

# Comparing and Selecting Performance Measures Using Rank Correlations

*Massimiliano Caporin*

*Department of Economics and Management "Marco Fanno", University of Padova*

*Francesco Lisi*

*Department of Statistical Sciences, University of Padova*

**Abstract** The financial economics literature proposes dozens of performance measures to be used, for instance, to compare, analyse, rank and select assets. There is thus a problem: which measures should be considered? We extend the current literature by comparing a large set of performance measures over more than one thousand of equities included in the Standard & Poor's 1500 index. We evaluate performance measures by mean of rank correlations, exploiting the possible dynamic evolution of the rank correlations, and proposing a method for the identification of the subset of measures which are not equivalent. Our empirical study highlights that recent and more flexible measures provide different asset ranks compared to classical approaches, and that the set of equivalent performance measures is not stable over time.

**JEL** C10, G11, C40

**Keywords** Performance measurement; rank correlations; comparing performance measures

**Correspondence** Massimiliano Caporin, Dipartimento di Scienze Economiche "Marco Fanno", Via del Santo, 33, 35123, Padova, Italy; e-mail: [massimiliano.caporin@unipd.it](mailto:massimiliano.caporin@unipd.it)

**Citation** Massimiliano Caporin and Francesco Lisi (2011). Comparing and Selecting Performance Measures Using Rank Correlations. *Economics: The Open-Access, Open-Assessment E-Journal*, Vol. 5, 2011-10. doi:10.5018/economics-ejournal.ja.2011-10. <http://dx.doi.org/10.5018/economics-ejournal.ja.2011-10>

## 1 Introduction

Since the pioneering works of Sharpe (1964 and 1966) and Treynor (1965), the topic of performance measurement has attracted considerable interest in the financial economic literature. From a general viewpoint, we may identify, among others, two fundamental topics covered by performance measurement. The first considers the returns of financial assets, and aims to define and interpret ratios or indices, the performance measures or reward-to-risk ratios, for the purpose of determining the assets' risk/return trade-off. The second analyses returns of managed portfolios and focuses on the introduction and use of models and approaches which make possible to infer the choices made by investment managers. For examples on the second topic see Knight and Satchell (2002) and the references therein, the literature on style analysis (see Sharpe, 1992, among others) and the contributions related to conditional CAPM approaches, including Ferson and Schadt (1996), Avramov and Chordia (2006).

This study deals with the first issue. We focus on the comparison of performance measures based on the returns of specific assets. The approaches proposed by this strand of the literature may be considered as tools for portfolio managers and agents facing investment decisions. Performance measures are here used as tools for selecting a relatively small number of assets with given features (such as small drawdowns or high return...) for a subsequent allocation possibly using a generalization of the Markowitz approach. Alternatively, performance measures may be used to select a number of assets for the direct application of naïve portfolio allocation rules, such as the equally weighted one (see De Miguel et al., 2009). The financial economics literature proposes also to use performance measures as objective functions for determining the weights of an optimal portfolio. We will not pursue this objective, but for an example of such an approach see Farinelli et al. (2008, 2009).

A relevant point is still open and has recently attracted some interest: which performance measure should be used? In fact, many reward-to-risk ratios have been proposed. Besides the well-known Sharpe, Sortino and Treynor indices, a number of alternative measures are available, such as the Omega index (Shadwick and Keating, 2002), the Rachev ratio (Rachev et al., 2003), and the *FT* ratios (Farinelli and Tibiletti, 2003), among others. Their number is increasing over time,

and new indices are designed to meet specific requirements, for example Pedersen and Rudholm-Alfvén (2003), or with the purpose of overcoming the limitations of the oldest measures. Some examples are given by the need of increasing the robustness of performance measures with respect to deviations from normality, or of introducing measures more appropriate for agents characterized by loss aversion (Gemmill et al., 2006) or by aggressiveness (Farinelli and Tibiletti, 2003).

The comparison of alternative performances, generally using rank correlations, have already been considered. In particular, we refer to Gemmill et al. (2006), Eling and Schuhmacher (2007), Eling (2008), and Eling et al. (2011). These contributions use a simple and effective approach for deciding which measures to use: in order to compare alternative indices, they verify whether they rank assets differently. Performance measures providing equivalent rankings are redundant and may thus be discarded. Following this method, we may identify a restricted set of performance measures carrying different information on the risk/return trade-off.

In this work we follow the empirical approach of Eling and Schuhmacher (2007) and provide three main contributions. The first one extends and completes the cited paper by broadening the set of performance measures to be compared. In particular, we include performance measures based on partial moments (Farinelli and Tibiletti (2003) and Rachev et al. (2003), as in Eling et al. (2011), and on loss aversions (Gemmill et al., 2006). In addition, we base our analysis on equities, rather than on managed portfolios as in Eling and Schuhmacher (2007), Eling (2008), and Eling et al. (2011). With respect to these issues, and differently from Eling and Schuhmacher (2007), we find cases of low rank correlation across performance measures, and then we argue that the equivalence relations may depend on the kind of assets considered and on the sample period. We also introduce four new performance measures: the expected return over range, where the risk measure is given by the maximum range; the VaR ratio, which is the ratio of the upper and lower quantiles of a given return distribution; and two performance measures derived from a utility function with loss aversion.

The second contribution is associated with a different topic: the stability over time in the rankings induced by different performance measures. We will try to answer the question: "Are rank correlations time-varying?" To that end, we compare the rank correlations computed both over samples of different length, and over rolling windows. Our analysis extends the studies of Eling and Schuhmacher

(2007), and Eling (2008) that did not consider the rolling approach but evaluate the rank correlations on the full sample and on a two or five years sample. We show that, for our data, the rank correlations are not time invariant and are influenced by the sample size. Therefore, on the one side, appropriate tools for comparing and selecting performance measures are needed, while, on the other side, these dynamics could be exploited within an asset management framework.

Building on this new evidence, for the third contribution of this work, we tackle the topic of the redundancy of the performance measures in a dynamic context. Given a set of  $N$  performances measures, we propose a way to reduce them in order to consider only those which really carry different information. In our empirical study we start, in the most general case, with 80 measures and, using a procedure based on the asymptotic distribution of the rank correlation coefficient, we conclude that 57 measures are redundant since they carry information similar to the 23 we select. In connection with the second outcome of this paper, we also infer that the set of performance measures carrying relevant information may be time-varying as well. This additional piece of information could be proven to be extremely relevant for periodic rotation or rebalancing of managed portfolios using asset screens.

Given that the allocation choices of portfolio managers and agents are generally taken at a low frequency (monthly to quarterly) in this paper we work with monthly data, but analysis at different frequencies may be considered. Moreover, we assume that the series of interest are characterized by deviations from normality (which, for equities, is one of the well-known stylized facts, see Cont, 2001, among others), and that the risk and reward measures presented below are estimated with their sample counterparts without introducing a parametric model.

The rest of the paper is organized as follows. Section 2 lists the performance measures that will be considered, describes the dataset, and discusses some problems connected to the selection bias. In Section 3 we report the results of the analysis concerning the correlations between different performance measures and we show how to obtain the set of the measures that are significantly different. Our final conclusions are presented in Section 4.

## 2 Performance Measures List and Dataset Description

From a general viewpoint, performance measures can be defined as ratios between a reward measure and a risk measure, and their value can be interpreted as the reward per unit of risk. Despite a general agreement on what a performance measure is, a number of choices are available for the reward and risk measures to be considered, as well as for the type of variables to be used for their evaluation.

In order to provide a general setup, we start by introducing some notation: we denote by  $R_{i,t}$  the (nominal) log-return of asset  $i$  in period  $t$ ;  $R_{f,t}$  is the risk-free investment return (it is time-varying since we consider it as a pure risk-free investment within each period);  $R_{B,t}$  identifies the return of a benchmark investment;  $X_{t=1}^T$  is the sequence of observations of the variable  $X_t$  from time 1 to time  $T$ ;  $E[X^p]$  is the moment of order  $p$  of  $X$ ;  $E[g(X)^p]$  is the moment of order  $p$  of the function  $g(X)$ ;  $\sigma[X]$  is the volatility of  $X$ ; and,  $E[X^p|Y]$  is the conditional moment of order  $p$  of  $X$ .

The performance measures presented below will be defined over a variable  $X_{i,t}$  that takes one of the following values

$$X_{i,t} = \begin{cases} R_{i,t} \\ R_{i,t} - R_{f,t} \\ R_{i,t} - R_{B,t} \end{cases} . \quad (1)$$

These cases represent three possible relevant dimensions for performance measurement, not necessarily mutually exclusive: nominal returns (relevant for agents focusing on purely risk investments), excess returns with respect to a risk-free returns (for investors considering also a risk-free investment), deviations from a benchmark (relevant within an active management framework). We now describe the performance measures we consider, grouping them from a statistical point of view (thus separating the use of general risk measures, from ratios based on partial moments and quantiles, and those derived from utility functions). Of course, different and more detailed classifications could have been used, Aftalion and Poncet (2003), Le Sourd (2007), and Cogneau and Hubner (2009a, 2009b). Nevertheless, we prefer to maintain a limited and simple structure, and our selection of performance measures includes quantities designed to capture deviations from normality as well as to take into account agents' behaviour.

