaum (1996) is especially significant because it shows that in the estimation of a non-linear model by GMM, the size of the Wald test is larger than the theoretical size, and this size grows with the number of restrictions.

A related analysis is the multivariate extension of the Behrens-Fisher problem. This procedure deals with tests for the comparison of means when the variances are not known and, in spite of the problem's simplicity, there is no simple solution to produce an optimal test. As demonstrated by Belloni and Didier (2008), in the multivariate extension of the Behrens-Fisher problem, when the sample size is small, relative to the number of restrictions, the Wald and LR tests cause excessive rejection of the null hypothesis, when compared to the asymptotic χ^2 values. The Lagrange Multiplier test is more robust in relation to this problem.

In order to correct the problems associated with testing a hypothesis with the GMM estimation in finite samples, when there is a large number of moment conditions and restrictions, we replace the asymptotic χ^2 distribution of such tests with a distribution constructed thanks to a block bootstrap procedure. This is similar to the procedure proposed by Ledoit and Wolf (2008). As we will shortly see, the bootstrap correction is essential in joint performance tests. Another justification for the use of the bootstrap is that it can work as a Bartlett correction for the test statistics, as is discussed by Lehmann and Romano (2005). For example, the usual likelihood ratio test has an error rate, in terms of rejection probabilities, of the order $O(T^{-1})$, whereas the likelihood ratio test's error rate, with the bootstrap correction, is of the order $O(t^{-2})$.

The good properties of the bootstrap estimator for the GMM-derived test statistics can be found in the work of Hall and Horowitz (1996). These properties are also addressed by Inoue and Shintani (2006), who demonstrate the properties of the bootstrap estimator when the GMM setting is based on HAC matrices.

The properties of the overidentification tests in models with non-linear moment conditions using the block-bootstrap were analyzed by Ouyssse (2010). Her results show that the block-bootstrap procedure using fixed-size blocks provides results than the stationary bootstrap procedures using random-size blocks.

The procedure for constructing the Wald and Distance GMM statistics by the bootstrap is similar to used by Ledoit and Wolf (2008), with an adjustment to the set of funds employed.

At each bootstrap iteration, one selects blocks of size m from all funds, until a bootstrap sample of a size equal to the original. On the basis of this replication, the Wald and Distance GMM statistics are constructed. In the case of the Wald statistic, we only have to estimate the model in the unrestricted form, while we only estimate the restricted model for the Distance GMM test.

From these estimations we obtain the bootstrap p -values in the usual manner (Efron and Tibshirani, 1993), and calculate the p -values for the Wald and Distance GMM tests as follows:

$$\begin{aligned} p_W &= B^{-1} \sum_{i=1}^B \mathbb{I}\{W < W_i^*\} \\ p_D &= B^{-1} \sum_{i=1}^B \mathbb{I}\{D < D_i^*\} \end{aligned} \quad (20)$$

where $\mathbb{I}$ is the indicator function, W and D are the values of the test statistics obtained from the original sample, W_i^* and D_i^* are the values in each block-bootstrap replication, and B denotes the number of bootstrap replications. Specific details on the implementation of these tests are provided in the empirical section that follows.

A method of analysis related to this joint performance test methodology is represented by the analyses that control for the false discovery rate, e.g. Romano et al. (2008). In this class of analyses, joint tests are constructed from an aggregation of the results of individual tests of hypotheses. This is usually done on the basis of the p -values calculated for each individual test, controlling for false discovery rate (FDR) problems, where the FDR is defined as the proportion of false rejections when the number of comparisons is large. Romano et al. (2008), in particular, propose mechanisms for controlling false discovery rates by taking into account the dependency among test statistics.

The joint test procedure proposed in the present paper may be viewed as a special case, in which it is possible to directly construct a p -value associated with the null hypothesis of interest. In our case, the p -value may be directly constructed for the null hypothesis $\frac{\mu_1}{\sigma_1} = \frac{\mu_2}{\sigma_2} = \dots = \frac{\mu_k}{\sigma_k}$ through the use of the moment conditions and the use of the full HAC matrix that estimates all covariances among the estimated parameters. This is a more simple way of controlling for the FDR than the Romano et al. (2008) procedures. We compare the results from using our methodology with this class of methods in the following section.

