Working Paper Series, N. 17, December 2010

On the stability of performance measures over time

Giovanna Menardi
Department of Statistical Sciences
University of Padua
Italy

Francesco Lisi
Department of Statistical Sciences
University of Padua
Italy

Abstract: Performance persistence is a relevant issue when evaluating the predictability of future results of managed portfolios. A related crucial aspect is the stability over time of the measure used to assess the performance, defined as the degree of association between the rankings of financial assets induced by the performance measure throughout subsequents periods. In this work a general class of possible criteria to measure stability is proposed. Then, the attention is focused on a specific index, whose asymptotic expected value and variance are derived under the null hypothesis of absence of stability. Furthermore, two statistical tests for evaluating the significance of stability are discussed. An application to a large set of US equity mutual funds shows that stability may remarkably vary, as the performance measure or the time widow width where it is computed change.

Keywords: performance measures, rankings, stability over time.

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Giovanna Menardi
tel: +39 049 827 4168
menardi@stat.unipd.it
http://www.stat.unipd.it/~menardi

On the stability of performance measures over time

Giovanna Menardi
Department of Statistical Sciences
University of Padua
Italy

Francesco Lisi
Department of Statistical Sciences
University of Padua
Italy

Abstract: Performance persistence is a relevant issue when evaluating the predictability of future results of managed portfolios. A related crucial aspect is the stability over time of the measure used to assess the performance, defined as the degree of association between the rankings of financial assets induced by the performance measure throughout subsequents periods. In this work a general class of possible criteria to measure stability is proposed. Then, the attention is focused on a specific index, whose asymptotic expected value and variance are derived under the null hypothesis of absence of stability. Furthermore, two statistical tests for evaluating the significance of stability are discussed. An application to a large set of US equity mutual funds shows that stability may remarkably vary, as the performance measure or the time widow width where it is computed change.

Keywords: performance measures, rankings, stability over time.

1 Introduction

The association between current performance and past patterns of managed portfolios (i.e. mutual funds) has been widely discussed in the recent years by the financial academic and practitioners' communities. A key topic of the research has, in particular, focused on analyzing the predictability of the behavior of financial assets in terms of performance persistence over time (see, among others, Carhart, 1997, Beckers and Thomas 2010). The inherent literature mainly aims at revealing any evidence of persistence in both an absolute and a relative sense (that is, for instance, whether winners in a period are the same at the next period or whether some managers show particular skills) and for different temporal evaluation horizons. The issue has lead to controversial results and the discussion is still far from being convergent to some unambiguous conclusions.

Hendricks *et al.* (1993), Goetzmann and Ibbotson (1994), Brown and Goetzmann (1995), Bollen and Busse (2005), for instance, have found evidence of persistence in mutual fund performance over short-term horizons. Instead, Grinblatt and Titman

(1992) and Elton *et al.* (1993, 1996b), enlighten the predictability of mutual funds returns over longer horizons. Moreover, while Grinblatt and Titman (1992) report that relative persistence over time is consistent with the ability of fund managers, Carhart (1997) shows results not supporting the existence of skilled or informed mutual fund portfolio managers.

The issue is even more crucial when the predictability of hedge funds is considered. Investments in hedge funds involve significant lock-up periods, implying that the investors need to have sufficient information about the performance of hedge funds over a long period before committing their money to them. Brown *et al.* (1999) show that there is virtually no persistence in their sample. In contrast, Park and Staum (1998) present some evidence for performance persistence among hedge funds. On the other hand, Agarwal and Naik (2000) document that persistence is revealed on the lower tail of the distribution of the performance that is, it is mostly driven by losers funds continuing to be losers rather than winners being winners.

Besides these controversial results, the practice of using information about past performance to gain some hints about the future behavior of financial assets is still very common: fund companies still pelt investors with advertisement touting past returns of their funds, and investors - not only private - often base their choice of investments on the observation of the past performance of financial products. Hence, a fair evaluation of performance persistence is capital.

Whatever approach is followed, the analysis of persistence is generally based on some specific performance measures, namely reward-to-risk ratios, providing information about the assets' behavior throughout time. However, dozens and dozens of measures have been introduced in the scientific and the practitioners' literature, each of them offering different information and specific perspectives about the trade-off between return level and risk exposure (see, for instance, the reviews of Le Sourde, 2007 and Aftalion and Poncet, 2003). Which measure to choose to evaluate the performance depends on several aspects, as the kind of assets, the context of the investment and the individual idea of performance (Hübner, 2007).

Moreover, it is important to analyze the quality of the performance measures by looking at their intrinsic properties. From a predictive perspective, this may be translated in requiring the ability of a performance measure in providing as much information as possible about the assets' future behavior. Following this direction, this work moves from the intention of analyzing the extent to which performance measures may be used for predictive purposes. In particular, this study rests on the idea that any lack of persistence resulting in analyzing the past performance of financial assets might be due not only to the lack of persistence of the assets themselves, but also to the use of an instrument to measure performance, characterized by a scarce predictive power. Therefore, unlike several previous studies about persistence in financial assets'performance, in this work it is the *stability* of the performance measures to be under analysis and not the measures themselves.

The issue has been receiving increased attention in the last few years. Broihanne *et al.* (2008), for instance, propose an efficiency index and investigate its robustness over different sets of funds. Hübner (2007) aims at measuring the precision of a performance index in reproducing the ranking induced by the correct asset pricing model and evaluates the stability of the rankings under alternative pricing models.

A similar evaluation is conducted by Bodson *et al.* (2008). In the same work, the authors address the problem of examining if there exist any measure able to uncover some persistence in performance in a two-periods framework.