## 2.1 Traditional Performance Measures and Other Unclassified Measures

This first set of performance measures contains the most known and traditional indices:

- the Sharpe ratio, introduced by Sharpe (1966, 1994):

$$Sh(X_{i,t}) = \frac{E[X_{i,t}]}{\sigma[X_{i,t}]}, \quad (2)$$

- the Treynor index, Treynor (1965), defined for nominal returns and excess returns only:

$$Tr(X_{i,t}) = \frac{E[X_{i,t}]}{\beta_i}, \quad (3)$$

where  $\beta_i$  is estimated through a CAPM regression;

- the Appraisal ratio, defined as:

$$AR(X_{i,t}) = \frac{\alpha_i}{\sigma[\varepsilon_{i,t}]}, \quad (4)$$

where  $\alpha_i$  is the intercept of a CAPM regression and  $\sigma[\varepsilon_{i,t}]$  denotes the volatility of the CAPM residuals;

- the expected return over Mean Absolute Deviation ratio of Konno and Yamazaki (1991):

$$ERMAD(X_{i,t}) = \frac{E[X_{i,t}]}{E[|X_{i,t} - E[X_{i,t}]|]}. \quad (5)$$

We also include here some performance measures which are not consistent with the following groups and are defined as ratios between the first order moment of  $X_{i,t}$  and a risk measure:

- the return over MiniMax ration of Young (1998):

$$ERMM(X_{i,t}) = \frac{E[X_{i,t}]}{\max(\max X_{t=1}^T, -\min X_{t=1}^T)}; \quad (6)$$

- the expected return over the range ratio, which, to our knowledge has never been considered in previous studies:

$$ERR(X_{i,t}) = \frac{E[X_{i,t}]}{\max X_{t=1}^T - \min X_{t=1}^T}. \quad (7)$$

Finally, we include here also the Risk Adjusted Performance (RAP), or M2 index of Modigliani and Modigliani (1997):

$$M2 = (E[R_{i,t}] - E[R_{B,t}]) \frac{\sigma[R_{B,t}]}{\sigma[R_{i,t}]} + E[R_{f,t}] - E[R_{B,t}]. \quad (8)$$

## 2.2 Measures Based on Drawdown

This set contains measures based on risk indices focusing on the drawdown, which is define as

$$D_t(X_{i,t}) = \min(D_{t-1} + X_{i,t}, 0) \quad D_0 = 0. \quad (9)$$

Given the observations for  $X_{i,t}$   $t = 1, \dots, T$ , the drawdown  $D_t(X_{i,t})$  or simply  $D_t$  represents, at time  $t$ , the maximum loss an investor may have suffered from 1 to  $t$ . Risk measures are defined ordering drawdowns and computing quantities such as the maximum drawdown,  $D_1(X_{i,t}) = \min D_{t=1}^T$ , or the second largest drawdown  $D_2(X_{i,t}) = \min(D_{t=1}^T - D_1(X_{i,t}))$ , and so on. We also assume  $D_1(X_{i,t}) < 0$ . We consider three indices based on drawdowns:

– the Calmar ratio of Young (1991):

$$CR(X_{i,t}) = \frac{E[X_{i,t}]}{-D_1(X_{i,t})}; \quad (10)$$

– the Sterling ratio, introduced by Kestner (1996):

$$SR(X_{i,t}; w) = \frac{E[X_{i,t}]}{-\frac{1}{w} \sum_{j=1}^w D_j(X_{i,t})}; \quad (11)$$

where  $w$  is a parameter that identifies the number of values used in the computation of the risk index;

– the Burke ratio, due to Burke (1994):

$$BR(X_{i,t}; w) = \frac{E[X_{i,t}]}{\left(\frac{1}{w} \sum_{j=1}^w [D_j(X_{i,t})]^2\right)^{\frac{1}{2}}}. \quad (12)$$

In the Burke and Sterling ratios, Eling and Schuhmacher (2007) fix the value of  $w$  between 1 and 10. Differently, we prefer linking the number of drawdowns to the sample dimension as  $w = \left\{ \left\lceil \frac{T}{20} \right\rceil, \left\lceil \frac{T}{10} \right\rceil \right\}$  where  $\lceil a \rceil$  denotes the nearest integer of  $a$ .

### 2.3 Measures Based on Partial Moments

We also analyze performance measures based on partial moments: - the Sortino ratio, Sortino and Van der Meer (1991):

$$Sr(X_{i,t}) = \frac{E[X_{i,t}]}{E\left[(\min(X_{i,t}, 0))^2\right]^{\frac{1}{2}}}; \quad (13)$$

- the Kappa 3 measure of Kaplan and Knowles (2004):

$$K3(X_{i,t}) = \frac{E[X_{i,t}]}{E\left[(\min(X_{i,t}, 0))^3\right]^{\frac{1}{3}}}. \quad (14)$$

- the Farinelli and Tibiletti (2003) ratio, or *FT* ratio:

$$FT(X_{i,t}; b, p, q) = \frac{E\left[\left((X_{i,t} - b)^+\right)^p\right]^{\frac{1}{p}}}{E\left[\left((X_{i,t} - b)^-\right)^q\right]^{\frac{1}{q}}} \quad (15)$$

where  $(X_{i,t} - b)^+ = \max(X_{i,t} - b, 0)$ ,  $(X_{i,t} - b)^- = \max(b - X_{i,t}, 0)$ . The threshold return level  $b$ , and the partial moment orders  $p$  and  $q$  are calibrated following Farinelli and Tibiletti (2003) in order to match them with possible investors' styles or preferences:  $p = 0.5$  and  $q = 2$  for a defensive investor;  $p = 1.5$  and  $q = 2$  for a conservative investor;  $p = q = 1$  for a moderate investor (note that this combination makes the  $FT(X_{i,t}; b, 1, 1)$  equivalent to the Omega index of Shadwick and Keating (2002));  $p = 2$  and  $q = 1.5$  for a growth investor;  $p = 3$  and  $q = 0.5$  for an aggressive investor; in addition,  $p = 1$  and  $q = 2$  defines the Upside Potential Ratio of Sortino et al. (1999). Finally, we consider the following cases for the threshold return,  $b = \{-0.02, 0, 0.02\}$ , where the  $-2\%$  and  $2\%$  values may represent the choices of a less risk averse and a more risk averse investor, respectively.

### 2.4 Measures Based on Quantiles

A class of performance measures similar to the previous one replaces partial moments with reward and variability measures based on quantiles (see Rachev



et al., 2003, Biglova et al., 2004, among others). At first, we define the following quantities: the Value-at-Risk at the  $\alpha$  confidence level is the quantity  $VaR(X_{i,t}; \alpha)$  such that  $P[X_{i,t} \leq VaR(X_{i,t}; \alpha)] = \alpha$ ; the Expected Shortfall  $ES(X_{i,t}; \alpha) = E[X_{i,t} | X_{i,t} \leq VaR(X_{i,t}; \alpha)]$ . We consider the following indices based on  $VaR(X_{i,t}; \alpha)$  and  $ES(X_{i,t}; \alpha)$ , with  $\alpha$  set equal to 5% or 10%:

– the Expected return over absolute VaR:

$$VR(X_{i,t}; \alpha) = \frac{E[X_{i,t}]}{|VaR(X_{i,t}; \alpha)|}; \quad (16)$$

– the VaR ratio, defined as:

$$VaRR(X_{i,t}; \alpha) = \frac{|VaR(-X_{i,t}; \alpha)|}{|VaR(X_{i,t}; \alpha)|}; \quad (17)$$

(to our best of our knowledge this index has never appeared in the literature);

– the Expected return over absolute Expected Shortfall, STARR, (Rachev et al., 2003):

$$STARR(X_{i,t}; \alpha) = \frac{E[X_{i,t}]}{|ES(X_{i,t}; \alpha)|}; \quad (18)$$

– the Generalized Rachev Ratios (Biglova et al., 2004):

$$GR(X_{i,t}; \alpha, p, q) = \frac{E[|X_{i,t}|^p | X_{i,t} \geq -VaR(-X_{i,t}; \alpha)]^{\frac{1}{p}}}{E[|X_{i,t}|^q | X_{i,t} \leq VaR(X_{i,t}; \alpha)]^{\frac{1}{q}}}, \quad (19)$$

where  $p > 0$  and  $q > 0$  are conditional moment orders calibrated as for the orders of the *FT* index. Note that the combination  $p = q = 1$  gives the simple Rachev Ratio (Biglova et al., 2004).

## 2.5 Measures Derived from Utility Functions

Some performance measures deviate from the general structure of reward-to-variability ratios. A relevant example is given by quantities derived from utility functions, and allowing the computation of risk-adjusted returns. The first we

consider is the Morningstar Risk-Adjusted Return, MRAR (Sharma, 2004, and Morningstar, 2007):

$$MRAR(X_{i,t}; \lambda) = \begin{cases} E \left[ \frac{(1 + X_{i,t})^{-\lambda}}{e^{E[\ln(1+X_{i,t})]}} \right]^{-\frac{12}{\lambda}} & \lambda > -1, \lambda \neq 0 \\ \lambda = 0 & \end{cases}, \quad (20)$$

where  $\lambda$  is a risk aversion coefficient. In the empirical part, we consider three risk aversion values: 2, 10 and 50.

