The cost of sustainability on optimal portfolio choices

Stefano Herzel, Marco Nicolosi, Cătălin Stărică

Abstract

We examine the impact of sustainability criteria, as measured by the KLD scores, on optimal portfolio selection performed on an investment universe containing the equities in the S&P500 index and covering the period between 1993 and 2008. The optimizations are done according to the Markowitz mean-variance approach while sustainability constraints are introduced by eliminating from the investment pool those assets that do not comply to different social responsibility criteria (screening). We compare the two efficient frontiers, i.e. the one without and the one with screening. A spanning test is performed to determine if the differences between the two types of efficient frontier are significant. We introduce a measure of how much an investor has to pay (through loss of return or through additional risk) in order to satisfy given sustainability criteria. The analysis is carried on separately on the three main dimensions of sustainability, namely Environmental, Social and Governance.

1 Introduction

Investment choices based on Socially Responsible (SR) criteria are assuming a greater relevance in today financial markets, both in terms of asset under management and of number of investors. Often, a SR portfolio strategy is implemented by eliminating from the investment set those companies that have issues related to at least one of the SR criteria. This procedure is called "screening" and can be implemented in different ways, depending on the emphasis put by the investor on different type of concerns, and on the depth (as percentage of the total number of assets or of market value) of the screening. The issues of concern are commonly classified under three main areas: Environmental (E), Social (S) and Governance (G), the so-called "three dimensions of SR". Our analysis is based on the scores provided by KLD Research and Analytics, a rating agency specialized in sustainability issues for equities traded in the US market.

What is the impact of different screening policies on the investment set? How strong are its effects in terms of capitalization? It is obvious that reducing the investment set diminishes the expected (risk-adjusted) returns, but can such a loss be measured? Do the effects of screening change with time or are they relatively stable? Does screening based on some of the SR dimensions have stronger effects for the investor? How do such effects depend on the level of screening and/or the level of risk that the investor is willing to bear?

Questions of this nature are natural for any portfolio manager or researcher with an interest in socially responsible investment. Quoting Kurtz (1997): "one of the most important questions, still unanswered about SR investment is: What does the efficient frontier for a socially responsible investor look like, and what does it imply for asset allocation". This paper is an attempt to provide an answer to these questions based on an investment exercise performed on an universe comprising an almost exhaustive subset of S&P500 firms and stretching on a sixteen year period from 1993 to 2008. We study how constraints based on the KLD sustainability scores affect portfolio choices in the classical mean-variance optimization framework. More concretely, we construct and compare quarterly the efficient frontier corresponding to the whole investment universe as well as those corresponding to smaller universes where firms with poor sustainability scores where removed.

Several measures of the impact of SR screening are put forth[1].

Earlier studies (such as Kinder at al. (1997)), compared the performance of SR indexes (like the Domini Social Index) to conventional ones (for example S&P500), reporting, in many cases, favorable results for SR screening[2]. After Kurtz (1997) observed that such approaches were flawed by the fact that they did not take into account differences in investment styles, namely capitalization, price to book ratios and dividends, the use of three (or four) factor models became a standard for this kind of analysis and a certain consensus of the findings started to emerge. Stone at al. (2001) studied the impact of SR screening on managed portfolios in the US equity market adopting the SR rating provided by KLD and found no significant differences between SR and not SR returns. Bauer et al. (2005) found, after controlling for the investment style, no evidence of significant differences in risk-adjusted returns between ethical and conventional funds for the 1990-2001 period. Statman and Gloushkov (2009) analyzed returns on US stocks rated by KLD during 1992-2007 and found no evidence that socially responsible investors had a return advantage relative to conventional investors[3]. Amenc and Lasurd (2008) analyzed the performances of sixty-two SR funds in the period 2002-2007 by computing their alpha with respect to the Fama-French model. In most of the cases they found a null or negative alpha, indicating that the SR funds did not create any value beyond that predicted by their respective exposures to the style factors. It seems hence relatively safe to state that the general consensus is that, after taking into account the specificity of the investment styles, the differences between the returns on SR instruments and conventional ones, both expected and realized, tend to vanish.

The vast majority of the literature on SR investment, and all of the above cited papers in particular, compares the returns of either singular assets, or of

[1]Needless to say, the results of any such analysis depend strongly on the choice of the model adopted for estimating the inputs of the optimization procedure. In our case, we used the three factor model of Fama and French (1992a) to estimate the covariance structure of the assets and we set the expected returns equal to the market implied ones.

[2]No general consensus on the effects of SR screening on financial returns exists in the early studies, some of them reporting a positive effect, others a neutral or even a negative one.

[3]More precisely, they observed that active screening, i.e. longing companies with positive ranking and shorting those with negative ranking, may induce some positive effects on returns, while passive screening only induces negative effects. The sum of the two strategies is usually null or slightly negative. Hence they concluded that for a SR manager it is important to perform some form of active screening of securities.

actively or passively managed portfolios, usually after filtering out the effect due to investment styles. Our paper extends hence the existing literature by a dynamical analysis of the effect of SR screening on the optimal mean variance allocation process. While the previous literature uses already constructed investment instruments, we put together ourselves the investment portfolios, mimicking the real-life situation faced by a portfolio manager that has to implement SR constraints in her investment policy. We construct and compare on a quarterly basis the efficient frontiers corresponding to an investment universe with and without SR screening. We analyze the time evolution of the "price of sustainability", defined as the loss in the Sharpe ratio due to shrinking of the investment set, for the three sustainability dimensions and for different levels of risk and of screening. We find that the "price of sustainability" is strongly related to the loss in capitalization[4] but also that it is surprisingly small comparing to the size of discharged market value: an "all-concern" screening, that eliminates all the companies that raise even a single issue of concern (more than 60% of the market capitalization) decreases the Sharpe ratio by no more than 7% for a medium level of risk (for details see Section 5).

2 The Price of Sustainability

In this section we formalize the optimization problem and propose a measure for the impact of screening on the optimal portfolios. We consider a set of N assets and denote their rates of return between time $t - \Delta t$ and time t by $R_t^i, i = 1, \ldots, N$. Let Σ_t be the covariance matrix and μ_t the vector of expected return of $R_t^i, i = 1, \ldots, N$ at time t. The optimal allocation problem

[4] We find that companies with a higher capitalization raise more "Social" concerns. For this reason, the S dimension is usually the one that produces a major loss in terms of capitalization, for any level of screening. This is the main reason why screening based on the S dimension has usually a stronger impact than other kinds of screening.

is given by

$$\min_{w} w'\Sigma_t w \tag{2.1}$$

$$w'\mu_t \geq R \tag{2.2}$$

$$\sum_{i=1}^{N} w_i = 1 \tag{2.3}$$

$$w \geq 0 \tag{2.4}$$

where w is a N-vector of portfolio weights, R is a parameter indicating the minimum level of acceptable expected return, (2.3) is a balance constraint and (2.4) is the short-selling constraint.