Figure 1
Return distribution

4. Empirical Analysis

The procedures were implemented using a sample of daily returns for 97 actively managed, unlevered, open end stock funds in Brazil, covering the period from July 31, 2006 to July 31, 2008, resulting in a total of 500 observations. The sample comprises the largest stock funds in operation in Brazil during that period. In other words, that is, the 97 funds had the largest net worth amounts in the period. Given the large number of funds, a few descriptive statistics are presented with the use of charts and image plots.

Daily fund returns are calculated as logs of daily fund share value ratios. In turn, a fund's share value on any given day is defined as the ratio of the fund's net worth to the number of existing shares. A fund's number of shares changes with share purchases and redemptions. A fund's net worth on that day is the sum of the market values of its positions on that day, minus an amount resulting from the application of a proportion of annual management fees and expenses. For example, if the management fee is equal to four percent per annum, an amount equal to four percent of the fund's net worth, divided by 365, is deducted from the market value of the fund's investment portfolio. Hence, share values correspond to net asset values in U.S.-market investment funds, and the rates of return closely represent the before-tax returns available to stock fund investors. The sample funds are considered unlevered because they belong to a class of stock funds where short selling or entrance into equity derivative transactions is not allowed. Finally, the open ended nature of the sample funds precludes their shares from being traded in organized stock markets.

Figure 1 displays a superposition of the distributions, which were non-parametrically estimated with kernel density for the actively managed funds in the sample. This figure makes it clear that the distributions of returns are very

Figure 2
Sharpe ratio distribution

Figure 3
Correlations

similar, except for a few outlying observations.

The distribution of the Sharpe ratios, is displayed in Figure 2, computed for the sample period and the funds covered, together with a Gaussian approximation. A reasonable dispersion in terms of Sharpe ratios is observed in the 0.02 to 0.08 range. It is also important to notice that the distribution of these ratios is quite different from a Gaussian distribution.

In order to provide evidence indicating the high level of dependency among fund performances, we have displayed the correlations among all sample funds in an image plot (Figure 3). In this plot, funds are arranged according to their Sharpe ratios, from the lowest to the highest, and we calculated the correlations for the returns of all funds. As the illustration makes clear, the general standard is that of highly correlated returns, providing support to the expectation that there

Table 1
Rejection proportions

α	10%	5%	1%
IID	0	0	0
GMM Newey-West	0.1968	0.0997	0.0288
GMM Andrews	0.1170	0.0531	0.0137
Bootstrap	0.0720	0.0363	0.0075

may not be performance differences for the sample funds.

4.1 Pairwise comparisons

In our first analysis, we used the pairwise comparison procedure proposed by Jobson and Korkie (1981) and Ledoit and Wolf (2008). Estimation and tests made use of a variance-covariance matrix unadjusted for heteroscedasticity and serial correlation (Jobson and Korkie, 1981), HAC matrices with distinct rules for bandwidth estimation (Newey and West, 1987b, for fixed bandwidth), the Andrews (1991) procedure and the block-bootstrap procedure proposed by Ledoit and Wolf (2008). Table 1 displays the total rejection proportions obtained in the pairwise comparisons using all the aforementioned tests at the ten, five and one percent significance level. The total proportion of p -values that are below the test's significance level is summarized for all tested fund combinations.

Figure 4 presents the image plot with the p -values obtained with the Jobson and Korkie (1981) performance comparison procedure, which assumes that returns are Gaussian IID. As indicated by this figure and in the first line of Table 1, the results show that, according to the Jobson and Korkie (1981) method, there are no significant differences among the investment funds analyzed. This is because the test indicates no rejection of the null hypothesis of fund performance equality at any considered significance level.