In this work we conjecture that performance persistence may be the results of two components: the first component is the aptitude of the financial assets to reproduce their behavior over time, the second one is the intrinsic stability of the instrument used to measure such behavior, namely the performance index. We, then, focus on the second component, to evaluate whether some structural stability does occur and if some measures show a larger stability.

Although stability is important for prediction, it is worth emphasizing that such property cannot be wished separately from other considerations, such as information provided by the measures, the final purpose of the analysis and the context where it is applied. Focusing on stability may be, instead, useful when it is considered as a tool for choosing among different and seemingly equivalent performance measures. A recent stream of research (see, among others, Eling and Schuhmacher, 2007 and Eling, 2008), for instance, advocates the virtual interchangeability of many performance measures in the evaluation financial assets: in such a context, addressing the choice of the performance measure toward the most stable one may be discriminant in providing more reliable results in the analysis of performance predictability.

In the same spirit of Bodson *et al.* (2008), we empirically evaluate the stability of performance measures on a large set of U.S. equity mutual funds. Unlike them, however, the analysis is carried out in a multi-period setting, and quite a large amount of performance measures are considered.

The objectives and contributions of this work may be summarized as follows: first, we give a definition of stability of a performance index over time and discuss some possible criteria to measure such stability (Section 2). Then, we investigate the statistical properties of a specific stability index and derive two possible tests to evaluate its significance (Section 3). Finally, we present an application to mutual funds data. In particular, Section 4.1 briefly reviews the performance measures whose stability is then investigated. Section 4.2 shows the results of the application. Some final remarks are given in Section 5.

2 How to measure the stability of performance measures

In order to evaluate if the choice of the performance measure may affect the predictability of future behavior of financial assets or, equivalently, if there exist performance indexes intrinsically more stable than others, a precise specification about the notion of stability is due.

In the following, we define *stability* of a performance measure the degree of similarity among the rankings induced by that performance measure throughout time on a population of financial assets. The maximum stability is reached when a performance measure produces unchanging rankings over the targeted population of assets, as time varies. When the rankings induced by a performance measure are random permutations over time, the measure is said *fully unstable* or simply *unstable*. Any measure that is not unstable has some degree of stability and its quantification is one of the goal of this work. While in principle, an *anti-stable* behavior, corresponding to a negative association of the rankings over the time, may be possible, such a situation is of little interest and will not be considered here.

It is worth noting that the provided notion of stability refers to an intrinsic property of a performance measure and, thus, should not depend on the specific set of assets or the specific period of time under consideration.

Given this definition, any criterion derived from a measure of association between ranks (observed over time) may be, in principle, considered as a candidate to assess the stability of a performance measure. Widely used statistics are, for instance, the rank correlation indexes, such as the Kendall's τ and the Spearman's ρ. The former is a linear function of the number of pairs which are in different orders in two rankings, the latter is the Pearson product moment correlation coefficient between the (untied) ranks. However, these measures compare the rankings induced by the performance measure at only two periods of time that is, in general, not enough for a good evaluation of persistence over time. Moreover, they are meant to detect both positive and negative association between ranks while, when evaluating the stability of a performance measure, a negative association between ranks should matter the same as the lack of association. Another issue is that the Kendall's τ takes into account only the number of inversions between ranks, without considering the size of the inversions that, in our context is surely relevant.

All these reasons motivate the choice of a different, although related, index. To introduce this new index, let $A_1, \ldots, A_i, \ldots, A_n$ be a set of n financial assets, whose historical returns are observed at times $1, 2, \ldots, T$ and suppose to partition the time window $[1, T]$ into P contiguous sub-periods having length l, so that $l \cdot P = T$; for instance, a 10 years long monthly time series may be divided into 5 adjacent windows each having length 24 months.

Now, suppose that M is a performance index, aimed at evaluating the returns per unit of risk of each asset in a given period and that $m_i^{(p)}$ is the estimate of M over the period p for the asset A_i, $p = 1, \ldots, P$, $i = 1, \ldots, n$. Finally, let $R_i^{(p)}$ denote the rank of of the asset A_i, induced by m over the set $A_1, \ldots,\ A_i, \ldots, A_n$ in the period p and consider the difference of ranking between adjacent periods, $d_i^{(p)} = R_i^{(p)} - R_i^{(p-1)}$.
According to the previous definition, a general criterion to measure the stability of the performance measure M is to consider the following index:

$$I_0(M) = \frac{1}{P-1} \sum_{p=2}^{P} \varphi\left(d_i^{(p)}\right), \tag{1}$$

where φ is a suitable function of the difference between rankings at two contiguous periods.

Some possible functional forms for $\varphi(\cdot)$ are:

- $\varphi\left(d_i^{(p)}\right) = \sum_{i=1}^{n} \left(d_i^{(p)}\right)^2$ (quadratic);

- $\varphi\left(d_i^{(p)}\right) = \sum_{i=1}^{n} \left|d_i^{(p)}\right|$ (linear);

- $\varphi\left(d_i^{(p)}\right) = \sum_{i=1}^{n} \left|d_i^{(p)}\right|^q$ (q-adic);
- $\varphi\left(d_i^{(p)}\right) = \max_i \left(d_i^{(p)}\right)$ (maximum);
- $\varphi\left(d_i^{(p)}\right) = d_\pi$ ($\pi-$quantile).