Gemmill et al. (2006) introduced a set of performance measures derived within a behavioral finance framework. Following the prospect theory of Kahnemann and Tversky (1979), the utility function is replaced by a value function displaying loss aversion and focusing on gains and losses at time  $t$  with respect to the beginning of period wealth  $W_{t-1}$ . The following equation defines the value function

$$V_t(X_{i,t}) = \begin{cases} (W_{t-1}X_{i,t})^p & X_{i,t} \geq 0 \\ -\lambda (-W_{t-1}X_{i,t})^q & X_{i,t} < 0 \end{cases}, \quad (21)$$

where  $p, q$  and  $\lambda$  are positive parameters and loss-aversion is included if  $\lambda > 1$ . The wealth evolves according to  $W_t = W_{t-1} (1 + R_{i,t})$ . The Value Function in (21) displays a 'House-Money' effect, as defined in Barberis et al. (2001), if the loss aversion coefficient depends on previous gains and losses, thus becoming time varying

$$\lambda_t = \lambda_0 + \lambda_1 W_{t-2} X_{i,t-1}. \quad (22)$$

Following Gemmill et al. (2006) we define a set of performance measures accounting for loss aversion. We first rewrite the value function as

$$V_t(X_{i,t}) = (W_{t-1}X_{i,t})^p I(X_{i,t} \geq 0) - \lambda (-W_{t-1}X_{i,t})^q I(X_{i,t} < 0), \quad (23)$$

where the first component identifies gains and the second losses. The expectation of the ratio between the two quantities is a performance measure as it can be considered a reward to variability quantity. Gemmill et al. (2006) suggest as performance measures the following ratios

$$LAP^S = \frac{E[(X_{i,t})^p | X_{i,t} \geq 0] P(X_{i,t} \geq 0)}{E[(-X_{i,t})^q | X_{i,t} < 0] (1 - P(X_{i,t} \geq 0))}, \quad (24)$$

$$LAP^H = \frac{E[(X_{i,t})^p | X_{i,t} \geq 0] P(X_{i,t} \geq 0)}{\lambda_t E[(-X_{i,t})^q | X_{i,t} < 0] (1 - P(X_{i,t} \geq 0))}, \quad (25)$$

where  $P(X_{i,t} \geq 0)$  is the probability of having returns above zero. We then introduce two alternative indices that take into account the evolution of the wealth in the evaluation of performances:

$$LAP^{WS} = \frac{\sum_{t=1}^T (W_{t-1} X_{i,t})^p}{\sum_{t=1}^T I(X_{i,t} \geq 0)} \left( \frac{\sum_{t=1}^T (-W_{t-1} X_{i,t})^q}{\sum_{t=1}^T I(X_{i,t} < 0)} \right)^{-1}, \quad (26)$$

$$LAP^{WH} = \frac{\sum_{t=1}^T (W_{t-1} X_{i,t})^p}{\sum_{t=1}^T I(X_{i,t} \geq 0)} \left( \frac{\sum_{t=1}^T \lambda_t (-W_{t-1} X_{i,t})^q}{\sum_{t=1}^T I(X_{i,t} < 0)} \right)^{-1}. \quad (27)$$

In the empirical analysis, we follow Gemmill et al. (2006) and set  $\lambda$  equal to 2.25. In addition, we set  $p$  and  $q$  to 0.75 and 0.95, as in Gemmill et al. (2006), and, we also consider the combinations used for the *FT* index.

## 2.6 Dataset Description

We compare the set of performance measures reported in Table 1 over a dataset that contains the stocks included in the S&P 1500 index. The index covers from large-cap to small-cap stocks and is thus heterogeneous with respect to the company market value. We retrieved from Datastream the monthly returns of the S&P 1500 components for the period January 1990 – October 2008. For these assets, the S&P 1500 index represents the appropriate equity benchmark, and the US 1-month Treasury Bill index, provided by Citigroup, is our proxy for the risk-free asset. For each asset, we consider logarithmic returns and excess returns over the returns of the risk-free asset and over the benchmark. As expected, these series are characterized by large deviations from normality.

Note that the index composition changes over time. Since our dataset includes the 1500 assets belonging to the S&P 1500 index at the end of October 2008, not all of them are available for the whole considered period (for example, in January 1990 only 754 assets out of 1500 were available). To deal with this problem, we

**Table 1:** List of performance measures considered. The first column reports the performance measure name as defined in Section 2. The other columns refer to the return type considered in the evaluation of the performance measures: the returns of a given asset, the excess returns with respect to a risk-free investment, and the deviations between the asset returns and the benchmark investment. When needed, beside the name of each performance measure we report the number of parameter combinations considered. The Treynor index, the Appraisal Ratio and the M2 index are not defined for deviations of asset returns with respect to benchmark returns. M2 is not defined for excess returns. In brackets we report the number of cases considered for each performance measure, deriving it from the parameter combinations previously discussed. For instance, the Burke and Sterling ratios have two different cases associated with the two values of the parameter  $w$ . Similarly, the Farinelli-Tibiletti ratios are included in eighteen different forms combining the six cases for the moment order pairs and the three thresholds. The LAP measures include 19 cases, obtained by combining the 4 performance measures described in the previous section, and 6 parameter combinations mimicking Farinelli and Tibiletti (2003) and Gemmill et al. (2006). The 5 cases of  $LAP^S$  computed with the  $FT$  index parameter combinations are not considered since these are equivalent to Omega measures.

Performance measures (cases)	Returns	Excess returns	Deviations from benchmark
Sharpe ratio	X	X	X
Treynor index	X	X	NA
Appraisal ratio	X	X	NA
Average R over MAD	X	X	X
Average R over MiniMax	X	X	X
Average R over Range	X	X	X
M2	X	NA	NA
Calmar ratio	X	X	X
Sterling ratio (2)	X	X	X
Burke ratio (2)	X	X	X
Sortino ratio	X	X	X
Kappa 3 measure	X	X	X
Farinelli-Tibiletti (18)	X	X	X
Average R over VaR (2)	X	X	X
Average R over ES (2)	X	X	X
VaR ratio (2)	X	X	X
Generalized Rachev Ratios (20)	X	X	X
MRAR (3)	X	X	X
LAP (19)	X	X	X
Total	80	79	77

followed two different strategies. At first, we focused on the last 120 observations of the sample, corresponding to the period November 1998 – October 2008: there are 1236 assets always included in the index over this range. On this reduced set of stocks, we performed a static analysis of the rank correlation using three different evaluation windows: November 1998 – October 2008 (120 monthly returns); November 2003 – October 2008 (60 monthly returns); November 2005 – October 2008 (36 monthly returns). This study allows a comparison of performance measures over time, avoiding possible effects due to a changing cross-sectional dimension. We also obtain some preliminary results on the window size effect and on the time-varying nature of the rankings.

In a second step, we focus on the entire sample (January 1990 – October 2008) and use a rolling approach to evaluate the stability of rankings over time. At this stage, the rank correlations are measured over a rolling window of 60 months for assets always available in each window. The number of assets is 754 in the first window and 1404 in the last one. This different approach allows a comparison of rank correlations over a number of periods.

The use of an increasing number of assets over time could be questionable. However, using only the 754 available for the entire sample period would have induced a sample selection bias in the analysis. Clearly, the optimal solution would have been to use the entire track record of all the S&P 1500 constituents, including dead or de-listed companies, but unfortunately this piece of information was not available from Datastream.

### **3 Empirical Analysis**

We compute the performance measures over the S&P 1500 constituents and compare them using the Spearman rank correlation ( $R_S$ ). We evaluate all reward and risk measures using their empirical counterparts. That is, we used sample moments and sample quantiles without using a dynamic parametric model for the returns density. These choices make our results comparable with those in Eling and Schumacher (2007) and Eling (2008). After the Z-transformation of Fisher (1915), the Spearman rank correlation has an asymptotic density which could be used to test the null hypothesis of independence between two variables. However, our purpose

is not to test independence, but rather to study the degree of correlation between ranks based on performance measures and, in particular, to detect measures that are highly correlated or concordant. Eling and Schuhmacher (2007) tested the null hypothesis  $R_S \leq p$  for different values of  $p$ . They found that for  $p = 0.917$  the null hypothesis was rejected for all assets. Note that the test cannot be applied under the null of unit correlation, i.e. perfect agreement, because, as claimed also by Eling and Schuhmacher (2007), in this case there is no discrepancy between the rankings induced by the performance measures and thus no variability.

In this work, we follow an approach similar to that of Eling and Schuhmacher (2007) and Eling (2008), but differing in the kind and in the number of assets used to compute performance measures. In fact, the database of Eling and Schuhmacher (2007), and Eling (2008) include only managed portfolios. In contrast, we focus on equities and, differently from the two previous studies, we always compute performance measures, and the associated rank correlations, across assets available over a common period. In addition, our results suggest that the threshold level  $p$  may depend on the asset type as well as on the sample dimension and on the set of selected performance measures.