Screening is the most straightforward way to introduce SR constraints in a portfolio choice. Companies not complying with given SR requirements are simply removed from the investment universe. Hence, for any investment instance, this approach yields two portfolios: the optimal "screened" portfolio, obtained by solving the allocation problem over the restrained investment universe and the optimal "reference" portfolio obtained by solving the allocation problem over the whole investment universe.

For a given time t, let $\pi_t(R)$ be the ratio between the Sharpe Ratios of the optimal "screened" portfolio and of the optimal "reference" portfolio, for a level of expected return R. Note that $\pi_t(R)$ is always positive and smaller than 1. We define the *sustainability price* at time t for a level R of expected return as

$$p_t(R) = 1 - \pi_t(R). \tag{2.5}$$

This quantity measures the relative Sharpe Ratio loss due to SR screening.

We focus our analysis on three levels of minimum expected returns (three different values of the parameter R), associated to three different levels of risk. The first one is the expected return of the global minimum variance (GMV) portfolio, which we will refer to as $R1$. Since we want to compare portfolios on different frontiers at the same level of expected return, the level $R1$ is set as the highest expected return among the GMV portfolios for all the frontiers considered. The second value $R2$ corresponds to the market level of expected return and represents a medium level of risk. The third value $R3$ corresponds to the highest level of risk considered and is chosen such that $R2$ is the average between this return and $R1$. Note that the values of $R1, R2$ and $R3$ change with time.

3 Model calibration

In this section we explain how we obtain the inputs of the allocation problem (2.1)-(2.4), that is the covariance matrix Σ_t and the vector of expected returns μ_t. To estimate Σ_t we adopt the Fama and French model (see Fama and French 1992a, 1992b, 1993) as, probably, it is the most popular factor model:

$$R_t^i - RF_t = \alpha^i + \beta_1^i\,(R_t^M - RF_t) + \beta_2^i\,SMB_t + \beta_3^i\,HML_t + \epsilon_t^i\ , \tag{3.6}$$

where RF_t the risk free return, R_t^M is the return of the market portfolio, SMB_t is the return of the small minus big market capitalization factor and HML_t is the return of the high minus low book to market value ratio factor. The ϵ_t^i are the usual i.i.d idiosyncratic error terms with zero mean. The coefficient α^i represents the extra expected return of the i-th company that is not directly explained by the sensitivities to the factors.

We estimated the factor loadings $\beta_1, \beta_2, \beta_3$ by regressing the excess returns of single companies on the mentioned risk factors[5]. We re-estimated the model every three months between 1993 and 2008, using monthly data on a window of length 5 years.

Let us denote with $\hat{\beta}_t$ and $\hat{\epsilon}_t$ the $N \times 3$ matrix of the factor loadings and the $N \times T$ matrix of the error terms as estimated at time t respectively. The estimation is based on a regression performed at time t over past T observations. Then the estimated covariance matrix Σ_t is given by

$$\hat{\Sigma}_t = \hat{\beta}_t\,\widehat{\mathrm{cov}_t(f)}\hat{\beta}_t' + \widehat{\mathrm{cov}_t(\hat{\epsilon})} \tag{3.7}$$

where $\widehat{\mathrm{cov}}_t(f)$ is the variance matrix of the risk factors and $\widehat{\mathrm{cov}}_t(\hat{\epsilon})$ is the (diagonal) variance matrix of the error term.

The vector of expected returns μ_t is set to match the market implied vector of expected returns

$$\hat{\mu}_t = RF_t + \hat{\lambda}_t\hat{\Sigma}_t w_t^{mkt} \tag{3.8}$$

where w_t^{mkt} is the N-dimensional vector of the relative market capitalizations, that is its i-th component is the ratio between the market capitalization of the i-th company and the total capitalization of the market at time t. The parameter λ_t represents the risk aversion of a representative agent. Since a

[5]The time series of the risk factors come from K. R. French's web site

change in λ re-scales the excess returns without affecting the efficient portfolios and their variances, the price of sustainability $p_t(R)$ defined by (2.5) does not depend of λ_t.

This choice of μ_t can be interpreted as an assumption of market equilibrium since it implies the efficiency of the market portfolio. It is a common way to avoid the pitfalls of statistical estimation of the expected returns, a notoriously difficult issue, adopted, for example, as the starting point of the popular Black-Litterman model (Black and Litterman 1992)[6].

4 Social responsibility criteria

KLD Research and Analytics, Inc. rates the social responsibility of US companies in seven areas: Corporate Governance, Community, Diversity, Employee Relations, Environmental, Human Rights, Products. For each of these areas, KLD produces a number of indicators that come, each, in two flavors: "strength" and "concern". The values these indicators take are 1 or 0. While a point in a given strength indicator means that the company has a meritorious behavior with respect to the criterion in question, a value of 1 in a concern indicator signals a weakness of the company relative to the criterion related to that indicator. A score of zero indicates that the company has not qualified neither for a strength nor for a concern. In addition to the seven major area already mentioned, KLD provides also negative ratings on controversial business issues like Alcohol, Gambling, Firearms, Military, Nuclear Power and Tobacco.

KLD's database is organized in annual spreadsheets. From 1991 to 2000 KLD research covered approximately 650 companies belonging to the Domini 400 Social Index and/or to the S&P500 index. From 2001 KLD expanded its coverage to include the largest 1000 US companies by market capitalization while from 2003, KLD provides ratings for the largest 3000 US firms. The sustainability data are assembled at the end of each year on the basis of company's public information, corporate social responsibility reporting and other information obtained trough direct engagement with the company (if there is any). The scores are published in the month of January of the following year. In our analysis the first portfolio allocation of each year is done in the beginning of the month of March. In this way we make sure that

[6] Their proposers argued that "...the only neutral means is the set of expected returns that would clear the market if all the investors had identical views".

the KLD data set containing the sustainability scores of the previous year has been already released. For our analysis we consider KLD data from 1992 to 2007 and consequently portfolio allocation from 1993 to 2008.

5 Empirical analysis

5.1 Data aggregation and screening

We start this section by describing how we processed the KLD data. For every firm covered, we aggregated the strength indicators and, separately, the concern indicators for each of the seven sustainability areas mentioned in the beginning of section 4. The aggregation was done by summing up all the indicators and dividing by the number of indicators involved. Then we aggregated the newly created indicators of the seven areas in the three classical dimensions of Environmental, Social and Governance. Since two of the seven sustainability areas are identified as Environment and Corporate Governance, the Social dimension, was obtained from the aggregation of the remaining 5 areas.