with d_π, the $\pi-$quantile of the $d_i^{(p)}$. Each of these forms could be appropriate in specific contexts. The quadratic function has the nice feature of being directly related to the Spearman's rank coefficient. However, as well as for other functions among those listed, it results in weighting more large variations. If, on the one hand, this is properly our intent, on the other hand it implies that the stability index might be dominated by few very large differences. This reason suggests us to derive our stability index from a linear φ, by taking the sum of the absolute differences between ranks:

$$I_0(M) = \frac{1}{P-1}\sum_{p=2}^{P}\frac{\sum_{i=1}^{n}|d_i^{(p)}|}{\varphi_{max}}, \tag{2}$$

where $\varphi_{max} = \lfloor\frac{1}{2}n^2\rfloor$ is a normalizing factor, corresponding to the maximum value of $\sum_i |d_i^{(p)}|$, that occurs when $R_i^{(p)} = n + 1 - R_i^{(p-1)}$, for $i = 1, \ldots, n$ and $p = 2, \ldots, P$. It follows that if ranks totally reverse from a period to the next, $I_0(M) = 1$, while it is equal to zero when all the $d_i^{(p)}$ are null, i.e. when rankings do not change over time. For intermediate situations we have $0 < I_0(M) < 1$.

The function φ chosen in (2) is known in the statistical literature as the Spearman's footrule. Dismissed by Kendall (1970) as a measure of association between ranks because of a lack of statistical properties, the interest in the Spearman's Footrole was renewed by Diaconis and Graham (1978) and recently again by Genest et al. (2010) after they found some interesting features of the index.

Being an average over different periods of the normalized sums of the absolute differences between ranks, $I_0(M)$ is potentially robust to the choice of the specific set of assets and periods of time considered. Also, note that $I_0(M)$ does not refer to the value assumed by the performance measure, but only to the relative ranking among assets. As a side effect, this index can be used only if we can assume that the set of assets is the same over the period $[1, T]$.

In order to be consistent with the given definition of stability - that should be maximum for a perfect association between rankings - the one's complement operator of I_0 is considered. Thus, the final formulation of the stability index is:

$$I(M) = (1 - I_0(M)). \tag{3}$$

It follows that $0 \leq I(M) \leq 1$, taking the index its minimum value when the rankings are inverted over adjacent periods (that is, when M is anti-stable) and maximum value when the rankings remain unaltered over time.

3 Testing the significance of the stability

Although the notion of stability should be ideally independent on the specific set of financial assets and periods considered, its measurement is, in practice, limited on a sample of assets, whose performances are observed over a given period of time. Hence, it is of interest wondering if, given a performance measure M, the observed stability has occurred by chance or it is due to a real predictive ability of the considered performance measure.

Following the notation introduced in the previous section, this goal may be pursued by performing a statistical test about the hypothesis H_0 : "*M is unstable*".

With respect to the statistic $I(M)$, defined in (3), the null hypothesis of complete absence of stability occurs when the $P-1$ $n-$tuples $R_i^{(p)}$, comprised in $I(M)$, are chosen independently and uniformly within the set S_n of all the possible permutations of the integers $1, 2, \ldots, n$.

Hence, capitalizing on results of Diaconis and Graham (1978), it is easy to show that, under the null hypothesis:

$$\begin{aligned} E(I(M)) &= 1 - \frac{\frac{(n-1)(n+1)}{3}}{\lfloor \frac{n^2}{2} \rfloor} \\ &= \begin{cases} \frac{1}{3} & \text{for odds } n \\ \\ \frac{1}{3} - \frac{2}{3n^2} & \text{for evens } n. \end{cases} \end{aligned}$$

and

$$\begin{aligned} Var(I(M)) &= \frac{1}{P-1} \frac{\frac{(n+1)(2n^2+7)}{45}}{\lfloor \frac{n^2}{2} \rfloor^2} \\ &= \begin{cases} \frac{1}{3}\frac{1}{P-1}\frac{4}{45}\frac{(2n^2+7)}{(n-1)(n^2-1)} & \text{for odds } n \\ \\ \frac{1}{P-1}\frac{4}{45}\frac{(2n^2+7)(n+1)}{n^4} & \text{for evens } n. \end{cases} \end{aligned}$$

Therefore, when $n \to \infty$, we have

$$\begin{aligned} E\left[\sqrt{n}(I(M) - 1/3)\right] &= 0 \\ Var\left[\sqrt{n}(I(M) - 1/3)\right] &= \frac{1}{P-1} \cdot \frac{8}{45} \end{aligned}$$

As a consequence, the hypothesis system to test can be formalized as follows:

$$\begin{cases} H_0 : I(M) \leq \frac{1}{3} \\ H_1 : I(M) > \frac{1}{3} \end{cases}$$

Two main procedures can be applied to test the null hypothesis:

1. An asymptotic test may be derived by taking advantage of the results due to Diaconis and Graham (1978) proving the Normality of the Spearman's footrule for large sample sizes. Being the (3) the one's complement of a weighted sum of Spearman's footrules, it follows that also $I(M)$ is asymptotically Normal. Hence, a standard z test may allow us to evaluate if a performance measure is significantly stable.

2. Alternatively, it would be possible to perform an exact test, by the means of permutations. Under the null hypothesis, indeed, the rankings are chosen independently and uniformly in the set S_n of all the possible permutations, thus being exchangeable at each period. The distribution of $I(M)$ may be then ideally obtained by estimating its value after permuting ranks at random at each period, for every possible permutation of ranks. However, being the cardinality of S_n equal to $(P-1)\cdot n!$, this is in practice unfeasible. Nonetheless, a Montecarlo permutation test may be performed, consisting in obtaining the approximated null distribution of I by the random selection of a large subset of all the possible permutations.