Another important issue not considered by Eling and Schuhmacher (2007), is the definition of the decision rule that specifies when two performance measures carry different pieces of information. Since Eling and Schuhmacher (2007) found only very high correlations between performance measures, they did not face the problem of defining what is a “low” rank correlation. Instead, for our data, in order to develop a decision rule, we define as “low” a rank correlation lower than 0.8, being aware this is an arbitrary choice. With such a choice, we do find evidence of “low” correlation. We note that, the limiting value we fix, is anyway much smaller than the average rank correlation reported in the study of Eling and Schuhmacher (2007). Since we only know the value of the sample rank correlation,  $\hat{R}_S$ , to define a precise threshold, we considered the asymptotic distribution of  $R_S$ . We thus considered the critical value, at 1% significance level, of the test  $H_0 : R_S \leq 0.8$  against  $H_1 : R_S > 0.8$ . In detail, if we denote by  $\rho$  the Fisher transformation of  $R_S$ ,  $\rho = \frac{1}{2} \ln\left(\frac{1+\hat{R}_S}{1-\hat{R}_S}\right)$  and by  $\hat{\rho}$  the corresponding sample quantity, asymptotically we have

$$\sqrt{N-2}\hat{\rho} \sim \mathcal{N}(\rho, 1). \quad (28)$$

**Table 2:** Rank correlations across selected performances measures - drawdowns and quantile based measures. The first column reports the pair of performance measures compared and the corresponding parameter value. The other columns report the rank correlations across three return types (asset returns, excess returns with respect to a risk free investment, and deviations between asset returns and a benchmark investment), and three sample dimensions (36, 60 and 120 months). Bold values identify rank correlations below the minimum threshold of 0.822 defined in Section 3, and denote relevant differences across the ranks induced by the two performance measures compared.

Rank correlations	Returns			Excess Returns			Deviations from benchmark		
	36	60	120	36	60	120	36	60	120
Window length	1.000	0.997	0.998	1.000	1.000	0.998	1.000	0.999	0.992
Sterling (5%) and (10%)	0.997	0.974	0.880	0.996	0.969	0.866	0.995	0.979	0.903
Burke (5%) and (10%)	0.993	0.991	0.985	0.992	0.992	0.990	0.994	0.994	0.987
VR index (5%) and (10%)	<b>0.727</b>	<b>0.699</b>	<b>0.600</b>	<b>0.728</b>	<b>0.701</b>	<b>0.611</b>	<b>0.795</b>	<b>0.743</b>	<b>0.586</b>
VaR Ratio (5%) and (10%)	0.997	0.998	0.997	0.998	0.998	0.998	0.998	0.999	0.998
STARR (5%) and (10%)									

Note that in our case, the rank correlation is computed between rankings induced by performance measures within the set of the  $N$  considered assets. Thus ?? holds for  $N$  large, since, in general, a large number of assets is analysed within equity screening programs. This allows us to define the required threshold for  $R_S$  as

$$R_S^*(\alpha) = \frac{\left(\exp\left(\ln\left(\frac{1+R_S}{1-R_S}\right) + 2Z_{1-\alpha}\sqrt{\frac{1}{N-2}}\right)\right) - 1}{\left(\exp\left(\ln\left(\frac{1+R_S}{1-R_S}\right) + 2Z_{1-\alpha}\sqrt{\frac{1}{N-2}}\right)\right) + 1}, \quad (29)$$

where  $Z_{1-\alpha}$  is the  $(1 - \alpha)$ -th quantile of a standard normal distribution. Such a quantity, corresponds thus to the critical value for the null hypothesis reported above. Such a choice allows a more direct interpretation of results, without resorting to the Fisher transformation of all quantities. In our analysis, with  $N = 1236$  in the static case, and  $\alpha = 1\%$ , the threshold (or critical value) defining the low correlation is 0.822. We further note that the sample size plays a relevant role in the definition of the critical value. For a very small number of assets, say below 50, the critical value would be quite large, easily leading to an acceptance of the null. However, the normal use of performance measures within an equity ranking program (an equity screening rule) involves evaluations over hundreds of assets, thus increasing the power of the test.

### 3.1 Within Group Analysis

In this section we report, analyze and comment on the rank correlation between performance measures that differ only for the parameters included in their definition. The purpose of this section is to provide a first reduction of the number of performance measures included in Table 1.

The first group we consider includes some measures based on drawdowns: the Sterling and Burke indices. These two quantities depend on the number of returns used for their computation. In the previous section we suggested the use of at least two values associated with 5% and 10% of the sample dimension. Given these two sets of performance measures, we evaluate whether the sample size used in the computation of the indices provides a different ranking across the assets. The results are reported in the first and second row of Table 2. The rank correlations



**Table 3:** Rank correlations across selected performances measures - Generalized Rachev Ratios. The first column reports the set of performance measures considered within each row. The first and second rows report the average rank correlation across the Generalized Rachev ratios for the parameter combinations associated to Moderate, Conservative, Growth and Defensive investors with a given quantile level. The third row reports the average rank correlation between the two groups associated to the first and second row. The other three rows report the average rank correlation between the Generalized Rachev ratios for Aggressive investors with respect to other indices. Columns from 2 to 10 contain the rank correlations values across three return types (asset returns, excess returns with respect to a risk free investment, and deviations between asset returns and a benchmark investment), and three sample dimensions (36, 60 and 120 months). Bold values identify rank correlations below the minimum threshold of 0.822 defined in Section 3, and denote relevant differences across the ranks induced by the performance measures compared within each row.

Rank correlations	Returns			Excess Returns			Deviations from benchmark		
	36	60	120	36	60	120	36	60	120
Window length									
Within GR (10%) excl. Aggressive	0.989	0.986	0.982	0.989	0.986	0.982	0.99	0.986	0.981
Within GR (5%) excl. Aggressive	0.997	0.995	0.992	0.997	0.995	0.992	0.998	0.995	0.991
Between GR (5%) and GR (10%) excl. Aggressive	0.957	0.955	0.955	0.956	0.955	0.956	0.962	0.956	0.955
GR Aggressive (10%) wrt other GR (10%)	0.941	0.898	0.848	0.955	0.912	0.871	<b>0.772</b>	<b>0.699</b>	<b>0.788</b>
GR Aggressive (5%) wrt other GR (5%)	<b>-0.474</b>	<b>0.233</b>	0.946	<b>-0.489</b>	<b>0.150</b>	0.937	<b>0.047</b>	<b>0.660</b>	0.954
GR Aggressive (5%) - GR Aggressive (10%)	<b>-0.452</b>	<b>0.192</b>	0.853	<b>-0.459</b>	<b>0.128</b>	0.868	<b>-0.072</b>	<b>0.415</b>	<b>0.810</b>

show evidence of equivalent informative content of the performance measures with respect to the number of returns used for the evaluation of the Burke and Sterling indices. Results do not change with respect to the sample dimension or to the return used for the evaluation. We conclude that there is no need to consider the Sterling and Burke indices computed over different numbers of drawdowns. This result confirms the findings of Eling and Schuhmacher (2007).

The second set of performance measures we analyze includes the quantile based measures, with the exclusion of the Generalized Rachev ratios. Table 2 reports the rank correlations between the VR index, the VaR ratios and the STARR ratio at the 5% and 10% quantile levels. Results show that the VR index and the STARR ratios should be considered with a single quantile level (rank correlation is always higher than 0.985) while the VaR ratio should be considered with both the 5% and 10% quantiles, given that the rank correlation is lower than 0.822 in all cases and also reaches a minimum close to 0.6 with a 10 years sample dimension (irrespective of the return used).

Table 3 reports the rank correlations across the Generalized Rachev ratios. We recall that we computed 10 different *GR* ratios combining five parameter combinations (Aggressive, Growth, Moderate, Conservative and Defensive) with two quantile levels (5% and 10%). We distinguished two groups, separating the effect of the Aggressive indices. Our analysis points out that this last parameter combination is the most sensible to the sample dimension, providing results different from the other *GR* ratios when the sample used is medium to small (3 or 5 years). The difference tends to be canceled with the sample set to 10 years, with the exclusion of the case of the evaluation of deviations from the benchmark. Differently, the other *GR* ratios (Growth to Defensive) are almost equivalent (the smallest rank correlation is equal to 0.955). In addition, the effect of the quantile level is minor. Building on these results, we chose to include the Moderate *GR* ratio at the 10% level when the sample dimension is large (10 years). In contrast, when the sample is small or medium, the *GR* for Aggressive investors will also be considered (again at the 10% level).

Following the performance measure groups previously introduced, we move then to measures based on partial moments that include the indices of Sortino, the Kappa 3 index and the *FT* ratios. Similarly to the Generalized Rachev ratios, we group the *FT* performance measures into two sets, separately considering the

**Table 4:** Rank correlations across selected performances measures - Farinelli-Tibiletti Ratios. The first column reports the set of performance measures considered within each row. We considered two groups: the first is composed by the parameter combinations associated to Defensive, Conservative, Moderate and Growth investors (in all possible pairs); the second group contains the pairs of performance measures involving at least one measure associated to Aggressive investors. "Within" stands for the average of rank correlations across the pairs of measures with the same minimum acceptable return. "Between" stands for the average rank correlations across the pairs of measures with different minimum acceptable returns. As an example, the line "Within -0.02" in the first group, contains the averages of rank correlations over the following pairs: UPR-Defensive, UPR-Conservative, UPR-Moderate, UPR-Growth, Defensive-Conservative, Defensive-Moderate, Defensive-Growth, Conservative-Moderate, Conservative-Growth, Moderate-Growth. Columns from 2 to 10 contain the rank correlations values across three return types (asset returns, excess returns with respect to a risk free investment, and deviations between asset returns and a benchmark investment), and three sample dimensions (36, 60 and 120 months). Bold values identify rank correlations below the minimum threshold of 0.822 defined in Section 3, and denote relevant differences across the ranks induced by the performance measures compared within each row.