At the end of this procedure each company has a strength and a concern score, standardized between 0 and 1, for each of the three macro-dimensions.

The returns for the companies considered in the optimization exercise were downloaded from Datastream[7]. For each investment exercise we kept only those companies for which there was a series of returns sufficiently long (at least 5 years) to estimate the covariance matrix. We also deleted from our set those firms that were not identifiable when matching the data of KLD and of Datastream [8] After pre-processing the data we kept an average of 470 companies per year, with around 460 companies for the first years and around 490 for the more recent years. The percentage of the market value lost with respect to the market value of the S&P500 index is 5% on average.

Let us describe now the screening process applied to the investment universe. The screening procedures have been applied separately to each dimension following two different approaches. The first approach, called "all-

[7]The data type used is “RI” that takes into account also for the dividends, assuming that dividends are re-invested.

[8]KLD identifies companies by the CUSIP code only since 1995, before that year it used just names, while Datastream adopts ISIN codes. Matching the two standards is not always trivial, we are grateful to Cristiana Manescu for helping us with this issue.

concern screening", removes from the universe of investments all those companies that raised at least one concern in the considered dimension. We will show below (Section 5.2 and Figure 1) that it is very invasive both in terms of loss of market value and in terms of number of companies discarded. For our analysis it has the additional drawback that the size of the investment set changes from year to year. Another point worth noticing is that whole industries might be excluded from the investment universe: for example, an "all-concern screening" along the E dimension will probably eliminate most of the companies in the Technology or Oil & Gas sectors.

The second approach we applied is a "partial screening" in which one eliminates from the investment universe a fixed given percentage of the companies for each industrial sector[9]. This approach keeps the size of the investment set constant. It has the advantage of preserving the diversification across industrial sectors and, for this reason, it is often adopted to build SR indexes, like for instance the popular Domini 400 Social Index.

We consider three different percentages for the "partial screening": 10%, 30% and 50% of companies were eliminated. For each of the sustainability criteria we sort the firms on the basis of their concern scores. Companies with a high negative concern are at the bottom of the list and are the first to be eliminated. If there is a tie in the number of concerns, we eliminate the companies with the smaller number of strengths. This procedure takes into account both concerns and strengths and yields a set of allocation of the same size over the whole time period under study. It is designed to yield a screened investment universe that has the same composition, in terms of industrial sectors, as the whole market.

5.2 Effects of screening on market values and sustainability prices

We start by examining the effect of screening on the investment universe. Figure 1 shows the market capitalization percentage eliminated (top panel) and the number of firms remaining (bottom panel) for "all-concern" screening along the three different dimensions of sustainability. The figure shows that such kind of screening can be extremely invasive. Screening along the S or G dimensions removes in recent years around 80% of S&P500 companies

[9]The sectors considered are: Basic Materials, Consumer Goods, Consumer Services, Financials, Healthcare, Industrials, Oil & Gas, Technology, Telecommunications, Utilities.

and almost 90% of market capitalization. The effect of screening based on the E scores is rather different: it appears to be constant through time and produces a loss of market capitalization of around 40%.

Figure 2 shows the market value percentage eliminated by partial screening, for different levels of screening. We observe that the effect of partial screening in terms of market value loss is rather constant in time, with screening along the Social dimension usually having a stronger impact.

A 10% screening eliminates up to 20% of market capitalization, while a 50% screening eliminates up to 70% for S dimension and 50% for E and G dimensions.

Next we discuss the results of the portfolio optimization exercise. Table 1 reports the mean and the 95-th percentile of the temporal distribution (quarterly observations from 1993 to 2008) of the price of sustainability. Screening is implemented for the main dimensions of sustainability separately (E, S, G) and for three different portfolios $R1$ $R2$, $R3$, corresponding to different classes of risk. This table provides an indication of the amount of Sharpe ratio loss due to different kind of screening, at different risk levels and along different sustainability dimensions. For example, an investor who plans to adopt a 10% screening on the environmental dimension at the lowest possible risk ($R1$) must be prepared to loose on average 1.5% of the Sharpe ratio but, if she is unlucky (95% percentile), the loss can be as big as 3.6%. As expected, her prospective loss decreases if she is willing to take on more risk choosing the levels $R2$ or $R3$. We note that the loss increases with the level of screening, but it remains relatively low for almost all cases (even at a 50% or all-concerns screening). Screening along the social dimension is usually more costly in terms of loss of Sharpe ratio than the other two. The loss is usually greater and more volatile for portfolios of type $R1$[10].

The time evolution of the sustainability prices $p_t(R)$ corresponding to the $R2$ level of returns for "all-concern screening" and for 10% screening (the two extremes of the screening level) are displayed in Figures 3 and 4 respectively. For the "all-concern screening" we see that screening along the E dimension produces the smallest drop in the Sharpe ratio: smaller than 1.3% in 95% of cases. Screening along the G and S dimensions can cause maximal reductions of at most 5.1% and 6.4%, respectively. These numbers

[10] As we will argue in Subsection 6.1, the results for portfolio $R1$ are affected by stronger estimation errors and are not always significant. For this reason we prefer to restrict the discussion to portfolios of type $R2$ and $R3$.

seem remarkably low, if one considers that the market capitalization loss can be as large as 90%.

The overall picture emerging from the analysis is that of an hierarchy in the percentage of Sharpe ratio loss due to screening based on sustainability scores *independently on the level of screening* with screening along the Social dimension being the most penalizing while that along the Environmental one the least. Screening on Governance scores yields a reduction of the Sharpe ratio closer to that of the S (E) screening for higher (lower) levels of screening.

We also note that portfolios with the $R2$ risk level are affected by a smaller drop in the Sharpe ratio. Sustainability prices for $R2$ and $R3$ levels of risk are rather stable during the whole period, especially for lower levels of screening.

6 Statistical Tests

This section presents some statistical tests on the results presented above. The first test is on the robustness of the sustainability prices with respect to variations in input data. Then we examine if discharging assets because of a rule based on KLD ratings produces significatively different results, in terms of the impacts on market capitalization and on sustainability price, from those obtained by a rule based on random exclusions. The third and the fourth tests compare the ex-post performances of the optimal portfolio choices at the risk levels $R1, R2, R3$ and for different levels of screening.