The two procedures turn out to be equivalent also for small sample sizes. The Jarque-Bera test accepts, at the 5% significance level, the hypothesis that distribution of the statistic $I(M)$, obtained through 10000 random permutations of the ranks, is Gaussian. This occurs, for example, for $n = 50, 100, 1000$ and for the value $P = 5, 10, 20$.

4 An application to mutual funds

In this section we analyze and compare the stability of 17 performance measures that are among the most commonly indexes used to evaluate the results of managed portfolios. In our analysis the measures are applied to 650 US equity mutual funds observed, at a monthly frequency, in the period from June 2001 to March 2010 (105 months).

4.1 The performance measures

Dozens of performance measures have been proposed in the literature in the last years, each one with specific characteristics, advantages and drawbacks. Here, we consider, for convenience, 17 measures, classified in five classes according to the approach used to evaluate the risk. Clearly, other choices of the selected performance measures would have been possible, as well as other criteria of classification. However, the adopted choices are, as far as we know, among the most commonly considered in the financial literature (see, for instance, Eling 2008 or Caporin and Lisi, 2009). In the following, the performance indexes included in the analysis are listed (see Table 1 for further details).

- Traditional measures: the Sharpe ratio (denoted, in the following as M_1), and some indexes derived from the Capital Asset Pricing Model (CAPM), namely the Treynor index (M_2), the Appraisal ratio (M_3) and the Jensen's Alpha (M_4). Being the intercept of the security market line, the Jensen's Alpha is not properly consistent with the definition of performance measure, but it has been included in the analysis because of its widespread employment in evaluating the past behavior of financial assets.

- Measures based on the Drawdown: the considered indexes belonging to this class are the Sterling ratio (M_5) and the Burke ratio (M_6).

- Measures based on partial moments: some performance indexes defining the risk by negative deviations of the realized returns have been taken into account in the present analysis. In particular, we have considered the Sortino ratio (M_7), the Kappa 3 index (M_8), the Upside Potential ratio (M_9), the Omega index (M_{10}) and two variants introduced by Farinelli e Tibiletti that compare an upper partial moment of order p with a lower partial moment of order q. The choice of (p, q) is associated to investors' styles or preferences. We have considered the two following parametrizations: ($p = 0.5$, $q = 2$, M_{11}) for a defensive investor and ($p = 3$, $q = 0.5$, M_{12}) for an aggressive strategy.
- Measures based on the Value at risk: performance indexes belonging to this class consider as a measure of the asset's risk the possible loss (or expected loss) which is not exceeded with a given probability $1 - \alpha$. The Var ratio (Caporin and Lisi, 2009) with parameters $(\alpha, 1-\alpha)$ and the Rachev ratio (α) have been considered in the analysis, with α set to 0.05 and 0.01). In the following, we refer to these measures respectively as M_{13}, M_{14}, M_{15}, M_{16}.
- The last measure considered in the analysis is an approximation of the Morningstar risk adjusted return measure (MRAR, M_{17})[1], whose definition is based on the utility theory. The λ parameter, governing the investor's level of risk aversion, has been set to 2 consistently with the Morningstar choice.

4.2 Analysis of the stability

In this section, we show the results deriving from an analysis of the stability of the performance measures previously described. The application was accomplished on a set of equity mutual funds belonging to a specific style-based category, namely the Large blend funds as defined by Morningstar. As it is known, Large blend funds tend to invest across the spectrum of U.S. industries and have portfolios that are fairly representative of the overall stock market in both size, growth rates, and price.

Starting from the monthly time series of the NAVs, we computed the excess returns on the basis of the monthly risk free investments[2]. The availability of the considered data led to have at our disposal 105 monthly returns of 630 mutual funds (observed from June 2001 to March 2010). Funds younger than 105 months and no more existing funds at the time of data collection were excluded from the analysis, in order to handle an unchanging sample of assets over the considered period. Thus, in principle, the resulting dataset suffers of uncompleteness and survivorship bias (see, e.g. Elton *et al.*, 1996a). However, focusing the interest on the performance measures (and not on the funds), the choice does not have any relevant impact on the results of the analysis.

In order to evaluate the stability, each performance measure $M_j (j = 1, \dots, 17)$ was estimated on P adjacent time windows having width $l = 7, 15, 21,$ and 35

[1] The exact computation of the Morningstar risk-adjusted return measure has been prevented by the lack of historical data of front loads, deferred loads, or redemption fees to compute the load adjusted returns. Hence returns are not adjusted for the impact of sales loads.

[2] Data have been drawn from the Treasury Constant Maturity Rates, source: http://www.treas.gov/offices/domestic-fince/debt-magement/interest-rate/index.html