The fund performance comparisons using the HAC estimators are displayed in Figure 5. This figure contains the results from two methods for constructing HAC estimators, the Newey and West (1987b) rule with fixed bandwidth, and the Andrews (1991) rule. In all the analyses we employed a Bartlett kernel function, but other kernel functions were tested, with similar results. The fixed bandwidth Newey-West estimator uses a rule based on the number of sample observations in order to determine bandwidth S_t , and the Bartlett kernel. This rule is given as:

$$S_t = \text{int}(4(T/100)^{2/9})$$

where int represents the integer part.

The Andrews (1991) method assumes that sample moments follow a first-order autoregressive process. In this method, an AR(1) process is fit for every moment condition (Equation 6), and from there, one estimates the correlation coefficients

Figure 4
Fund performance – IID Procedure

$\hat{\rho}_i$ and the residual variance ($\hat{\sigma}_i^2$), where z denotes the number of instruments used and n is the number of moment conditions. The rule proposed in Andrews (1991) for the Bartlett kernel indicates that the optimal choice of S_t is given by:

$$S_t = \text{int}(1.1447(\hat{\alpha}T)^{1/3})$$

with

$$\hat{\alpha} = \left(\sum_{i=1}^{zn} \frac{4\hat{\rho}_i^2 \hat{\sigma}_i^4}{(1 - \hat{\rho}_i)^6 (1 + \hat{\rho}_i)^2} \right) / \left(\sum_{i=1}^{zn} \frac{\hat{\sigma}_i^4}{(1 - \hat{\rho}_i)^4} \right) \quad (21)$$

The results displayed in Figure 5 and lines two and three of Table 1 show that these two corrections produce results that differ from those generated by the Jobson and Korkie (1981) method. The procedure using the Andrews (1991) rule generates a number of rejections, which is similar to the test's nominal size (rejection proportions of 0.1170, 0.0531 and 0.0137 for the ten, five and one percent significance levels). This is evidence in favor of the null hypothesis of equal fund performance. The use of the Newey and West (1987b) rule, however, in general leads to a number of rejections that are approximately twice that of the test's nominal size (rejection proportions of 0.1968, 0.0997 and 0.0288 for the ten, five and one % significance levels). This is evidence of the existence of funds with significantly different performances. However, as was previously discussed, tests of hypotheses based on HAC matrices can be severely distorted in finite samples, and in that situation we use the Ledoit and Wolf (2008) bootstrap procedure.

Through the Ledoit and Wolf (2008) block-bootstrap procedure, we performed all pairwise comparisons. We used 5,000 bootstrap replications, as well as the calibration procedure for the optimal block size and test size, which was also described by Ledoit and Wolf (2008). In this procedure, the optimal average block size is

Figure 5
Fund Performance : HAC Procedure

Figure 6
Bootstrap

equal to three. The results (Figure 6) show that the block-bootstrap correction does not change the results obtained by the procedure using HAC matrices in any significant way. In this case, the number of rejections is close to the test's nominal size, with rejection proportions equal to 0.0720, 0.0363 and 0.0075 for the ten, five and one % significance levels.

The overall results indicate that IID procedures, HAC matrices using the Andrews rule and the Ledoit and Wolf (2008) bootstrap method generate rejection proportions similar to, or below the nominal significance level. This outcome can be interpreted as evidence in favor of the null hypothesis in Neyman-Pearson testing framework, whereas the HAC procedure using the Newey-West rule produces rejection proportions higher than the test's nominal size. This is evidence against the null of performance equality, however; we must remember the procedure's

Figure 7
Adjusted P -Values – GMM Newey-West

possible distortion problems associated with finite samples.