Rank correlations	Returns			Excess Returns			Deviations from benchmark		
	36	60	120	36	60	120	36	60	120
Window length	0.941	0.944	0.914	0.938	0.942	0.907	0.956	0.960	0.927
Within and Between UPR, Defensive, Conservative, Moderate and Growth									
Within -0.02	0.941	0.944	0.914	0.938	0.942	0.907	0.956	0.960	0.927
Within 0	0.915	0.919	0.871	0.912	0.917	0.872	0.928	0.926	0.882
Within 0.02	0.915	0.926	0.922	0.918	0.929	0.931	0.902	0.907	0.915
Between -0.02 and 0	0.865	0.843	<b>0.716</b>	0.848	0.825	<b>0.682</b>	0.881	0.843	<b>0.725</b>
Between -0.02 and 0.02	<b>0.557</b>	<b>0.428</b>	<b>0.147</b>	<b>0.534</b>	<b>0.406</b>	<b>0.140</b>	<b>0.581</b>	<b>0.414</b>	<b>0.213</b>
Between 0 and 0.02	<b>0.790</b>	<b>0.736</b>	<b>0.649</b>	<b>0.794</b>	<b>0.744</b>	<b>0.682</b>	<b>0.798</b>	<b>0.735</b>	<b>0.701</b>
Within Aggressive and Between Aggressive and Defensive, Conservative, Moderate and Growth									
Within -0.02	<b>-0.268</b>	<b>0.045</b>	<b>0.498</b>	<b>-0.365</b>	<b>-0.100</b>	<b>0.356</b>	<b>-0.111</b>	<b>0.067</b>	<b>0.470</b>
Within 0	<b>0.383</b>	<b>0.05</b>	<b>-0.108</b>	<b>0.532</b>	<b>0.255</b>	<b>0.163</b>	<b>0.014</b>	<b>0.027</b>	<b>-0.011</b>
Within 0.02	0.857	0.853	<b>0.819</b>	0.874	0.864	0.840	0.852	0.847	0.823
Between -0.02 and 0	<b>0.058</b>	<b>0.025</b>	<b>0.103</b>	<b>0.086</b>	<b>0.057</b>	<b>0.152</b>	<b>-0.047</b>	<b>0.025</b>	<b>0.146</b>
Between -0.02 and 0.02	<b>0.125</b>	<b>0.104</b>	<b>-0.029</b>	<b>0.112</b>	<b>0.073</b>	<b>-0.041</b>	<b>0.171</b>	<b>0.09</b>	<b>0.018</b>
Between 0 and 0.02	<b>0.513</b>	<b>0.349</b>	<b>0.242</b>	<b>0.574</b>	<b>0.421</b>	<b>0.333</b>	<b>0.386</b>	<b>0.346</b>	<b>0.269</b>

Aggressive parameter combination. The results are reported in Table 4, where the first group includes the parameter combinations Growth, Moderate, Conservative, Defensive as well as the Upside Potential Ratio (which is a special case of the *FT* index as we previously argued). Our analysis shows that these parameter combinations do not provide additional information or relevant differences in the ranking of the underlying assets (first to third rows). The result is marginally influenced by the sample length and the kind of return used in the evaluation of the indices. On the other hand, the threshold used in the index construction matters, making the indices sensibly different in terms of assets ranking (fourth to sixth rows of Table 4). In fact, the rank correlations across indices computed over different thresholds are generally small and always lower than 0.822. When considering the Aggressive parameter combination, the rank correlations are always small, and sometimes negative (this is a consequence of limited relevance given to the risk by that parameter combination). In addition, they are affected by both the sample dimension and the return type. Summarizing, we suggest considering the *FT* Moderate index (or Omega index) together with the Aggressive parameter combination, under all three of the thresholds considered. For the Sortino and Kappa 3 indices, the rank correlation with respect to the Omega index is higher than 0.98 and therefore the two indices are not considered.

Moving to the performance measures based on utility functions (Table 5), we first note that the MRAR indices with risk aversion set to 10 and 50 are almost equivalent. Therefore, we decide to focus on the measure with risk aversion set to 2 and 10. By contrast, in the LAP measures, the Hwang-Satchell, Moderate and Growth parameter combinations are almost equivalent while the Conservative case is very close to them. In order to provide a selection of measures which is limited, internally consistent, and that maximizes the difference across parameter combinations, we suggest focusing on the cases Defensive, Moderate and Aggressive. Within each group, we suggest considering all performance measures even if the Moderate case reports a high within-group average rank correlation.

Finally, we consider a further group composed by most of the traditional performance measures. Table 6 includes the rank correlation of these indices with the ranking induced by the Sharpe ratio. As we may observe, all indices are almost identical to the Sharpe ratio in terms of ranking of the assets. Some minor exceptions are the Appraisal ratio and the M2 index for the 3 year sample.

**Table 5:** Rank correlations across selected performances measures - utility based performance measures. The first column reports the set of performance measures considered within each row. We separately consider the MRAR measures and the Loss Aversion Performance measures. The last one are grouped depending on their parameters in Hwang-Satchell (HS), Defensive, Conservative, Moderate, Growth and Aggressive. Apart the Hwang-Satchell case, the groups do not include the LAP<sup>S</sup> measure which is equivalent to the FT Moderate Index with threshold set at zero. For MRAR measures we report the rank correlation coefficients. For LAP groups we report the average rank correlation within each group and between each pair of groups. Columns from 2 to 10 contain the rank correlations values across three return types (asset returns, excess returns with respect to a risk free investment, and deviations between asset returns and a benchmark investment), and three sample dimensions (36, 60 and 120 months). Bold values identify rank correlations below the minimum threshold of 0.822 defined in Section 3, and denote relevant differences across the ranks induced by the performance measures compared within each row.

Rank correlations	Returns			Excess Returns			Deviations from benchmark		
	36	60	120	36	60	120	36	60	120
Window length									
MRAR 2 - MRAR 10	<b>0.252</b>	<b>-0.076</b>	<b>-0.006</b>	<b>0.462</b>	<b>0.068</b>	<b>0.256</b>	<b>0.021</b>	<b>-0.124</b>	<b>0.024</b>
MRAR 2 - MRAR 50	<b>0.168</b>	<b>-0.132</b>	<b>0.003</b>	<b>0.404</b>	<b>0.047</b>	<b>0.260</b>	<b>-0.153</b>	<b>-0.160</b>	<b>0.052</b>
MRAR 10 - MRAR 50	0.894	0.858	0.955	0.940	0.922	0.974	<b>0.707</b>	<b>0.806</b>	0.97
LAP - Within HS	0.951	0.939	<b>0.685</b>	0.950	0.938	<b>0.685</b>	0.932	0.932	<b>0.715</b>
LAP - Within Defensive	<b>-0.067</b>	<b>-0.054</b>	<b>0.081</b>	<b>-0.052</b>	<b>-0.046</b>	<b>0.074</b>	<b>-0.101</b>	<b>-0.069</b>	<b>0.052</b>
LAP - Within Conservative	0.813	<b>0.759</b>	<b>0.534</b>	0.853	<b>0.776</b>	<b>0.510</b>	<b>0.777</b>	<b>0.767</b>	<b>0.494</b>
LAP - Within Moderate	0.958	0.927	<b>0.670</b>	0.957	0.929	<b>0.665</b>	0.944	0.924	<b>0.666</b>
LAP - Within Growth	0.897	0.864	<b>0.432</b>	0.898	0.861	<b>0.446</b>	0.850	0.837	<b>0.469</b>
LAP - Within Aggressive	<b>0.523</b>	<b>0.591</b>	<b>0.508</b>	<b>0.477</b>	<b>0.580</b>	<b>0.520</b>	<b>0.408</b>	<b>0.514</b>	<b>0.517</b>
LAP - Between HS-Defensive	<b>0.081</b>	<b>0.027</b>	<b>-0.121</b>	<b>0.078</b>	<b>0.012</b>	<b>-0.125</b>	<b>0.061</b>	<b>-0.008</b>	<b>-0.145</b>
LAP - Between HS-Conservative	<b>0.741</b>	<b>0.698</b>	<b>0.154</b>	<b>0.751</b>	<b>0.703</b>	<b>0.123</b>	<b>0.684</b>	<b>0.634</b>	<b>0.064</b>
LAP - Between HS-Moderate	0.948	0.928	<b>0.671</b>	0.947	0.929	<b>0.669</b>	0.935	0.925	<b>0.686</b>
LAP - Between HS-Growth	0.835	0.838	<b>0.562</b>	0.831	0.837	<b>0.567</b>	<b>0.787</b>	<b>0.807</b>	<b>0.590</b>
LAP - Between HS-Aggressive	<b>0.339</b>	<b>0.434</b>	<b>0.477</b>	<b>0.322</b>	<b>0.439</b>	<b>0.506</b>	<b>0.240</b>	<b>0.367</b>	<b>0.531</b>
LAP - Between Defensive-Conservative	<b>0.156</b>	<b>0.104</b>	<b>0.173</b>	<b>0.141</b>	<b>0.097</b>	<b>0.161</b>	<b>0.100</b>	<b>0.071</b>	<b>0.156</b>
LAP - Between Defensive-Moderate	<b>0.108</b>	<b>0.059</b>	<b>-0.023</b>	<b>0.103</b>	<b>0.045</b>	<b>-0.031</b>	<b>0.076</b>	<b>0.013</b>	<b>-0.061</b>
LAP - Between Defensive-Growth	<b>0.052</b>	<b>0.008</b>	<b>-0.042</b>	<b>0.044</b>	<b>-0.003</b>	<b>-0.048</b>	<b>0.036</b>	<b>-0.021</b>	<b>-0.060</b>
LAP - Between Defensive-Aggressive	<b>-0.051</b>	<b>-0.074</b>	<b>-0.145</b>	<b>-0.055</b>	<b>-0.077</b>	<b>-0.146</b>	<b>-0.037</b>	<b>-0.078</b>	<b>-0.139</b>
LAP - Between Conservative-Moderate	0.817	<b>0.783</b>	<b>0.417</b>	0.832	<b>0.791</b>	<b>0.388</b>	<b>0.769</b>	<b>0.738</b>	<b>0.326</b>
LAP - Between Conservative-Growth	<b>0.734</b>	<b>0.672</b>	<b>0.253</b>	<b>0.745</b>	<b>0.679</b>	<b>0.237</b>	<b>0.673</b>	<b>0.623</b>	<b>0.205</b>
LAP - Between Conservative-Aggressive	<b>0.128</b>	<b>0.134</b>	<b>-0.179</b>	<b>0.127</b>	<b>0.146</b>	<b>-0.177</b>	<b>0.028</b>	<b>0.043</b>	<b>-0.199</b>
LAP - Between Moderate-Growth	0.831	<b>0.819</b>	<b>0.536</b>	0.824	<b>0.817</b>	<b>0.541</b>	<b>0.787</b>	<b>0.791</b>	<b>0.568</b>
LAP - Between Moderate-Aggressive	<b>0.264</b>	<b>0.333</b>	<b>0.257</b>	<b>0.249</b>	<b>0.341</b>	<b>0.284</b>	<b>0.174</b>	<b>0.277</b>	<b>0.319</b>
LAP - Between Growth-Aggressive	<b>0.485</b>	<b>0.556</b>	<b>0.442</b>	<b>0.463</b>	<b>0.551</b>	<b>0.461</b>	<b>0.392</b>	<b>0.481</b>	<b>0.467</b>