Sharpe ratio	M_1	$\frac{\overline{r}-r_f}{\sigma(r_t)}$	r_f is the risk-free rate, $\overline{r}$ and $\sigma(r_t)$ are the mean and the standard deviation of r_t
Treynor index	M_2	$\frac{\overline{r}-r_f}{\beta}$	β is the estimated coefficient of $(r_t - r_{f,t}) = \alpha + \beta(r_{B,t} - r_{f,t}) + \varepsilon_t$
Appraisal ratio	M_3	$\frac{\alpha}{\sigma_\varepsilon}$	α and σ_ε are the estimated intercept and the standard deviation of the residuals of the CAPM line, as defined above.
Jensen's Alpha	M_4	α	
Sterling ratio	M_5	$\frac{\overline{r}-r_f}{\sum_{s=1}^{S}\frac{-MD_s}{S}}$	MD_s is the denotes the s^{th} lowest return; S has been set to the nearest integer to $l/10$
Burke ratio	M_6	$\frac{\overline{r}-r_f}{\sqrt{\sum_{s=1}^{S}\frac{MD_s^2}{S}}}$	
Sortino ratio	M_7	$\frac{\overline{r}-\tau}{\sqrt{LPM_2(\tau)}}$	τ is the minimum acceptable return, set to 0; $LPM_k = \sum_{t=1}^{l}\frac{max(\tau-r_t,0)^k}{l}$
Kappa 3 index	M_8	$\frac{\overline{r}-\tau}{\sqrt[3]{LPM_3(\tau)}}$	
Upside potential ratio	M_9	$\frac{HPM_1(\tau)}{\sqrt{LPM_2(\tau)}}$	$HPM_k = \sum_{t=1}^{l}\frac{max(r_t-\tau,0)^k}{l}$
Omega index	M_{10}	$\frac{\overline{r}-\tau}{LPM_1(\tau)} + 1$	
Farinelli-Tibiletti index	M_{11}, M_{12}	$\frac{\left[\frac{\sum_{t=1}^{l} max(r_t-\tau,0)^p}{l}\right]^{\frac{1}{p}}}{\left[\frac{\sum_{t=1}^{l} max(tau-r_t,0)^q}{l}\right]^{\frac{1}{q}}}$	$\alpha = 0.05, 0.01$
VaR ratio	M_{13}, M_{14}	$\frac{\lvert Var(-r_t,\alpha)\rvert}{\lvert Var(r_t,\alpha)\rvert}$	$-VaR(r_t,\alpha)$ is the $\alpha-$quantile ($\alpha = 0.05, 0.01$) of the distribution of r_t
Rachev ratio	M_{15}, M_{16}	$\frac{\sum_t r_t \mathbf{1}_{[r_t \geq -Var(-r_t,\alpha)]}(r_t)}{\sum_t r_t \mathbf{1}_{[r_t \leq Var(r_t,\alpha)]}(r_t)}$	
Pseudo-MRAR(λ)	M_{17}	$\left[\frac{\sum_{t=1}^{l}(1+r'_t)^{-\lambda}}{l}\right]^{-\frac{12}{\lambda}} - 1$	$r'_t = \frac{1+r_t}{1+r_f} - 1$

Table 1: Performance indexes included in the analysis of stability. r_t denotes the return of the considered asset at time t, $r_{f,t}$ is the risk free rate, $r_{B,t}$ is the return of the market index.

months, hence resulting in $P = 15, 7, 5,$ and 3 periods, respectively. The time windows widths could seem unusual and, in fact, they are. Undeniably, the choice does not correspond to any specific financial meaning and reflects, instead, the need to guarantee the comparability of results which can be obtained only if the performance are evaluated exactly on the same global period. Nonetheless, the considered

	P=3, l=35	P=5, l=21	P=7, l=15	P=15, l=7
M_1	0.419 (<0.001)	0.461 (<0.001)	0.484 (<0.001)	0.393 (<0.001)
M_2	0.467 (<0.001)	0.454 (<0.001)	0.511 (<0.001)	0.398 (<0.001)
M_3	0.463 (<0.001)	0.509 (<0.001)	0.523 (<0.001)	0.483 (<0.001)
M_4	0.463 (<0.001)	0.460 (<0.001)	0.509 (<0.001)	0.399 (<0.001)
M_5	0.424 (<0.001)	0.472 (<0.001)	0.419 (<0.001)	0.391 (<0.001)
M_6	0.387 (<0.001)	0.461 (<0.001)	0.444 (<0.001)	0.372 (<0.001)
M_7	0.435 (<0.001)	0.471 (<0.001)	0.493 (<0.001)	0.379 (<0.001)
M_8	0.431 (<0.001)	0.473 (<0.001)	0.491 (<0.001)	0.380 (<0.001)
M_9	0.452 (<0.001)	0.452 (<0.001)	0.474 (<0.001)	0.376 (<0.001)
M_{10}	0.423 (<0.001)	0.457 (<0.001)	0.468 (<0.001)	0.386 (<0.001)
M_{11}	0.424 (<0.001)	0.450 (<0.001)	0.453 (<0.001)	0.379 (<0.001)
M_{12}	0.351 (0.067)	0.406 (<0.001)	0.408 (<0.001)	0.370 (<0.001)
M_{13}	0.388 (<0.001)	0.396 (<0.001)	0.382 (<0.001)	0.357 (<0.001)
M_{14}	0.365 (0.004)	0.376 (<0.001)	0.379 (<0.001)	0.348 (0.001)
M_{15}	0.374 (<0.001)	0.337 (0.320)	0.372 (<0.001)	0.348 (<0.001)
M_{16}	0.346 (0.154)	0.356 (0.003)	0.372 (<0.001)	0.348 (<0.001)
M_{17}	0.454 (<0.001)	0.484 (<0.001)	0.419 (<0.001)	0.437 (<0.001)

Table 2: Stability estimates of performance measures for different pairs of (P, l) and corresponding p-values in brackets. Time series returns referred to 105 months have been used.

window widths can be deemed as approximations of periods more often used to evaluate the performance of a managed portfolios. Indeed, the same analysis was conducted with respect to periods of length 6, 12, 24 and 36 months (without satisfying the constrain of using the same global period to conduct the analysis) and results were basically equivalent.