6.1 Robustness

Here we study the robustness of the sustainability price with respect to variations of the efficient frontiers due to estimation errors. The efficient frontiers depend on the estimated covariance matrices, that is on the loading factors computed by the regressions. One possible way to assess the robustness of our measures is therefore to consider the asymptotic distribution of the loading factors as estimated by regression (3.6) and then to simulate new values for them. From the simulated values of beta we recompute a new covariance matrix as in (3.7) and a new vector of expected returns as in (3.8), then we perform a new optimization. Repeating such procedure a number of times gives us a confidence interval for the sustainability prices. Figure 6 shows the results of this straightforward (but computationally intensive) approach to the case of 10% E-screening for the last month of allocation (December

2008). The three bars correspond to the simulated distribution of the loss of Sharpe ratio for the three level of expected returns $R1$ $R2$ and $R3$. The candles show the median and the 25-th and 75-th percentiles. Outliers are represented as crossed points, while the circles indicate the loss of Sharpe ratio measured for the three original portfolios, before resampling. It is obvious that results are more robust with respect to the variability of input estimates for portfolios $R2$ and $R3$, but very sensitive to the input estimates for portfolio $R1$. Similar results apply also to other cases and make us believe that the high volatility of sustainability prices of portfolios $R1$ is due to estimation error.

6.2 Random screening

Is there anything special with a screening based on SR criteria? What happens when the same number of assets is eliminated by following different rules? To answer such questions we applied a procedure similar to the "partial screening" described in Subsection 5.1 with the only difference that the firms were excluded from the investment set by a random choice and not because of poor KLD scores. We repeated this procedure, which we called "random screening", 100 times for each year of the sample. Finally, we computed the percentage of market capitalization and of Sharpe Ratio lost after each "random screening" to compare them with the corresponding quantities computed after the "partial screening" based on KLD scores on the E, S, and G dimensions.

The results of the experiment are shown in Figure 5 for the case of the 10% screening and risk level $R2$. The bottom panel of Figure 5 represents, like the top panel of Figure 2, the loss in Market Value for a 10% screening. However, in this case such a loss is compared to the distribution of the random sample (the candles show the median and the 25-th and 75-th percentiles). The figure clearly shows that the losses in capitalization due to the S screening (represented by crosses) are, for most of the years, higher than those produced by the other two dimensions and than those due to random screening. Also the losses due to E screening (represented by circles) are usually very high compared to the random sample. The fact that screening according to the S and E criteria eliminates such a relevant amount of market capitalization is reflected in the loss of Sharpe Ratio for the level $R2$, shown in the top panel of Figure 5, that is the analogous to Figure 4 with the addition of the distribution candles. Here we see that the losses due to S or E screening are higher than most of those due to random screening. The G screening has a lower impact on the loss of market value and, consequently, on the cost of sustainability, although for the last years of the sample, it also produced a few outliers.

6.3 Spanning test

The goal of this paper is to compare the mean-variance efficient frontier corresponding to the unrestricted investment universe to that corresponding to a smaller universe that excludes those assets that do not satisfy some sustainability conditions. A spanning test, see Huberman and Kandel (1987) and De Roon and Nijma (2001), is a statistical tool to check whether the difference between two frontiers is statistically significant or only due to sampling error. The test tells if the efficient frontier built on an investment universe changes significantly when other assets are added to the investment pool. Due to the numerous instances of testing resulting from the dynamic approach that we are advocating and from working with different screening rules, we will apply the spanning test in a particular set-up that we now describe.

Let us consider a market with three so-called "conventional" assets plus three so-called "sustainable" assets. The three "conventional" assets are the efficient portfolios computed over the whole investment universe for the three levels of expected returns $R1$, $R2$ and $R3$, while the three "sustainable" assets are the corresponding portfolios computed over the screened universe, for a given level of screening and a given sustainability dimension[11]. We apply the spanning test to answer the question whether the mean-variance frontier of the "sustainable" assets remains efficient when the investment opportunities are increased to include also the "conventional" assets or, in other words, whether adding the three "conventional" assets to the group of three "sustainable" ones significantly enhances the opportunities of diversification.

In the standard framework of spanning test, one starts from an investment universe with K risky assets to which one adds N assets. The first K assets are usually called the benchmark assets while the additional N assets are referred to as the test assets. In our case, the benchmark assets are the $K = 3$ portfolios optimized over the screened universe while the test ones are the $N = 3$ efficient portfolios computed over the whole investment universe for

[11] The returns of such portfolios are in fact the returns of dynamic strategies (or managed funds) re-balanced quarterly over the period considered.

the three levels of expected returns $R1$, $R2$ and $R3$. If there exists exactly one portfolio for which investors can not improve their opportunities, then the two frontiers have only one point in common and we say that there is intersection. If there are no improvement possibilities by adding the "conventional assets" to the set of choices, the two frontiers coincide and we say that there is spanning.

Let R_{t+1} be the vector of the returns of the "sustainable" assets and r_{t+1} the vector of the returns of the "conventional" assets. As explained in Huberman and Kandel (1987) and in De Roon and Nijman (2001), the starting point of the spanning test is the linear relation

$$r_{t+1} = a + BR_{t+1} + \epsilon_{t+1} \ . \tag{6.9}$$

where a is a N-dimensional vector and B is a N by K matrix.

As shown in De Roon and Nijman (2001), in absence of short-sale constraints, spanning implies that the following conditions hold

$$a = 0, \qquad Bi_K - i_N = 0 \ , \tag{6.10}$$

where i_K and i_N are vectors of ones whose dimensions are respectively K and N. The test statistic for spanning is hence a Wald statistic that is asymptotically distributed as a χ^2_{2N} with $2N$ degrees of freedom.

In the case of short-sale restrictions on the benchmark assets (see De Roon et al. 2001) one has to divide the frontier of the benchmarks ("sustainable" in our case) in sub-parts and to test spanning for each single part. The constrained frontier has to be thought as the sum of parts of unconstrained frontiers, where each part is built over the subset of assets for which constraints are not binding. Then the responsible constrained frontier spans the frontier over the whole set of assets if there is spanning for each single sub-part. When short sales are forbidden also for the test ("conventional" in our case) assets, De Roon et al. (2001) show that to test the spanning for the generic single sub-part of the frontier corresponds to test that $2N$ inequality constraints hold. Suppose we consider the j-th sub-part of the frontier and let us indicate with $R^{[j]}$ the returns of the subset of assets over which that part of frontier is computed. Suppose that the size of such a subset is $L^{[j]} < K$. Then one should run the regression (6.9) with $R^{[j]}$ as independent variables and testing for the following inequality constraints:

$$a + \eta_{max}(Bi_L - i_N) \leq 0,$$
$$a + \eta_{min}(Bi_L - i_N) \leq 0, \tag{6.11}$$

where η_{min} and η_{max} are the zero beta returns of the portfolios at the edges of the j-th sub-part of the frontier we are considering (as shown in De Roon et al. 2001) and the parameters depend on the index $[j]$.

