Style Consistency and Survival Probability in the Hedge Funds' Industry

P.-A. Barès[a], R. Gibson[b] and S. Gyger[a]

February 13, 2001

[a] *Institute of Theoretical Physics*
Swiss Federal Institute of Technology Lausanne EPFL
1015 Lausanne – Switzerland
E-mail / phone number : Sebastien.Gyger@epfl.ch   0041 21 693 34 17
E-mail / phone number : Pierre-Antoine.Bares@epfl.ch   0041 21 693 44 79

[b] *Swiss Banking Institute*
University of Zürich
8032 Zürich – Switzerland
E-mail / phone number : RGibson@isb.unizh.ch   0041 1 634 29 69

We gratefully thank the Institute of Theoretical Physics of EPFL and Olivier d'Auriol Asset Management SA (1009 Pully – Switzerland, `mailto:OlivierdAuriol@urbanet.ch`) for partial funding of this research project. The authors also acknowledge Financial Risk Management (FRM) for providing part of their hedge funds database and O. d'Auriol, J. Brandon, R. Chatagny and Prof. B. Liang for valuable comments and suggestions.

Style Consistency and Survival Probability in the Hedge Funds' Industry

Abstract

This study focuses on two problems that affect the choice of alternative investments, that is the style consistency of the manager and his survival probability.

We first present a new quantitative approach to describe fund managers style consistency. We show, through hard and fuzzy clustering, that the investment style of a manager may depart over time from his reported style.

Second, we apply a survival analysis method based on the Kaplan-Meier estimator, that takes into account the right-censorship of the data. A conditional survival analysis suggests that funds' investment styles, the size of their assets under management, their beta and style consistency can significantly affect their survival probability.

Quantitative measures of style consistency and conditional survival probabilities thus offer useful information when selecting and monitoring hedge fund managers.

I Introduction

Most hedge funds managers follow specialized strategies and restrict their activities to particular style investments. The strategy describes the objectives, the assets and the trading mechanisms adopted by a manager to add value when he creates and manages a portfolio.

Authors in the hedge funds' literature usually rely upon and condition their analysis on the trading styles described in the hedge fund industry when they analyze funds' performance and their survivorship bias (Ackermann, McEnally and Ravenscraft (1999), Agarwal and Naik (2000a), Brown, Goetzmann and Ibbotson (1999), Brown, Goetzmann and Park (1997), Liang (1999), Schneeweis and Spurgin (1997)). This approach seems reasonable enough since a manager should see no reason to expose himself and his clients to alternative processes he is not completely familiar with. Moreover, his disclosed strategy is checked and analyzed through thorough discussions and due diligence processes. Nevertheless, the managers' behavior often diverges from his reported strategy. For instance, it has been shown in Brown, Goetzmann and Park (1997) that good performers in the first half of the year reduce volatility of their portfolios in the second half, while poor performers increase it.

The main objective of this study is to propose a framework to study the style consistency and the survival probability of hedge fund managers. Secondly, we would like to determine whether style consistency is an important factor in conditioning the survival probability of hedge funds. It is however important to mention that many other factors can influence a hedge fund manager, such as their dominant investment style, their assets under management or their implicit directional exposure.

We thus introduce and examine a new feature of hedge funds investment practices : that is their style consistency over time, also known as style drift. Similarly to performance persistence, examined in past hedge funds studies (Agarwal and Naik (2000b), Brown, Goetzmann and Park (1997)), we believe that a style consistency analysis shall bring new perception to the hedge funds industry, which suffers from a notorious lack of transparency. Using clustering analysis, our first objective is to build reliable and representative hedge funds indices, that can be used as non-biased performance benchmarks. The formalism is then extended to fuzzy clustering using concepts borrowed from information theory. In this latter case, strict membership to a group is relaxed to allow the managers to be associated probabilistically to each cluster.

We introduce two measures to precisely quantify the consistency of a manager's strategy over time. In particular, we show that *Trading* as well as *Long/Short* are the most consistent though more volatile investement styles.