For the sake of comparison, we used the corrections for the p -values obtained through multiple comparison procedures with the Bonferroni, Holm (1979), Hochberg (1988), Hommel (1988) controls for the Family-wise Error Rate, as well as the Benjamini and Hochberg (1995) corrections for False Discovery Rate and the Benjamini and Yekutieli (2001) correction for dependency FDR in the p -values generated by the Newey-West GMM procedure. These procedures were used because they were the cases in which the rejection probabilities were higher. The p -values corrected for all pairwise comparisons are displayed in Figure 7, where BH denotes the Benjamini and Hochberg (1995) procedure and BY is the Benjamini and Yekutieli (2001) procedure. Although in this situation the only procedure containing the appropriate control for dependency among tests is that proposed by Benjamini and Yekutieli (2001), it is possible to observe that all the corrections for individual p -values in Fig. 7 provide support for the null hypothesis of performance equality for all funds analyzed. This is shown by the fact that the lowest p -value obtained in all of the types of corrections is equal to 0.091 for only one of the comparisons performed, with all other p -values being higher than 0.3.

4.2 Joint performance test

In accordance with the methodology proposed in Section 3, the joint testing procedure using GMM was implemented for the actively managed stock funds. GMM estimation was based on the moment conditions given in Equation 6, representing a set of 2×97 moment conditions in our sample.

We first used the Wald test (Equation 7) without the HAC matrix. The value obtained for the Wald statistic in our sample was equal to 2.865, well below the asymptotic critical value of $\chi^2(96)$, which is equal to 119.870 for the 95 percent significance level. The p -value for this statistic is virtually equal to one, indicating

Figure 8
Fund Performance : Wald Test – IID

the validity of the null hypothesis of equal fund performance. However, due to the power and size problems of the Wald test, we used the bootstrap procedure discussed in Section 3, in order to obtain the bootstrap p -value for this hypothesis. In the determination of the optimal block size, we adopted the Politis and Romano (1992) procedure, which also points to an optimal size of three observations per block in the circular block-bootstrap procedure. The bootstrap procedure was based on 1,000 replications.

Figure 8 displays the distribution for the obtained bootstrap statistic. The vertical line indicates the value of the Wald statistic produced for the original sample. The bootstrap p -value (Equation 20) was equal to 0.805, providing support for the null hypothesis of performance equality among all funds.

However, the Wald test procedure may have distorted properties in finite samples, as was discussed in Section 3. Therefore, we adopted the procedure suggested by Hansen (2006), to use HAC matrices in the construction of the Wald statistic. We used an HAC matrix with the same Newey-West HAC estimator with fixed bandwidth that was employed in the preceding section.

The estimated value for the Wald statistic using the HAC matrix is 1269.77, showing that this correction affects the test statistic in a significant way. This is mainly thanks to the inclusion of the correlations among each fund's parameters. In order to check the distribution of this statistic in finite samples using the HAC matrix, we adopted the same bootstrap procedure that was previously discussed. This procedure was done to control for the potential problems with HAC matrices in finite samples. Figure 9 shows the distribution of the Wald-HAC statistic obtained from the bootstrap. The illustration indicates that the finite sample distribution of this statistic with the HAC correction differs significantly from the

Figure 9
Fund Performance : Wald Test – HAC

Table 2
Restricted model

	Coefficient	Std. error	<i>t</i> -Statistic	<i>p</i> -value
S_{res}	0.048678	0.019346	2.516130	0.0119

asymptotic distribution, thus emphasizing the importance of the bootstrap correction. After the bootstrap procedure, we obtained a *p*-value of one, which is evident in Figure 9. Once more, this provides support for the null hypothesis of fund performance equality. This discrepancy between the asymptotic values and the finite sample distribution may be linked to the invariance problems of the Wald test, as is discussed by Gregory and Veall (1985) and Hansen (2006).

In order to avoid the problems with the Wald test, we implemented the GMM Distance test (Newey and West, 1987a), following the Hansen (2006) proposal, thus entailing the use of the HAC matrix. In the present case, the construction of the GMM Distance statistic requires only the model's estimation under the alternative hypothesis, with the imposition of the performance equality restrictions among all funds. Table 2 presents the estimated coefficients under the null hypothesis of fund performance equality. The value of the Sharpe ratio under this restriction is equal to 0.0486, which is close to the sample mean of the Sharpe ratios for all sample funds (0.049).