Let us note that the quantity $a + \eta(Bi_L - i_N)$ for a generic η is nothing else than the vector of Jensen's alphas of the test assets with respect to the efficient portfolio of the L benchmark assets, whose zero-beta return is η.

The restrictions (6.11) can be tested using a Wald test that in the case of inequality constraints was studied in Kodde and Palm (1986). The test statistic, under the null hypothesis is asymptotically distributed as a mixture of χ^2 with $2N$ degrees of freedom (see De Roon et al. 2001).

The results of spanning test are shown in Table 2 for all levels of screening and for all sustainability dimensions. We separate the case of allowed short selling (columns labeled "ss") from that of short selling not allowed ("no ss"). The Wald statistic ξ is reported together with its p-value. Following De Roon et al. (2001) we choose an upper and a lower bound to the zero beta return of portfolios along the efficient frontier corresponding to the screened investment universe. We chose the lower bound to be $\eta = 0$ meaning that investors have the possibility to invest their money in a risk-free asset with zero net return. The upper bound for η is just the return of the global minimum variance portfolio when short selling is allowed, corresponding to the zero-beta return given by the intercept of the asymptote to the mean-variance frontier. When short selling is not allowed the upper bound has to be computed in order to take into account the effect of the constraints.

If short selling is allowed, the null hypothesis of spanning can be rejected in most of the cases. It means that the efficient frontier corresponding to the screened portfolios and the one obtained over the whole investing universe are significantly different. For the lowest level of screening the spanning hypothesis is rejected only for the E dimension. The levels of significance of the test vary. In particular the test is highly significant for the highest levels of screening and also for the E dimension independently of the level of screening.

When short selling is not allowed, spanning can not be rejected when the screening is performed along the for S and G dimensions. This means that an investor who allocates her wealth imposing constraints on the social or corporate governance performance of the firms, does not improve the performance of her portfolio by removing the CSR constraints. A significant difference is instead found when the screening is performed along the E dimension even at medium levels of screening. That is, even though short selling is not allowed, restricting the investment universe based on environmental scores can significantly reduce the performance of the investment.

6.4 Comparison of realized returns

In this subsection we compare the realized returns of the screened and non-screened portfolios. More precisely, for each level of screening and for each sustainability dimension we compare the Sharpe ratios computed on the realized returns of the portfolios constructed from the whole investment universe to the corresponding ratios of portfolios built from screened assets. We perform two different tests for the difference in the Sharpe ratios. The first one is a classical t-test. In order to compare the Sharpe ratios, the realized returns were first standardized by local standard deviation estimates[12]. Then the expected values of the standardized series of realized returns were compared. The test revealed no significant differences.

We also implement a robust test for differences in the Sharpe ratios, developed in Ledoit end Wolf (2008). The test is based on a construction of a studentized bootstrapped confidence interval. The test is particularly appropriate when the returns have heavy tails or when there is some autocorrelation between the returns or their squares. The method also takes into account finite sample effects. The test is proven to be more robust than other tests for the difference of Sharpe ratio (see for example Jobson and Korkie 1981).

This test confirms the results of the previous one: the differences in ex-post Sharpe ratios between the two classes of portfolios are never significant. This is not in contrast to the responses of the spanning test, because the spanning test examine all possible combinations of the portfolios. In fact, suppose that Sharpe ratios are not very different, but the investment can be improved by diversification, then the spanning test would reject the hypoth-

[12] The standard deviations were computed monthly using the observations of the previous year and an exponential filter with a decay parameter $\lambda = 0.9$.

esis of spanning, as it happens for the case of allowed short selling.

7 Conclusions

We studied the effect of several kinds of Social Responsible screening on optimal mean-variance portfolios during the period from 1993 to 2008 for firms belonging to the S&P500 index. The analysis was based on the sustainability scores provided by KLD, on the variances of the returns estimated by the Fama-French three factor model and on the expected returns obtained from the market neutral assumption of the Black-Litterman model.

We observed that the market capitalization eliminated by screening increases with time, because firms with higher capitalization are raising more concerns, according to the KLD scores, especially in the Social and Governance dimensions, in more recent times. To summarize the effects of screening we observe that the 10% partial screening eliminates up to 20% of the market capitalization, while the 50% partial screening discharges up to 70% of it. All-concerns screening may be even stronger: for example, in the year 2007, the all-concerns screening in the Social dimension eliminates almost 90% of the market capitalization. In general, Social screening has the strongest impact in terms of capitalization because it tends to eliminate firms with higher market values.

We proposed a measure that we called "price of sustainability", based on the loss of Sharpe Ratio after the screening. It measures ex-ante, that is basing on the currently available information, what is the effect of the screening on optimal portfolio choices. We observed that the level of the "price of sustainability" is usually rather small, which can be unexpected if compared to the extent of the loss in market capitalization. We analyzed three levels of portfolio risk, which we called $R1$, $R2$ and $R3$, with $R1$ representing the minimum risk, $R2$ the risk of the market portfolio and $R3$ the most risky one. We observed that portfolios at the medium risk level $R2$ have the smallest sustainability price and that sustainability prices for $R2$ and $R3$ are rather stable during the period of observation, especially for lower levels of screening Comparing the three dimensions of sustainability we noted that the Social one produces the highest sustainability price, while the smallest one is obtained by the Environmental dimension, independently of the level of screening. Screening based on Governance has an intermediate cost, usually comparable to that based on the Social dimension for higher levels of screening, and to the Environmental one for smaller levels.

We performed several tests on our results. We checked the robustness of the sustainability price with respect to estimation error by performing a test based on simulating the loading factors of the three factor model, obtaining that the sustainability prices at levels $R2$ and $R3$ are robust, while at level $R1$ they are very sensible to input estimates. This explains the higher volatility of the time series of sustainability prices at level $R1$. To examine wether there is anything special with respect to the screening based on KLD scores we compared it to a random rule and concluded that the Social screening in general exclude a much larger fraction of market value than any random choice. This is also reflected in the price of sustainability. The ex-post analysis of the realized returns of socially responsible and of conventional portfolios corresponding to the risk levels $R1$, $R2$ and $R3$ shows that there is not a significant difference in terms of Sharpe ratios. A spanning test shows that the inclusion of conventional portfolios may improve the investment opportunities by diversification, but only when short selling is allowed.

The general picture emerging from our analysis is that restricting the investment set for Socially Responsible reasons has a strong effect on market capitalization but at a small price in terms of Sharpe Ratio. From the point of view of a responsible investor this is good news.