According to Fung and Hsieh (2000b), hedge funds' managers are generally reluctant to disclose their activities publicly, hence little is known about the risk exposures of hedge fund strategies. In that respect, our methodology answers part of the risk assessment issue and allows one to discriminate between managers who deviate and those who consistently follow a specified investment style.

Our second motivation is to examine the managers' survival probability conditioned on different quantities of interest, such as their style consistency, their assets under management and their implicit directional exposure. For that purpose, the right-censorship of the data is formally taken into account using Kaplan-Meier estimator to compute the proportion of funds surviving. We find no strong evidence of a difference in the survival function between managers running different strategies. On the opposite, managers who have either less assets under management or have a positive implicit directional exposure or whose dominant strategy is more consistent, are more likely to disappear. We also show that given two managers alive at time t, the younger one is less likely to disappear due to poor performance. This finding is not inconsistent with Agarwal and Naik (2000b) result which states that the younger the fund the more likely it is to disappear. As it turns out, we will see that the style consistency of the managers is a feature that significantly influence the survival probability of hedge fund managers.

The structure of the paper is the following : the review of the literature and the data description are to be found in Section II and III respectively. In Section IV, we present a hard clustering technique that leads to a unique partitioning of the managers into distinct groups. We end up with four different clusters and show that each one can be associated to a single style through a bijective relationship. Moreover, the formalism can be used to distinguish between managers who make use of leverage from those who do not. For instance, we observe that managers following an *Event Driven* strategy have a very conservative attitude towards the use of leverage. Further, we extend the analysis to fuzzy clustering. In the latter case, each manager is associated in probability to each cluster. Therefore, transitions of managers among clusters or investment styles can be precisely measured. Finally, conditional survival analysis is developed in Section V. The analysis suggests that the investment styles of the funds, the size of their assets under management, their beta and style consistency can significantly affect their survival probability. Concluding remarks are to be found in Section VI.

II Survey of the literature

Style analysis has gained much interest in recent hedge funds studies. Fung and Hsieh (1997a) show that hedge funds strategies are different from mutual fund strategies and highly dynamic. They propose additional style factors (trading strategy factors) to extend the traditional asset class factor model (location factors), in order to capture the stylistic features of alternative investments. Agarwal and Naik (2000c) propose a distinct model based on location factors enhanced by passive option strategies. In both cases, authors show that they are able to explain a significant amount of variation in the hedge fund returns over time.

The assumption under return-based style analysis is that the manager's investment style is consistent over time. We show in Section IV that, to some extent, hedge funds do not exhibit coherence over time with regard to their investment style commitment. Therefore, caution must be given to style analysis.

In their studies, Fung and Hsieh (1997a) and (1999) use principal component analysis to build five dominant investments styles, that are closely related to the commonly referred style categories.

Though their approach leads to very interesting and promising results, we believe that principal component analysis might not be entirely satisfactory when applied to hedge funds. Not only because it accounts for only part of the cross-sectional return variance of the sample (under 50% for the 5 dominant components) but especially because of its construction. Dominant principal components are orthogonal, i.e. not correlated, to each other and cannot be considered as investable quantities, since the normalization procedure[1] does not meet the requirements of an investable hedge funds portfolio or hedge funds index construction. To overcome that problem, Fung & Hsieh build five investable style factors (that satisfy the short selling constraint and the sum of weights equal to 1) that replicate as closely as possible (on a correlation basis) the five dominant principal components. Each of these investable style factors is next associated with one of the database provider qualitative style categories. This additional procedure leads to indices which might not preserve the main property of the principal components : that is their explanatory power through variance maximization, although the diversity of managers among style factors is certainly preserved.