The value of the GMM Distance statistic for this sample was equal to 83.651, corresponding to a *p*-value of 0.8161 for the null of fund performance equality using the $\chi^2_{.95}(96)$ asymptotic distribution. Once more, this gives support to the null of equal performance. In order to check the finite sample properties, we adopted the

Figure 10
Fund Performance : GMM Distance Test

same block-bootstrap procedure used in previous analyses. This procedure deals with the finite sample problems in this test, and is equivalent to the J validity test of overidentification conditions (e.g. Anatolyev and Gospodinov, 2011).

The distribution of the GMM Distance statistic is provided in Figure 10, where the vertical line indicates the statistic obtained in the original sample. The bootstrap p -value (Equation 20) obtained with this procedure was 0.1768, and this value also supports the validity of the null hypothesis of equal fund performance.

We may compare the results with the T^2 statistic proposed by Leung and Wong (2010). In our sample, the value of the statistic (Equation 12) is 5515.523, and the critical value for rejection at an α of 0.95 (Equation 13) is 152.9081, leading to the rejection of the null hypothesis of Sharpe ratio equality. However, this statistic assumes Gaussian IID returns, and its power may be affected in finite samples. In order to obtain the finite sample distribution for the T^2 statistic proposed by Leung and Wong (2010), we ran the same block-bootstrap procedure adopted for the other joint test statistics.

As in the case of the Wald test, the finite sample distribution obtained from the bootstrap is quite different from the asymptotic distribution. In this case, through the block-bootstrap procedure we obtained a p -value of 0.860, which once more provides support for the null of fund performance equality. Since the Leung and Wong (2010) T^2 statistic is nothing more than a transformation of the Wald statistic, it suffers from the same power and size distortion problems of the Wald test when there is a large number of restrictions and moment conditions. This consequence points to the need for a correction in finite samples. The Leung and Wong (2010) T^2 statistic also suffers from a lack of robustness to the assumptions of normality and IID returns.

Figure 11

Fund Performance : Leung and Wong (2010) – Leung and Wong (2010) T^2 Test

5. Conclusion

The evaluation of investment fund performance is an important procedure in finance, and the development of robust statistical methodologies for that evaluation is a highly relevant line of research. In this paper, we have proposed a generalization of pairwise fund comparison methodologies (e.g. Jobson and Korkie, 1981, Ledoit and Wolf, 2008) to permit the joint evaluation of the performance of many investment funds using their Sharpe ratios. In addition, the present analysis sheds light on the relative value of the use of different investment strategies for the funds offered to the investing public.

This methodology may be implemented with the tools developed for GMM estimation. For this reason we are able to use the properties of this semi-parametric estimator, which does not require distributional assumptions and is robust to autocorrelation and heteroscedasticity problems, thanks to the use of HAC weighting matrices. Even though GMM-based estimators and tests have asymptotic optimality properties, it is possible to obtain distributions and tests with the correct power and size in finite samples through block-bootstrap procedures (e.g. Ledoit and Wolf, 2008, Gregory and Veall, 1985). This is the exact route we have taken in this paper. In particular, the nice properties of the GMM Distance mean that this test is the appropriate methodology for the joint fund performance evaluation. Herein, the results obtained with this procedure show that tests based on the asymptotic distribution and the bootstrap-corrected test point to the same result, in contrast with the Wald test.

The GMM methodology facilitates the direct construction of a joint test statistic and, similarly to the Romano et al. (2008) procedure, it also incorporates the

dependency among the individual statistics, through the HAC matrix obtained for the moment condition system.

The results of our empirical analysis show that there are no significant performance (risk-adjusted return) differences among the investment funds examined in the sample period. The pairwise comparison procedures indicate a number of rejections that are similar to the test's nominal size, providing evidence in favor of a homogeneous fund performance hypothesis. Once more, this evidence is supported by the proposed joint test procedures, with their bootstrap corrections.