For each choice of l and P, the stability index I was computed for all the measures $M_j, j = 1, ..., 17$. Descriptive results are listed in Table 2, while Figure 1 displays the boxplots of the distributions of $I(M_j)$. The distributions were obtained by following a bootstrap-style approach, consisting of the following steps: first, a subset of mutual funds was sampled without replacement from the data (the subset size was set to the 2/3 of the original sample size, namely to 420 funds); the performance measures were then estimated on the selected subsample for each time window $p = 1, \ldots, P$ and the stability index computed. The selection of a subsample of funds, (instead of a standard bootstrap sample, drawn with replacement from the data and having size $n = 630$) is motivated by the need of avoiding ties in the ranks, almost certainly occurring in case of selection with replacement. The empirical distributions of the stability index were then obtained by repeatedly iterating the described procedure for each performance measure.

The level of stability of the considered measures is not particularly high, ranging between 0.337 and 0.523. The more robust the risk measure used to adjust the return level is, the stabler the performance index tends to be, and the performance indexes using measures of risk based on the same approach, generally cluster also according to their level of stability. Thus, indexes hinging on the Value at risk and on the Drawdown are the least stable. Performance indexes based on a measure of risk related to an estimate of a partial moment get an average stability while indexes derived from the CAPM are generally the stablest one. However, this behavior also depends on the length of the periods where measures are estimated. For instance, among the performance measures computed on periods with length 21 months, the Sterling ratio and the Burke ratio show an average level of stability similar to other measures based on the CAPM or the partial moments.

Interestingly, from our analysis the measure that performs better in terms of stability turns out to be the Appraisal ratio, whatever time window width is chosen to compute I. This result is consistent with findings of Ornelas *et al.* (2009) and Zakamouline (2010).

The Farinelli and Tibiletti ratio with parameters $p = 3, q = 0.5$ shows high instability with respect to the other measures based on partial moments; indeed, this behavior conforms with what has been already noted, since that parametrization is associated to an aggressive style of investment, corresponding to an high propensity to the risk. The pseudo-Morningstar risk-adjusted return generally shows an relative high level of stability, leaving most of the other measures standing. However, this trend does not occur for all the considered time windows (when $l = 15$ months the stability of MRAR is quite low).

The analysis also enlightens some interesting issues concerning the best choice of the window width to make the measures stabler, and hence more predictive. Except for a few inversions, all the measures show larger stability when they are computed over periods having length set to 15 months. Also, relatively large values of stability result from the use of $l = 21$, while the association between rankings over the time is remarkably lower when the window widths are shorter than one year ($l = 7$) or longer than two years ($l = 35$). Hence, it turns out that measures perform more stably in the short-medium term.

Regarding the distribution of the stability, the symmetric shape of the boxplots indicates that positive and negative departures from the median stability having the same size are equally likely to occur. Moreover, for a given time window width, the stability of the performance measures shows comparable levels of variability around the median, thus suggesting that distributions of the stability of distinct performance measures differ for the location only. A further feature of these distributions, valuable to be noted, concerns how the distributions vary when different time window widths and number of periods are considered: the average level of stability tends to follow a bell-shaped curve, firstly increasing with the window width, reaching a maximum value when the index is computed over 5 and 7 periods, and then decreasing again for a larger window width and a smaller number of periods; as expected, instead, the variability of the stability index decreases with the number of periods considered.

As far as the significance of the stability indexes is concerned, the $p-$values associated to the application of the permutation test for the null hypothesis of *instability*, described in the previous section, are reported, in brackets, in Table 2 (the use of a z test yields to the same conclusions). Generally the stability of the performance measures tends to be significant, although with some exceptions. Again, the length of the period where measures were computed seems to be relevant, suggesting that there could be an 'optimal' window width for forecasting purposes. Indeed, when few periods (large window widths) are considered, the stability of performance indexes related to some extreme measure of risk is compromised at a significance level of 5%. Coherently with previous results, the window width that produces the largest departure from instability is $l = 15$, with the most significant level of stability given by the Appraisal ratio.

This behavior is even more evident in Figure 2, where the null distributions of the test statistic are plotted for the different choices of l and P and the observed statistics are reported on the histograms. It can be noted that, as far as the variability of the statistic increases for decreasing values of P, the observed statistics tend to be more likely under the null hypothesis of instability.

5 Concluding remarks

In this work, the attention has been focused on discussing the role of the stability of a performance measure over time. An index of stability has been proposed and its properties analyzed. Furthermore, we have conjectured that different measures might have different intrinsic degree of stability. The empirical analysis has shown evidence that this conjecture was right. It has turned out that the stability of the performance measures, evaluated by using the proposed criterion, varies across different time horizons and, in general, it departs significantly from its expected value under the hypothesis of instability. Even if the results deriving from our analysis cannot undisguisedly govern the choice of the performance measure, some useful remarks may be highlighted. Remarkable differences have resulted among the stabilities of different performances indexes, and the analysis has pointed out that the more robust the approach used to evaluate the risk is, the more stable the performance measure is.

Clearly, the choice of the performance measure should depend on several considerations, such as the investors'aptitude to the return-risk level and also other properties should be taken into account. For instance, the considered benchmark-based performance measures (namely the Appraisal ratio, the Treynor index and the Jensen's alpha) in our analysis have been estimated by considering the use of a single index market model. Instead, it is known that to capture all the systematic sources of excess returns it would be more realistic the use of a multi-factor index model. However, even if the choice of the performance index cannot be solely driven by considerations about the stability, this work has strengthened the idea that such choice is not inconsequential and, *ceteris paribus*, taking into account the property of stability, may help in providing more reliable results in the analysis of performance persistence.