In this study we are also concerned with style analysis. We focus on another widely used technique, called cluster analysis. Its explanatory power differs from the one obtained through principal component analysis. In multivariate statistics, the principal component method is used to find the linear combinations of statistical variables with maximum variances, while discarding those with small variances and hence reducing the number of variables to be treated. Those linear combinations stem directly from the diagonalization of the variables' covariance matrix[2]. The explanatory power of cluster analysis is weaker, since cluster analysis is used to assign the variables in different clusters, in such a way that two variables belonging to the same cluster are sufficiently "similar", whereas two variables lying in different clusters are dissimilar. The matrix of the similarity between each couple of variables is essential to clustering analysis. It is called distance or dissimilarity matrix and is measured in the d-dimensional space of observation[3]. A

[1]The normalization procedure for principal component analysis : $\{\vec{w} \in \mathrm{R}^N \mid \sum_{i=1}^N w_i^2 = 1 \text{ and } -1 \leq w_i \leq 1\}$ differs from the normalization procedure for hedge funds portfolio construction : $\{\vec{w} \in \mathrm{R}^N \mid \sum_{i=1}^N w_i = 1 \text{ and } 0 \leq w_i \leq 1\}$.

[2]If the variables are elliptically distributed, then statistical inference can be applied on the eigenvalues (the amount of variance that is explained) and eigenvectors (the rotated variables).

[3]The role of the variables and the observations is played by managers and their monthly returns respectively. In the following both terminologies are used without distinction.

detailed description will be given in Section IV. Cluster analysis is performed with an iterative algorithm that searches to minimize the average distance between each variable and its cluster representative (its mean or median for instance). The procedure is appealing since two managers using different strategies are expected to lie further apart than managers using the same one. Therefore, the latter will belong to the same group while the former will not. It has to be noted that cluster analysis is solely based on how far apart the variables lie from each other and thus does not require any prior assumption about the hedge funds returns distribution.

We then follow one of Fung and Hsieh (1999) objectives and form cluster based style indices that can easily be mapped into the data provider style indices.

Our second scope of interest lies in the survival probabilities of alternative investments. Survival has received considerable attention in former hedge fund's studies (Ackermann, McEnally and Ravenscraft (1999), Brown, Goetzmann and Ibbotson (1999), Brown, Goetzmann and Park (1997), Fung and Hsieh (1997b)), in which the attrition rate and the survivorship bias are computed. The attrition rate is defined as the percentage of funds which disappear in a given period (generally 1 year). It is documented in Brown, Goetzmann and Ibbotson (1999) that about 20% of the offshore funds disappear each year. The same drop out rate is found by Fung and Hsieh (1997b) for Commodity Trading Advisors (CTA), while Brown, Goetzmann and Park (1997) discovered an attrition rate of 15% in the hedge funds' industry.

The survivorship bias is the difference in returns between a portfolio that contains defunct funds from one that does not. It has been shown that hedge funds *ex-post* performance is overestimated in case one does not take into account funds that have disappeared. Survivorship bias for hedge funds is estimated around 3.0% per year over periods running from 1989 to 1995 and 1994 to 1997 by Fung and Hsieh (1997b) and Brown, Goetzmann and Park (1997) respectively. An exhaustive review of the literature may be found in Fung and Hsieh (2000a).

The procedure used in Brown, Goetzmann and Park (1997) consists in estimating the so-called fraction of funds surviving after t months by dividing the number of funds present in the database after t months by the total number of funds. A manager may choose no longer to report because he has reached his fund's target asset size while continuing to operate or he may no longer report due to poor performance and leave the business (Ackermann, McEnally and Ravenscraft (1999), FRM (1998)). The former is said to be right-censored (as well as the one still alive at the end of the survey), while in the sequel the latter is said to be dead or defunct. Both type of managers should not be treated the same way, since one continues to operate his fund, while the other has ceased his activity and therefore the above procedure is questionable. In the analysis of mutual funds, Lunde, Timmermann and Blake (1999) for example, uses a Cox regression model to estimate the relationship between a funds' age and its hazard rate. Other important factors that lead to premature disappearance, such as inferior relative performance, are also identified.

Our contribution follows a different path and formally takes into account the right-censorship of the data. We use a non-parametric method, based on Kaplan-Meier estimator, that has proved its efficiency in numerous research applications[4] and that is mathematically well suited to deal with right-censored data. The survival probabilities are then conditioned on several variables of interest. Our final objective is to examine whether these probabilities are significantly affected by the style consistency of the managers, the size of the assets under their management or their beta.