The price of sustainability

		$R1$		$R2$		$R3$	
		mean	95%	mean	95%	mean	95%
	E	1.5	3.6	0.4	0.8	0.2	0.6
10%	S	1.9	4.8	0.5	1.1	0.5	1.5
screening	G	1.3	2.8	0.3	0.6	0.3	0.7
	E	7.3	11.3	0.6	1.1	0.5	1.0
30%	S	6.3	11.8	1.0	1.9	1.3	3.2
screening	G	7.3	11.4	0.5	1.3	0.8	1.5
	E	13.2	18.1	0.7	1.3	0.9	1.6
50%	S	11.2	16.8	1.6	2.5	2.9	8.0
screening	G	13.2	19.9	0.8	1.7	2.2	5.8
	E	8.2	14.9	0.7	1.3	0.4	1.0
all-concerns	S	20.1	36.7	2.8	6.4	5.0	12.2
screening	G	14.7	36.6	2.8	5.1	4.8	16.7

Table 1 *Mean and 95-th percentile of the time series (quarterly observations from 1993 to 2008) of the price of sustainability for different types of screening and different levels of risk. Screening is implemented for the three dimensions of sustainability, Environmental (E), Social (S) and Governance (G). The risk levels correspond to the global minimum variance portfolio ($R1$), the portfolio at the level of the market portfolio ($R2$) and a portfolio ($R3$) symmetric to $R1$ with respect to the market portfolio.*

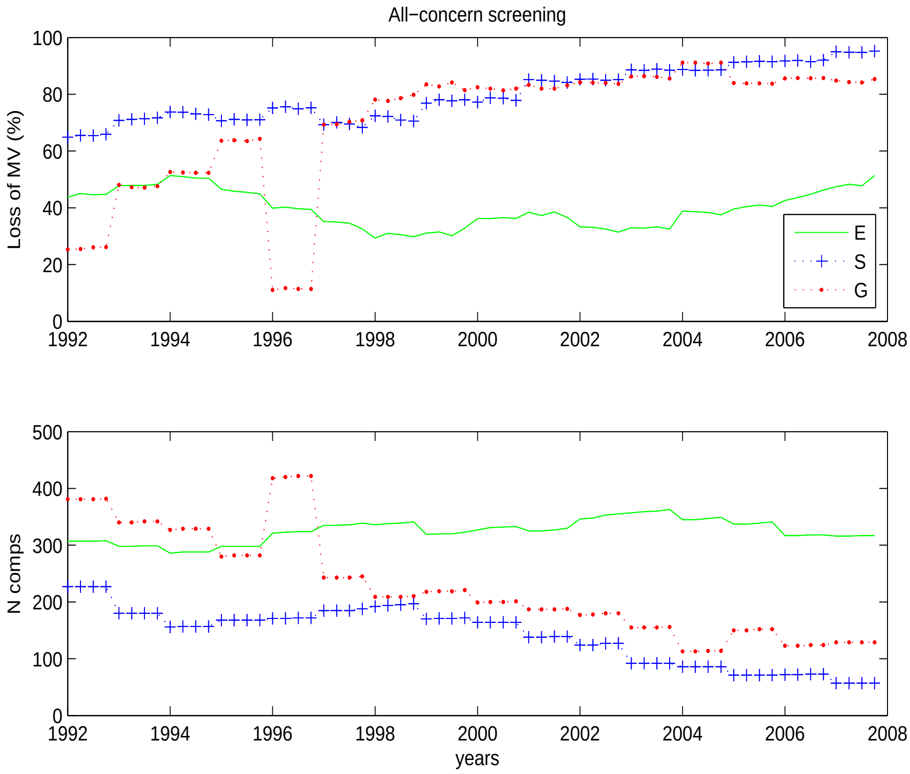

Figure 1 *Effect of all-concern screening during the period 1992-2007. Top: percentage of market value discharged by all-concern screenings on the Environmental (solid line), Social (crossed line) and Governance (dotted line) dimensions. Bottom: number of firms that have not been discharged by the screening.*