**Table 6:** Rank correlation of traditional and similar performance measures with the Sharpe ratio over different sample length. The first column reports the performance measure which is compared to the Sharpe ratio. Columns from 2 to 10 contain the rank correlations values across three return types (asset returns, excess returns with respect to a risk free investment, and deviations between asset returns and a benchmark investment), and three sample dimensions (36, 60 and 120 months). Bold values identify rank correlations below the minimum threshold of 0.822 defined in Section 3, and denote relevant differences across the ranks induced by the performance measures reported in each row and the Sharpe ratio.

Rank correlations	Returns			Excess Returns			Deviations from benchmark		
	36	60	120	36	60	120	36	60	120
Window length	0.950	0.934	0.858	0.945	0.940	0.906	—	—	—
Treynor	0.950	0.934	0.858	0.945	0.940	0.906	—	—	—
Appraisal ratio	<b>0.792</b>	0.940	0.906	<b>0.555</b>	0.915	0.893	—	—	—
ERMAD	0.999	0.999	0.997	0.999	0.999	0.998	0.999	0.999	0.998
ERR	0.995	0.992	0.966	0.994	0.994	0.985	0.994	0.989	0.978
ERMM	0.991	0.987	0.956	0.990	0.990	0.978	0.992	0.986	0.969
M2	<b>0.719</b>	0.928	0.964	—	—	—	—	—	—

Overall, we may infer that the Treynor index, the Appraisal ratio and the indices replacing the standard deviation in the Sharpe with a proxy are all equivalent. We thus suggest introducing in the following analysis only the Sharpe ratio. Notably, this result is in line with the findings of Eling and Schuhmacher (2007). In our case, the rank correlations are not as high as shown by these authors. Furthermore, our results point out that the equivalence across the selected performance measures is not influenced by the return used for the evaluation and only scarcely affected by the sample dimension.

After this within-group analysis, we select the following performance measures: the Sharpe ratio; the Calmar ratio; the Sterling Ratio and the Burke ratio computed over the 5% of the sample dimension; the VR index and the STARR at the 5% quantile; the VaR ratio at both the 5% and 10% quantiles; the Generalized Rachev ratio with Moderate parameter combination at the 10% quantile level (one single index - the Aggressive index is included only if the evaluation window is small); the FT Moderate and Aggressive indices under all three threshold levels (6 indices); the MRAR index with risk aversion set to 2 and 10; and the LAP measures for Defensive, Moderate and Aggressive parameter combinations (9 indices). On the whole, the total number of selected measures is 26.

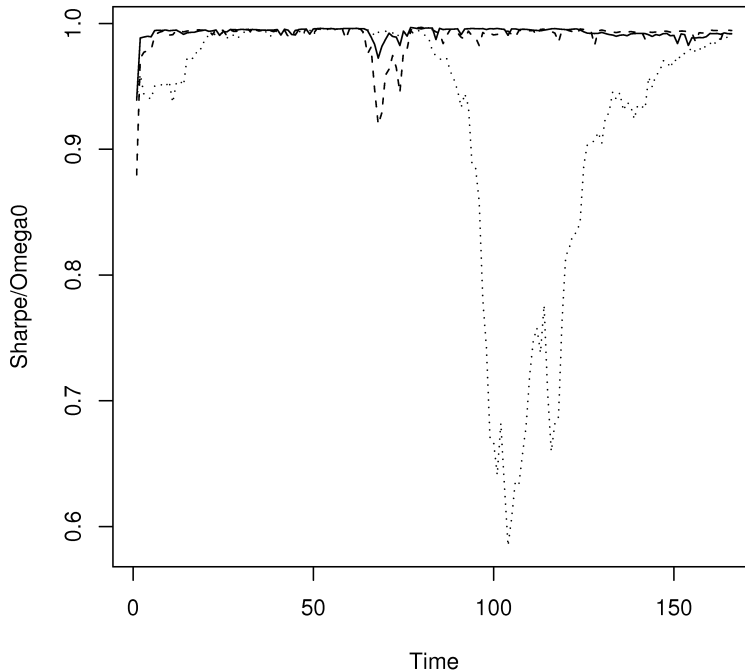
### **3.2 Descriptive and Rolling Analysis of Selected Measures**

We run additional correlation analysis on the reduced set of performance measures identified in the previous section. As a first outcome, we highlight that some of the measures are still highly correlated. In particular, we report in Table 7 the correlation between the Sharpe ratio and some selected measures. As shown in the table, we may infer that the Calmar ratio, the Sterling ratio (5%), the VR Index (5%), and the STARR (5%) are all equivalent to the Sharpe ratio. These findings confirm the results of Eling and Schuhmacher (2007) and are in line with the findings of Ortobelli et al. (2005) showing that traditional risk measures induce indifference across performance measures where the reward index is the average return. However, we obtain rather different rank correlations for Omega, with values going down to 0.536 and high rank correlation for long samples (120 months) only. Note that these differences are pronounced if we compute the Omega

**Table 7:** Rank correlation of selected measures with the Sharpe ratio. The first column reports the performance measure which is compared to the Sharpe ratio. Columns from 2 to 10 contain the rank correlations values across three return types (asset returns, excess returns with respect to a risk free investment, and deviations between asset returns and a benchmark investment), and three sample dimensions (36, 60 and 120 months). Bold values identify rank correlations below the minimum threshold of 0.822 defined in Section 3, and denote relevant differences across the ranks induced by the performance measures reported in each row and the Sharpe ratio.

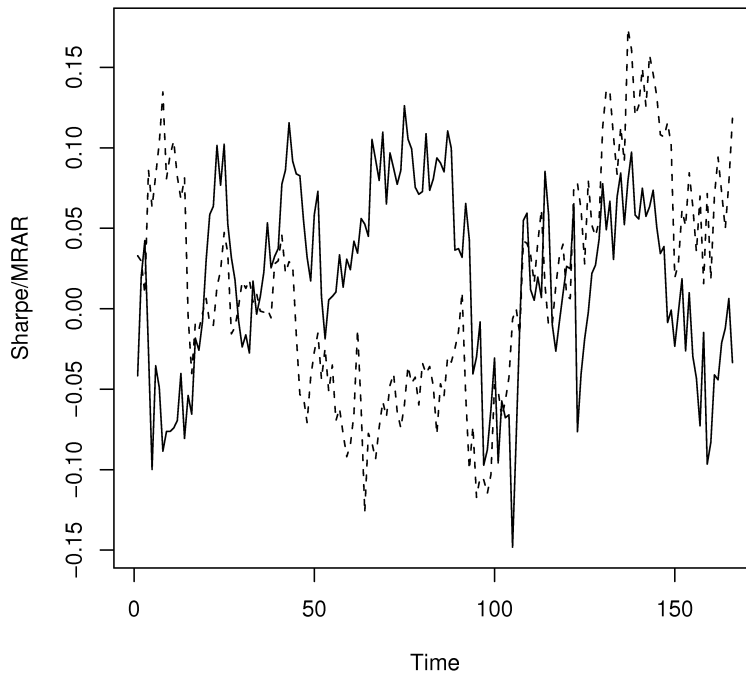
Rank correlations	Returns			Excess Returns			Deviations from benchmark		
	36	60	120	36	60	120	36	60	120
Window length	0.983	0.977	0.930	0.982	0.978	0.965	0.987	0.977	0.946
Calmar ratio	0.982	0.978	0.932	<b>0.785</b>	0.903	0.912	0.833	0.951	0.911
Sterling ratio (5%)	0.993	0.991	0.983	0.993	0.993	0.992	0.994	0.994	0.990
VR Index (5%)	0.996	0.994	0.983	0.997	0.996	0.994	0.995	0.996	0.993
STARR (5%)	<b>0.723</b>	0.933	0.991	<b>0.532</b>	0.866	0.991	0.865	0.907	0.993
Omega									





**Figure 1:** Ratio Sharpe/Omega for returns (full line), excess return (dashed line) and excess returns from benchmark (dotted line).