III Data description

FRM's hedge fund database contains data on 2992 funds at the end of April 1999, managed by more than 1500 managers[5]. We present in Table 1 the number of funds, the assets under management, the management and incentive fees as well as a summary of each style category monthly returns properties. Comprehensive information for each manager, including detailed strategy descriptions and historical performance is stored in the database. The vast majority of the information is obtained directly from the funds or their administrators.

[4] Medicine, biology, public health, epidemiology, engineering and demography.

[5] Financial Risk Management (FRM) is an independent research- based investment services company, specializing in constructing portfolios of hedge funds to achieve absolute return investment objectives.

Hedge fund managers employ a wide variety of methods for generating returns. In this study, we label *styles* or *strategies* the core characteristic of the hedge fund return generating processes. The style describes the methodology that managers follow when creating and managing their portfolios. FRM uses six distinct designations, namely : *Trading* (468), *Long and Short Market Hedged* (660), *Event Driven* (273), *Relative Value* (377), *Market Directional* (331) and *Multi-Strategy* (199). Each of them[6] represents a common investment style, that usually encloses lower level style characteristics. For instance, *Trading* groups *Discretionary, Macro* and *Systematic Trading* subcategories.

Managers are reporting their monthly performance on a fund-by-fund basis in a *net return* form. The *net return* of a fund over a period is defined as the change in the fund's value over that period, as a percentage of the starting value of the fund, adjusted for subscription and redemption, after periodic fees have been charged.

Hedge funds' managers usually charge two kind of fees : a management fee and an incentive fee. The management fee is based on a percentage of the assets in the funds. 1725 funds out of 2021 that report management fee charge between 1% and 2% each year (mean = 1.4%, median = 1.3%). The incentive fee gives the hedge fund manager a percentage of the profits earned by his fund. 1369 funds out of 1975 that report incentive fee charge 20% annually (mean = 18.1%, median = 20.0%), while 1345 are subject to high-water mark conditions. The use of a high water mark requires a manager to perform above the highest previous level before earning additional incentive fees.

The assets under management for a fund are defined as the net assets managed by the fund manager and invested in the specified fund. Until April 1999, 1681 funds have reported their assets under management and totalized more than $190 billions. We find that 131 hedge funds out of 2308 (5.7%) have ceased to report performance in their first year of activity[7]. Hedge funds may stop reporting for several reasons. First, because the fund has reached its target asset size while continuing to operate, second, because a fund may have left the business due to poor performance and third because the manager decides to concentrate his activity on his *offshore* fund instead of the *onshore* one (for legal and tax purposes).

Our database keeps track of funds, organized in different structures (*offshore* or *onshore*) and run by the same manager. This is a significant feature since it allows us to link appropriate funds between them. On the other hand, it does not distinguish between funds that stop reporting due to self-selection or discontinuation (to use Ackermann, McEnally and Ravenscraft (1999) terminology). Similarly to other hedge funds data collectors, FRM has recognized the importance of collecting and have kept records of defunct funds since 1994.

We decide to consider the manager rather than the fund as the standard quantity of interest. First, because he is the one who might be hired for his potential skills and second because he might have managed several funds in his career. Consequently, funds managed by the same manager with the same strategy are linked together from start to finish, with the objective of monitoring a good performance track record, as long as possible for statistical accuracy. Managers are reporting their performance on a monthly basis and all figures are net of fees. We impose a performance track record of a least 36 months[8] for a manager to enter our statistical studies in the sequel. Moreover we focus on managers with a primary non or slightly directional hedged strategy (*Multi-strategy* and *Directional* strategies are left aside).

Figure 1 shows the number of managers, classified according to the four major non or slightly directional FRM style's categories, namely : *Trading, Long/Short, Event Driven* and *Relative Value.* Because of their atypical market exposure as well as a favorable financial environment, hedge funds have grown exponentially in both size and number since 1990. We note that *Relative Value* is the style that has received the most attention since its number of managers has increased

[6] In brackets the number of hedge funds (for a total of 2308), following a specific strategy, reporting on a monthly basis with at least 1 return observation.