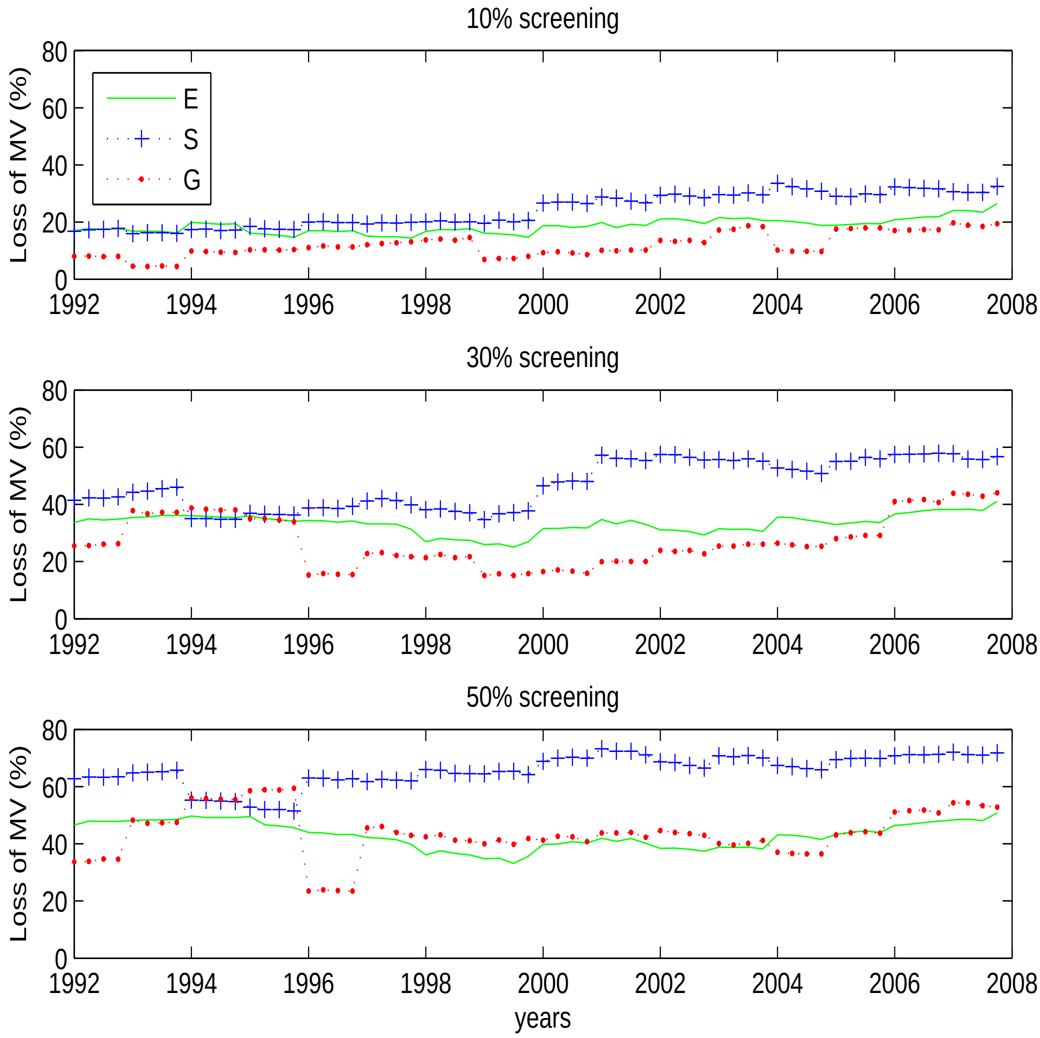

Figure 2 *Percentage of market value eliminated by different levels of partial screening during the period 1992-2007 on the Environmental (solid line), Social (crossed line) and Governance (dotted line) dimensions. Top:* 10% *partial screening. Middle:* 30% *partial screening. Bottom:* 50% *partial screening.*

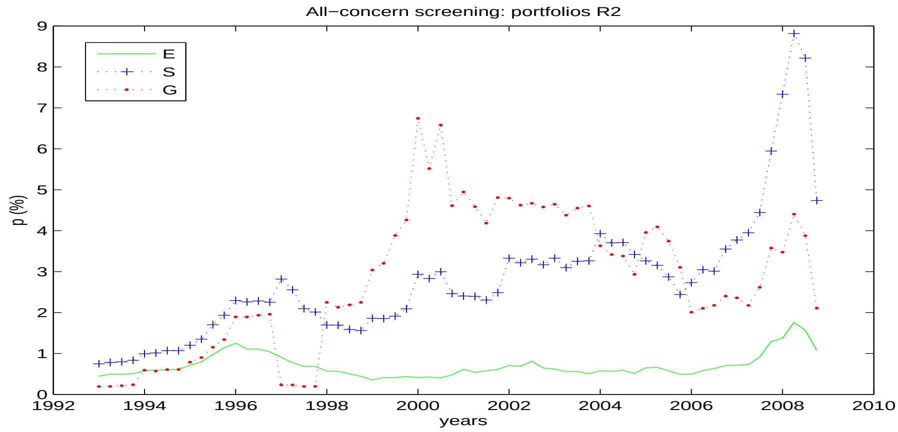

Figure 3 *Price of sustainability for portfolios at the level of risk R2 for the all-concern screening on the Environmental (solid line), Social (crossed line) and Governance (dotted line) dimensions.*

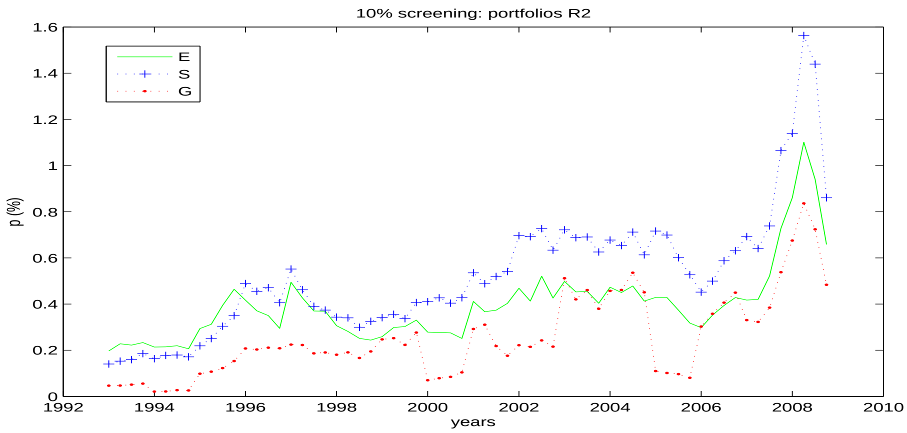

Figure 4 *Price of sustainability for portfolios at the level of risk R2 for a 10% partial screening on the Environmental (solid line), Social (crossed line) and Governance (dotted line) dimensions.*

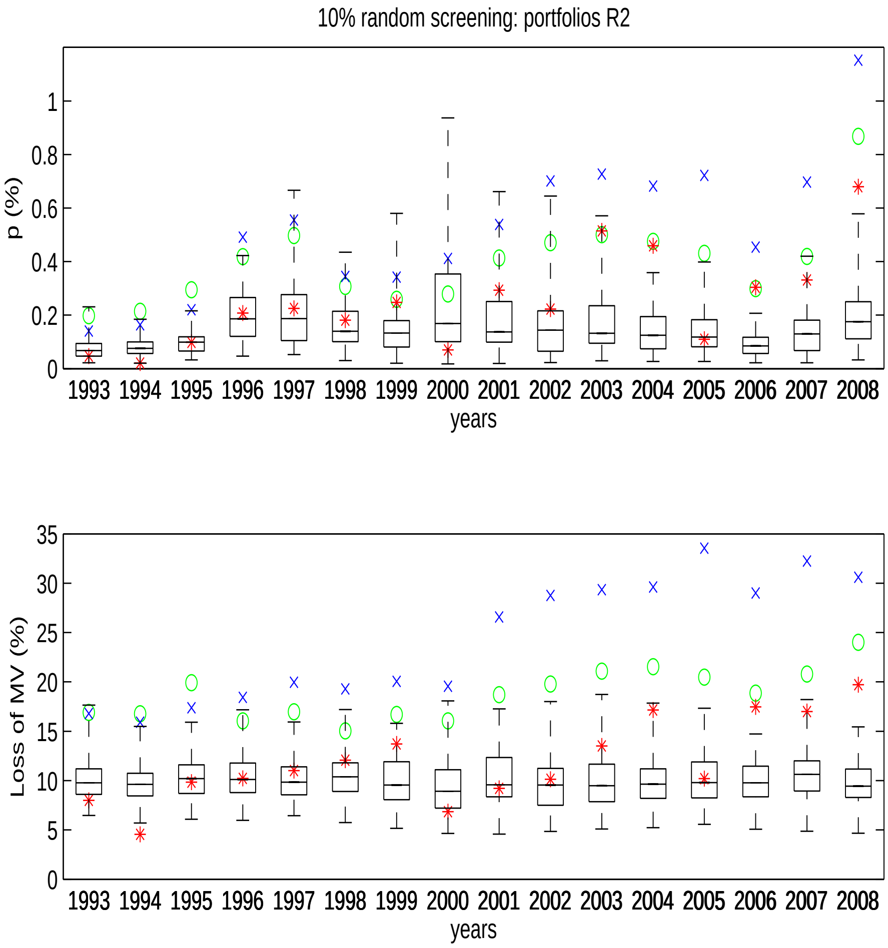

Figure 5 *Price of sustainability (top panel) and percentage of Market Value discharged by screening (bottom panel) for portfolios of type R2. The candles show the median, the 25-th and the 75-th percentiles of 100 samples obtained by a 10% random screening. The corresponding values for the non-random 10% partial screening are marked by a circle for the Environment, by a cross for the Social and by a star for the Governance.*