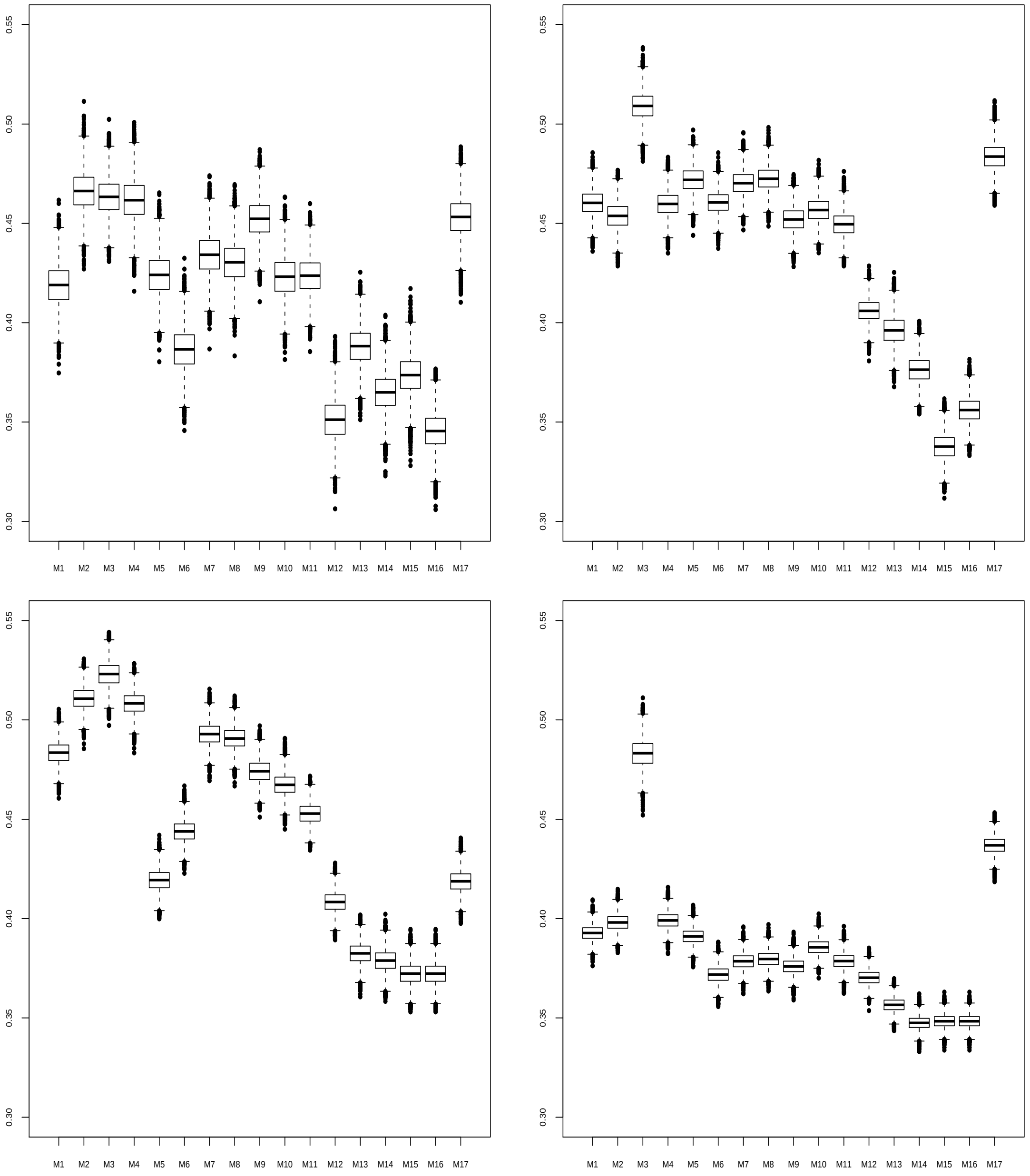

Figure 1: From the left to the right and from the top to the bottom the boxplots of the pseudo-bootstrap distributions of the stabilities are reported, with $P = 3, 5, 7, 15$ periods and $l = 35, 21, 15, 7$ months respectively. Each boxplot depicts a 5$-$numbers summary of the distribution of a performance index: the limits of the box and the thick solid line represent the first quartile, the third quartile and the median of the distribution, respectively. The the plot 'whiskers' extend out from each box to the most extreme data points which are away from the box no more than 1.5 times the interquartile distance. The remaining points are considered as outliers.