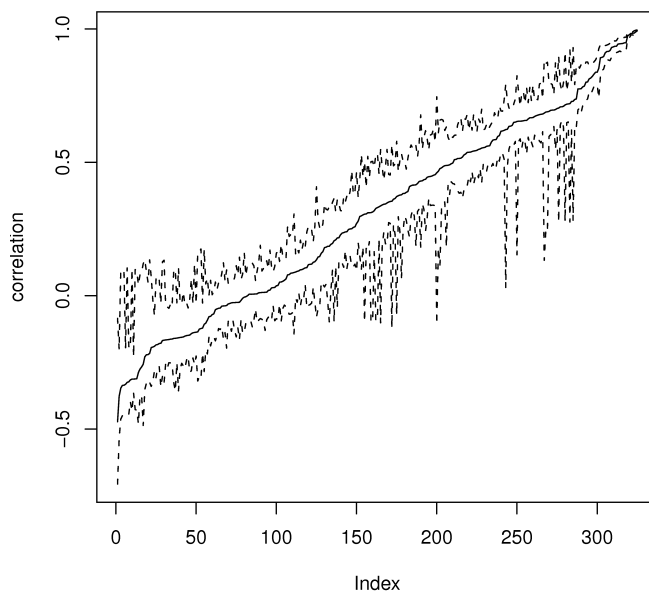
over Excess Returns or Deviations from the benchmark, while in the case of asset returns the Omega (with a zero threshold) is very close to the Sharpe, as in Eling and Schuhmacher (2007). Such a result points out that ranking of performance measures and their equivalence may be influenced by the kind of assets considered, the return type (nominal or excess return), the estimation window, and the sample period. To shed some light on the last motivation, and given the purposes of this paper, we perform a rolling analysis of the rank correlation across the reduced set of selected performance measures. Considering all the 1500 stocks in the S&P Index at the end of October 2008, and available over the range January 1990 to October 2008 (226 observations), we compute the rank correlation over 23 performance



**Figure 2:** Ratio Sharpe/MRAR(2) (full line) and Sharpe/MRAR(10) (dashed line).

measures (we drop from the set the Sterling ratio (5%), the VR Index (5%), and the STARR (5%)) on a rolling window of 60 months, obtaining 166 instances of the rank correlation matrix.

Across the performance measures with the highest average rank correlations, some pairs evidence a clear instability. This is the case for Omega with zero threshold and the Sharpe ratio when computed on deviations from the benchmark index (see Figure 1). Even though the global level of correlation is around 0.90, there are periods where the rank correlation is below 0.70 and periods where it is much higher than 0.90. Furthermore, this behaviour does not seem random but shows a clear persistence, which cannot be entirely associated to the rolling



**Figure 3:** Average rank correlations (full line), 5% and 95% quantiles (dashed lines) for each pair of selected performance measures. The average and quantiles are computed with respect to the time index. The rank correlations are ordered with respect to their sample average.

approach we follow. On the contrary, it does not seem to be dependent on the return type used for computing the performance measures. In fact, the instability is reduced when we consider simple returns or returns in excess from the risk-free return. However, this result is not shared by all the performance measures. As an example, let us consider the rank correlations between the pairs Sharpe-MRAR(2), and Sharpe-MRAR(10) computed using simple returns. Figure 2 shows that both MRAR(2) and MRAR(10) have low rank correlation with respect to the Sharpe ratio, but with relatively large changes over time, with a range going from about  $-0.15$  to about  $0.15$ . Similar results have been also observed for other pairs of performance measures and provide evidence of dynamics in the rank correlations. They also suggest that the use of one single index should be avoided given that, over

time, alternative performance measures may provide different informative contents, which could be relevant for selecting the optimal assets in a more appropriate way. In Figure 3, for each pair of performance measures (351 cases), we report the average, the 5% and 95% quantiles of the rank correlations computed on simple returns. These quantities have been evaluated using the time series of rank correlations computed over the entire set of 60 months rolling windows (166 observations). Data are ordered with respect to the average rank correlation. The graph clearly shows that rank correlations have, in many cases, a strong variation over time. Similar behaviours are obtained for excess returns with respect to a risk free or to the benchmark portfolio. In addition, we explore the relation between the sample length, the return type and the rank correlation levels. For this purpose, we run simple linear regressions across the rank correlations computed over different combinations of return types and sample periods. Let  $R_S(X_t, T)$  denote the set of rank correlations computed over the returns  $X_{i,t}$   $i = 1, 2, \dots, N$  using a sample of dimension  $T = 36, 60, 120$ . We consider the cross-sectional linear regressions across all different pairs of  $R_S(X_t, T)$  by varying the return type and the sample dimension. We obtain nine possible sets  $R_S(X_t, T)$  (three return type and three sample size) and 45 regressions of the form  $R_S(X_t, T) = \beta_0 + \beta_1 R'_S(X'_t, T') + \varepsilon$  where  $R'_S(X'_t, T')$  differs from  $R_S(X_t, T)$  either for the return type ( $X'_t \neq X_t$ ), the sample size ( $T' \neq T$ ), or for both. We then compare the  $R^2$  of the regressions and find that the sample size induces some change over rank correlations computed using the same return type. In fact, when  $X'_t = X_t$ , the  $R^2$  for the regressions with  $T = 120$  and  $T' = 60$  are the lowest, reaching a minimum of 0.57, which is still considerable. This is a somewhat expected result given that over shorter intervals the performance measures may be more sensitive to extreme returns. However, interesting observations emerge when comparing the rank correlations computed over the same sample dimensions ( $T = T'$ ) using different return types. In this case, we note that the return type plays an extremely limited role in the evaluation of rank correlations. The  $R^2$  of these regressions range from 0.93 to 0.99, without any clear difference across returns. As a result, we conclude that the choice of returns is not relevant within a selection process of assets (the simple return without any benchmark or risk-free asset can be used), while the use of at least 60 months could be suggested in order to reduce the impact of extreme returns.

## 4 Conclusions

A typical problem of portfolio management is to select some assets within a large group to build an optimal portfolio. One of the approaches followed is to create a screening rule, whose purpose is to order or rank assets. Within this framework, performance measures could do the task. Nevertheless, a different problem emerges: which measures to use? To answer this question, we followed the approach of Eling and Schuhmacher (2007) and compared performance measures using rank correlations. Such an approach allows evaluating the cross-sectional equivalence between the ranks induced by different performance measures. In this paper we have generalized the study of Eling and Schuhmacher (2007) by enlarging the selection of performance measures compared and exploring the dynamic properties of rank correlations. We have shown that performance measures based on partial moments and loss aversion are generally different from the traditional ones (including the Sharpe ratio). While the main result of Eling and Schuhmacher (2007) was that the most common performance measures induce very close rankings, we now show evidence that more flexible measures provide different rankings. Therefore, our results are more general than those in Eling and Schuhmacher (2007) and have shown evidence of differences in the ranking obtained by alternative performance measures. This finding might be read as an a bi-product of the different information extracted from the returns distribution by the performance measures we consider (some risk measures take into account higher order moments, or just the tail behaviour). Different rankings are, therefore and up to some point, the outcome of different informative content of performance measures.

As an additional finding, we have highlighted a changing behavior in rank correlations, even across pairs considered equivalent by Eling and Schuhmacher (2007). Our results suggest that different performance measures carry different information about asset returns, and also with respect to their relation with the risk-free asset and/or with the benchmark portfolio. As a consequence, if a set of performance measures is used to analyse, monitor and select assets, two elements should be considered: the possible equivalence between the measures included in the set; and the need of regularly checking and updating the set of performance measures since equivalence relations may vary over time. We have proposed an approach based on the rank correlations for selecting the set of performance

measures which are providing reasonably different asset rankings. The method we introduced is based on a statistical test based on the Fisher transformation of rank correlations.

The results we provided within this study could be made even more general by further enlarging the set of performance measures, for instance introducing generalizations of the Jensen Alpha obtained from factor models. Alternatively, the measures listed in Cogneau and Hubner (2009a, 2009b) could be considered.

Our findings could be exploited by building rules for asset screening based on an optimal combination of performance measures, for instance following the approach of Hwang and Salmon (2003). From a different viewpoint, the ranks obtained from each non-redundant performance measure might be used to define long or short positions. The first would be chosen among the assets with the highest ranks, while the second from the assets with lowest ranks. Combination of performance measures would result in an increase in the efficiency of those choices.