Finally, the results represent strong evidence for attesting that representative actively managed Brazilian stock funds cannot be considered as reasonable substitute investment vehicles. This provides evidence, through a novel and more consistent methodology, for the so-called “herding behavior” hypothesis. The results do not even support the recommendation that investors choose funds with lower management fees, since the returns used in the computation of returns for our sample Sharpe ratios are already adjusted for management fees.

References

- Anatolyev, S. (2012). Inference in regression models with many regressors. *Journal of Econometrics*, 170:368–382.
- Anatolyev, S. & Gospodinov, N. (2011). *Methods for Estimation and Inference in Modern Econometrics*. Chapman and Hall/CRC.
- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.
- Barras, L., Scaillet, O., & Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance*, 65:179–216.
- Belloni, a. & Didier, G. (2008). On the Behrens-Fisher problem: A globally convergent algorithm and a finite-sample study of the Wald, LR and LM tests. *Annals of Statistics*, 14:294–308.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300.
- Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.
- Berndt, E. & Savin, N. (1977). Conflict among criteria for testing hypotheses in the multivariate linear regression model. *Econometrica*, 45:1263–1277.

- Burnside, C. & Eichenbaum, M. (1996). Small sample properties of GMM-based Wald tests. *Journal of Economic & Business Statistics*, 14:294–308.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to Bootstrap*. Chapman & Hall.
- Fama, E. & French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance*, 65:1915–1947.
- Galea, M. & Vilca, F. (2010). Hypotheses testing for comparing means and variances of correlated responses in the symmetric non-normal case. *Chilean Journal of Statistics*, 1(2):113–123.
- Gregory, A. W. & Veall, M. (1985). Formulating Wald tests of nonlinear restrictions. *Econometrica*, 53:1465–1468.
- Grinblatt, M., Titman, S., & Wermers, R. (1995). Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. *American Economic Review*, 85:1088–1105.
- Hall, P. & Horowitz, J. (1996). Bootstrap critical values for tests based on generalized method of moments estimators. *Econometrica*, 64:891–916.
- Hansen, B. (2006). *Econometric Theory and Practice*, chapter Edgeworth Expansions for the Wald and GMM Statistics for Nonlinear Restrictions. Cambridge University Press.
- Hansen, L. P. (1982). Large sample properties of Generalized Method of Moments estimators. *Econometrica*, 50:1029–1054.
- Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics*, 14:262–280.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–803.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.
- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383–386.
- Inoue, A. & Shintani, M. (2006). Bootstrapping GMM estimators for time series. *Journal of Econometrics*, 133(2):531 – 555.

- Jobson, J. D. & Korkie, B. M. (1981). Performance hypothesis testing with Sharpe and Treynor measures. *Journal of Finance*, 36:889–908.
- Kosowski, R., Timmermann, A., & Wermers, R. (2006). Can mutual fund ‘stars’ really pick stocks? New evidence from a Bootstrap analysis. *Journal of Finance*, 61:2551–2588.
- Lakonishok, J., Shleifer, A., & Vishny, R. W. (1992). The impact of institutional trading on stock prices. *Journal of Financial Economics*, 32:23–43.
- Ledoit, O. & Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 19:850–859.
- Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.
- Leroy, S. F. (1989). Efficient capital markets and martingales. *Journal of Economic Literature*, 27:1583–61.
- Leung, P. & Wong, W.-K. (2010). On testing the equality of the multiple sharpe ratios, with application on the evaluation of iShares. SSRN.
- Newey, W. K. & West, K. D. (1987a). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 28:777–787.
- Newey, W. K. & West, K. D. (1987b). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.
- Ouysse, R. (2010). Finite sample properties of bootstrap GMM for nonlinear conditional moment models. *InterStat Journal*.
- Politis, D. N. & Romano, J. P. (1992). *Exploring the Limits of Bootstrap*, chapter A circular block resampling procedure for stationary data, pages 263–270. Wiley.
- Romano, J., Shaikh, A. M., & Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17:417–442.
- Rothenberg, T. J. (2006). *Handbook of Econometrics*, chapter Approximating the distributions of econometric estimators and test statistics. Elsevier.