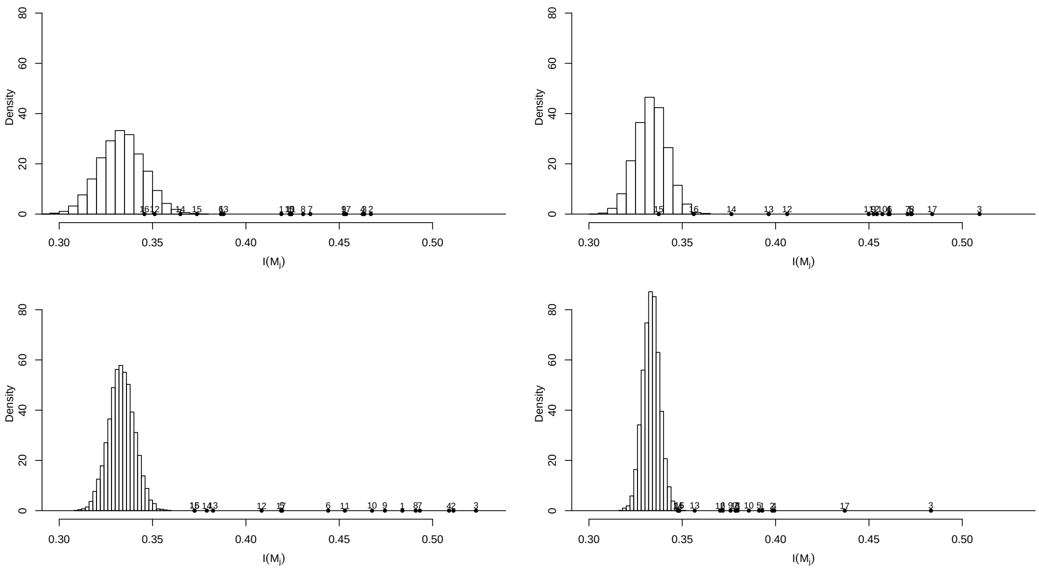

Figure 2: From the left to the right and from the top to the bottom: null distributions of $I(M_j)$ for window width l=35,21,15,7 (and P=3,5,7,15). On the x axis the observed statistics referred to the considered performance measures M_j, are reported and denoted by j, $j = 1, \ldots, 17$.

References

[1] Aftalion, F., and Poncet, P. (2003). Les techniques de mesure de performance. Paris: Economica.

[2] Agarwal, V. and Naik, N. Y. (2000). Multi-Period Performance Persistence Analysis of Hedge Funds. *The Journal of Financial and Quantitative Analysis*, 35(3), 327-342.

[3] Beckers, S. and Thomas, J. A. (2010). On the persistence of style returns. *Journal of Portoflios Management*, 37(1), 15-30.

[4] Bodson, L., Coën, A., Hübner G. (2008). How stable are the major performance measures?. *The Journal of Performance Measurement*, 13(1), 21-30.

[5] Bollen, N. P. B., and Busse, J. A. (2005). Short-term persistence in mutual fund performance. *Review of Financial Studies*, 18(2), 569-597.

[6] Broihanne, M. E., Merli, M. and Hübner, G. (2008), On the robustness of mutual funds ranking with an index of relative efficiency. *Banque & Marchés*, 94, 32-42.

[7] Brown, S. J. and Goetzmann, W. N. (1995). Performance Persistence. *The Journal of Finance*, 50(2), 679-698.

[8] Brown, S. J., Goetzmann, W. N. and Ibbotson, R. G. (1999). Offshore Hedge Funds: Survival and Performance 1989–95. *Journal of Business*, 72, 91-117.

[9] Caporin, M. and Lisi, F. (2009), Comparing and Selecting Performance Measures for Ranking Assets. Available at SSRN: http://ssrn.com/abstract=1393163.

[10] Carhart, M., (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 5782.

[11] Diaconis, P. and Graham, R. L. (1978). Spearman's Footrule as a measure of Disarray. *Journal of the Royal Statistical Society (B)*, 39, 262-268.

[12] Eling, M. (2008). Does the Measure Matter in the Mutual Fund Industry?. *Financial Analysts Journal*, 64(3), 54-66.

[13] Eling, M., Schuhmacher, F. (2007). Does the Choice of Performance Measure Influence the Evaluation of Hedge Funds? *Journal of Banking and Finance*, 31, 2632-2647.

[14] Genest, C., Nelehov, J. and Ben Ghorbal, N. (2010). Spearman's footrule and Gini's gamma: A review with complements. *Journal of Nonparametric Statistics*, 22, 937-954.

[15] Elton, E. J., Gruber M. J. and Blake, C. R. (1996a). Survivorship bias and mutual fund performance. *The Review of Financial Studies*, 9(4), 1097-1120.

[16] Elton, E. J., Gruber M. J. and Blake, C. R. (1996b). The Persistence of Risk-Adjusted Mutual Fund Performance. *Journal of Business*, 69(2), 133-157.

[17] Elton, E. J., Gruber M. J., Das S., and Hlavka, M. (1993). Efficiency with Costly Information: A Reinterpretation of Evidence from Managed Portfolios. *Review of Financial Studies*, 6(1), 1-22.

[18] Goetzmann, W. N. and Ibbotson, R. G. (1994). Do Winners Repeat? Patterns in Mutual Fund Behavior. *Journal of Portfolio Management*, 20(2), 9-18.

[19] Grinblatt, M. and Titman, S. (1992). The Persistence of Mutual Fund Performance. *Journal of Finance*, 47(5), 1977-1984.

[20] Hendricks, D., Jayendu P. and Zeckhauser, R. (1993), Hot Hands in Mutual Funds: Short Run Persistence of Relative Performance, 1974-1988. *Journal of Finance*, 48(1), 93-130.

[21] Hübner G. (2007). How Do Performance Measures Perform?.*The Journal of Portfolio Management*, 33, 64-74.

[22] Kendall, M. G. (1970). Rank Correlation Methods, Griffin.

[23] Le Sourd, V. (2007). Performance Measurement for Traditional Investment. *Financial Analysts Journal*, 58(4), 36-52.

[24] Ornelas, J.R., de Almeida Silva, A.F., Fernandes J.L. (2009). Yes, The Choice of Performance Measure Does Matter For Ranking of US Mutual Funds. Available at at SSRN: http://ssrn.com/abstract=1403916.

[25] Park, J. M. and Staum, J. C. (1998). Performance Persistence in the Alternative Investment Industry. *Technical report*, PARADIGM Capital Management.

[26] Zakamoulilne V. (2010). The Choice of Performance Measure Does Influence the Evaluation of Hedge Funds. Available at SSRN: http://ssrn.com/abstract=1403246.