**Acknowledgements.** Both authors acknowledge financial support from the University of Padova grant CPDA073598. We also thank the Editor, two anonymous referees, Michele Doronzo, Enrico Schumann, and the participants to the CMS2009 conference for comments on previous versions of the current paper. Usual disclaimers apply.

## References

- Aftalion, F., and Poncet, P. (2003). *Les techniques de mesure de performance*. Economica, Paris.
- Avramov, D., and Chordia, T. (2006). Asset pricing models and financial market anomalies. *Review of Financial Studies*, 19 (3): 1001–1040. <http://ideas.repec.org/a/oup/rfinst/v19y2006i3p1001-1040.html>
- Barberis, N., Hwang, M., and Santos, T. (2001). Prospect theory and asset prices, *Quarterly Journal of Economics*, 116 (1): 1–53. <http://ideas.repec.org/a/tpr/qjecon/v116y2001i1p1-53.html>
- Biglova, A., Ortobelli, S., Rachev, S., and Stoyanov, S. (2004). Different approaches to risk estimation in portfolio theory, *Journal of Portfolio Management*, 31 (1): 103–112. <http://www.ijournals.com/doi/abs/10.3905/jpm.2004.443328>
- Burke, G. (1994). A sharper Sharpe ratio, *Futures*, 23 (3): 56.
- Cogneau, P., and Hubner, G. (2009a). The (more than) 100 ways to measure portfolio performance. Part 1: standardized risk-adjusted measures, *Journal of Performance Measurement*, 13 (4). <http://orbi.ulg.ac.be/handle/2268/2782>
- Cogneau, P., and Hubner, G. (2009b). The (more than) 100 ways to measure portfolio performance. Part 2: special measures and comparison, *Journal of Performance Measurement*, 14 (1). <http://orbi.ulg.ac.be/handle/2268/30494>
- Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues, *Quantitative Finance*, 1 (2): 223–236. <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.5992>
- De Miguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, 22 (5): 1915–1953. <http://ideas.repec.org/a/oup/rfinst/v22y2009i5p1915-1953.html>

- Eling, M. (2008). Does the measure matters in the mutual fund industry? *Financial Analyst Journal*, 64 (3): 54–66.  
[http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1266187](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1266187)
- Eling, M., and F. Schuhmacher (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking and Finance*, 31: 2632–2647. <http://ideas.repec.org/a/eee/jbfina/v31y2007i9p2632-2647.html>
- Eling, M., Farinelli, S., Rossello, D., and Tibiletti, L. (2011). One-size or tailor-made performance ratios for ranking hedge funds, *Journal of Derivatives and Hedge Funds*, 16: 267–277. <http://www.palgrave-journals.com/jdhf/journal/v16/n4/full/jdhf201020a.html>
- Farinelli, S., and Tibiletti, L. (2003). Upside and downside risk with a benchmark, *Atlantic Economic Journal*, Anthology Section, 31 (4): 387. <http://ideas.repec.org/a/kap/atlecj/v31y2003i4p387-387.html>
- Farinelli, S., Ferreira, M., Rossello, D., Thoeny, M., and Tibiletti, L. (2008). Beyond Sharpe ratio: optimal asset allocation using different performance ratios, *Journal of Banking and Finance*, 32: 2057–2063. <http://ideas.repec.org/a/eee/jbfina/v32y2008i10p2057-2063.html>
- Farinelli, S., Ferreira, M., Rossello, D., Thoeny, M., and Tibiletti, L. (2009). Optimal asset allocation aid system: from "one-size" vs "taylor-made" performance ratio, *European Journal of Operational Research*, 192: 209–215.
- Ferson, W.E., and Schadt, R. (1996). Measuring fund strategy and performance in changing economic conditions, *Journal of Finance*, 51: 425–461. <http://ideas.repec.org/a/bla/jfinan/v51y1996i2p425-61.html>
- Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population, *Biometrika*, 10: 507–521. <http://www.jstor.org/stable/2331838>
- Gemmill, G., Hwang, S., and Salmon, M. (2006). Performance measurement with loss aversion, *Journal of Asset Management*, 7 (3): 190–207. <http://www.palgrave-journals.com/jam/journal/v7/n3/abs/2240213a.html>



- Hwang, S., and Salmon, M. (2003). An analysis of performance measures using copulae, in: Knigth, J., and Satchell, S. (eds), *Performance measurement in finance: firms, funds and managers*, Butterworth-Heinemann Finance, Quantitative Finance Series.
- Jensen, M. (1968). The performance of mutual funds in the period 1945-1968, *Journal of Finance*, 23 (2): 389–416.  
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.139.6166>
- Kahnemann, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica*, 47: 263–291.  
<http://ideas.repec.org/a/ecm/emetrp/v47y1979i2p263-91.html>
- Kaplan, P.D., and Knowles, J.A. (2004). Kappa: A Generalized Downside Risk-Adjusted Performance Measure, Morningstar Associates and York Hedge Fund Strategies, January.
- Kestner, L.N. (1996). Getting a handle on true performance, *Futures*, 25 (1): 44–46. <http://www.allbusiness.com/personal-finance/investing-trading-futures/538174-1.html>
- Knigth, J., and Satchell, S. (eds) (2002). *Performance measurement in finance: firms, funds and managers*, Butterworth-Heinemann Finance, Quantitative Finance Series.
- Konno, H., and Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its application to Tokyo stock market, *Management Science*, 37: 519–531. <http://www.jstor.org/stable/2632458>
- Le Sourd, V. (2007). Performance measurement for traditional investment. *Financial Analysts Journal*, 58 (4): 36–52.
- Lintner, J. (1965). The valuation of risky assets and the selection of risky investment in stock portfolios and capital budgets, *Review of Economics and Statistics*, 47: 13–37. <http://www.jstor.org/stable/1924119>
- Markowitz, H. (1959). *Portfolio selection: efficient diversification of investments*, John Wiley.

- Modigliani, F., and Modigliani, L. (1997). Risk-adjusted performance – how to measure it and why, *Journal of Portfolio Management*, 23 (2): 45–54.
- Mossin, J. (1969). Security pricing and investment criteria in competitive markets, *American Economic Review*, 59: 749–756. <http://ideas.repec.org/a/aea/aecrev/v59y1969i5p749-56.html>
- Ortobelli, S., Rachev, S., Stoyanov, S., Fabozzi, F.J., and Biglova, A. (2005). The proper use of risk measures in portfolio theory, *International Journal of Theoretical and Applied Finance*, 8 (8): 1107–1133. <http://ideas.repec.org/a/wsi/ijtafx/v08y2005i08p1107-1133.html>
- Pedersen, C.S., and Rudholm-Alfvén, T. (2003). Selecting a risk-adjusted shareholder performance measure, *Journal of Asset Management*, 4 (3): 152–172. <http://www.palgrave-journals.com/jam/journal/v4/n3/abs/2240101a.html>
- Rachev, S., Martin, D., and Siboulet, F. (2003). Phi-alpha optimal portfolios and Extreme Risk Management, *Wilmott Magazine of Finance*, November, 70–83. [http://www.wilmott.com/pdfs/060530\\_martin.pdf](http://www.wilmott.com/pdfs/060530_martin.pdf)
- Shadwick, W.F., and Keating, C. (2002). A universal performance measure, *Journal of Performance Measurement*, 6 (3): 59–84.
- Sharma, M. (2004). A.I.R.A.P. - Alternative RAPMs for alternative investments, *Journal of Investment Management*, 3 (4). <https://www.joim.com/abstract.asp?ArtID=116>
- Sharpe, W.F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance*, 19: 425–442. <http://www.jstor.org/stable/2977928>
- Sharpe, W.F. (1966). Mutual fund performance, *Journal of Business*, 39 (1): 119–138. <http://finance.martinsewell.com/fund-performance/Sharpe1966.pdf>
- Sharpe, W.F. (1992). Asset allocation: management style and performance measurement, *Journal of Portfolio Management*, 18 (2): 7–19. <http://www.ijournals.com/doi/abs/10.3905/jpm.1992.409394>

- Sharpe, W.F. (1994). The Sharpe ratio, *Journal of Portfolio Management*, Fall: 45–58. <http://www.stanford.edu/wfsharpe/art/sr/sr.htm>
- Sortino, F.A. (2001). *Managing downside risk in financial markets*, Butterworth-Heinemann Finance, Oxford.
- Sortino, F.A., and van der Meer, R. (1991). Downside risk, *Journal of Portfolio Management*, 17 (Spring): 27–31. <http://www.ijournals.com/doi/abs/10.3905/jpm.1991.409343>
- Sortino, F.A., van der Meer, R., and Plantinga, A. (1999). The Dutch triangle, *Journal of Portfolio Management*, 26 (Fall): 50–58.
- Treynor, J.L. (1965). How to rate management of investment funds, *Harvard Business Review*, 43 (1): 63–75.
- Young, T.W. (1991). Calmar ratio: A smoother tool, *Futures*, 20 (1): 40. <http://www.allbusiness.com/business-finance/equity-funding-stock/261555-1.html>
- Young, M.R. (1998). A MiniMax portfolio selection rule with linear programming solution, *Management Science*, 44: 673–683. <http://www.jstor.org/stable/2634472>

Please note:

You are most sincerely encouraged to participate in the open assessment of this article. You can do so by either recommending the article or by posting your comments.

Please go to:

<http://dx.doi.org/10.5018/economics-ejournal.ja.2011-10>

The Editor