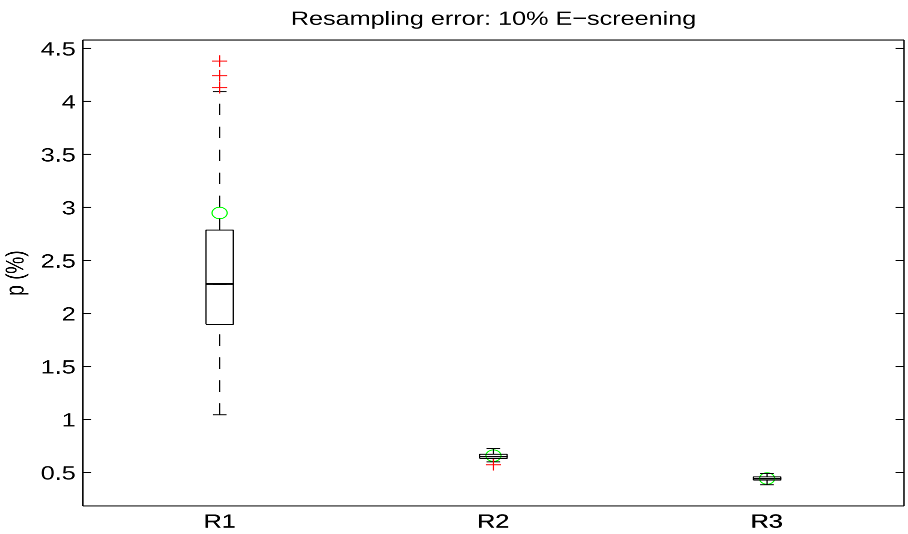

Figure 6 *Test of robustness for the* 10% *partial screening in the Environmental dimension. The candles represent the median, 25-th and 75-th percentiles while the crosses represent outliers of the sustainability prices for portfolios of type* $R1$ $R2$ *and* $R3$ *over 100 simulated scenarios. The circles correspond to the prices of sustainability of the* 10% *partial screening in the Environmental dimension.*

Spanning test

	E		S		G	
	ss	no ss	ss	no ss	ss	no ss
10% screening	16.82 (0.01) [***]	6.65 (0.12)	4.00 (0.68)	1.16 (0.65)	8.58 (0.20)	2.80 (0.42)
30% screening	37.80 (0.00) [***]	8.53 (0.06) [*]	11.93 (0.06) [*]	1.01 (0.66)	12.06 (0.06) [*]	3.86 (0.30)
50% screening	48.36 (0.00) [***]	7.36 (0.10) [*]	24.07 (0.00) [***]	1.58 (0.57)	38.43 (0.00) [***]	3.66 (0.33)
All-con screening	52.55 (0.00) [***]	10.56 (0.03) [**]	31.57 (0.00) [***]	1.43 (0.57)	45.57 (0.00) [***]	2.05 (0.50)

Table 2 *Spanning test for different types of screening and sustainability dimensions. The table represents the cases of short selling (ss) and of no short selling (no ss). In each case it is reported the Wald statistic ξ and the p-value (in parentheses). Rejection of the null hypothesis at the level of* 10%, 5% *and* 1% *is labeled by* $[*]$, $[**]$ *and* $[***]$, *respectively.*

References

[1] Amenc N., Le Sourd V. (2008), "Socially responsible investment performance in France", EDHEC Risk and Asset Management Research Centre

[2] Bauer R., Kees Koedijk and Roger Otten (2005), "International evidence on ethical mutual fund performance and investment style", "Journal of Banking & Finance", "29", 1751 - 1767"

[3] Black F., Litterman R. (1992), "Global Portfolio Optimization", Financial Analyst Journal, September/October, pp. 28-43

[4] De Roon F.A., Nijman T.E. (2001), "Testing for mean-variance spanning: a survay", Journal fo Empirical Finance, 8, pp. 111-155

[5] De Roon F.A., Nijman T.E., Werker B.J.M. (2001), " Testing for Mean-Variance Spanning with Short Sales Constraints and Transaction Costs: The Case of Emerging Markets", The Journal fo Finance, LVI, No.2 (April 2001), pp. 721-742

[6] Eugene F. Fama, Kenneth R. French (1992a), *The cross section of expected stock returns*, Journal of Finance 47, 427-465

[7] Eugene F. Fama, Kenneth R. French (1992b), *The economic fundamental of size and book-to-market equity,*

[8] Eugene F. Fama, Kenneth R. French (1993), *Common risk factors in the returns on stocks and bonds*, Journal of Financial Economics 33, 3-56

[9] Jobson, J.D., Korkie, B.M. (1981), "Performance hypothesis testing with the Sharpe and Treynor measures", Journal of Finance 36, 889-908

[10] D. A. Kodde, F. C. Palm (1986), "Wald criteria for jointly testing equality and inequality restrictions", Econometrica, Vol. 54, N.5, pp. 1243-1248

[11] Peter D. Kinder, Amy L. Domini (1997), "Social screening: Paradigms old and new", The Journal of Investing, Winter 1997, Pag 12-19

[12] L. Kurtz (1997), "No effect, or no net effects? Studies on socially responsible investing", The Journal of Investing , pp. 37-49

[13] G. Huberman, Kandel S. (Sep. 1987), "Mean-Variance Spanning", The Journal of Finance, XLII, No.4 , pp. 873-888

[14] O. Ledoit, M. Wolf (2008), "Robust performance hypothesis testing with Sharpe ratio", Journal of Empirical Finance, 15 850-859

[15] M. Statman, D. Gkushkov, B. Investors (2009), "The wages of social responsibility", Financial Analysts Journal, 65(4): 774-800

[16] Bernell K. Stone, John B. Guerard Jr (2001), "Socially responsible Investment Screening: Strong Empirical Evidence of No Significant Cost for Actively Managed Value-Focused Portfolios", Dept. of Finance. Provo, Brigham Young University