[7] This rate is much smaller than those computed for other databases (see section II), and is comparable with the 5% drop out in mutual funds.

[8] This window width of 36 months is a compromise between two mutually exclusive requirements : the selection of a great number of managers and the display of relevant statistical time-varying estimates. For comparison purposes, Fung and Hsieh (1997a), Fung and Hsieh (1997b) conducted their analysis on 297 funds that had returns over a common 36 months period.

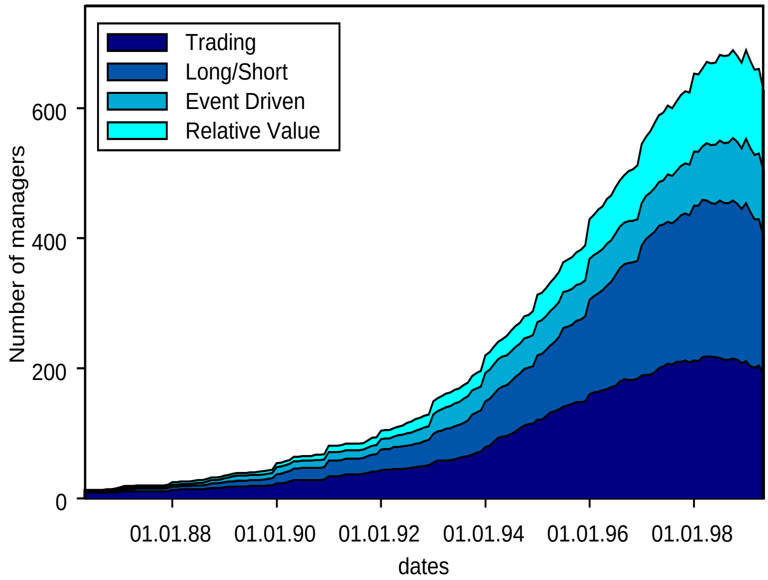


Figure 1: Number of managers in time, classified according to their specified strategy (with at least a 3 year track record in the FRM database) .

from 27 in January 1994 to 124 in April 1999. The slight decrease of managers' number from September 1998 is twofold. It is first due to negative events that affected the entire hedge funds community (Emerging Markets crisis : Russian and Asian) and second to the time it takes to incorporate new managers into the database.

IV Manager style consistency

In this Section, we apply cluster analysis to study the style consistency of hedge funds managers.

Cluster analysis is a descriptive tool that tends to define groups embedded in a data set. Notice first that it does not require any prior knowledge or assumption on the returns distribution and second that it is solely based on the managers' performance track records, as measured by his monthly net returns. This framework seems well adapted to our needs since two managers using different strategies are expected to lie further apart (in their d- dimensional space of returns) than two managers using the same one. Therefore, the former will belong to distinct groups while the latter will belong to the same one. Briefly stated, conditioned on d monthly returns measured for n managers, the whole procedure consists of dividing the basket of managers into k different groups through the minimization of a cost function. To our knowledge, such a procedure has never been applied to analyze hedge funds strategies. The cost function is specific to the type of clustering. Hard and fuzzy clustering along with their related cost functions are further described in Sections IV.A and IV.B respectively.

A natural application of hard clustering is the computation of reliable hedge funds indices. The important issue of leverage will also be investigated in the context of index building. For its part, fuzzy clustering allows the careful analysis of the managers style's consistency. We define two measures in Section IV.C that may prove useful to (in)validate the time consistency of the managers disclosed strategy and help investors faced with an investment decision.

IV.A Hard Clustering

We assume n managers represented by d monthly returns. This data set can be organized in a n by p matrix $X = (x_{if})_{i=1,\ldots,n}^{f=1,\ldots,d}$. The degree of dissimilarity between managers is the variable of interest, hence the necessity to define a distance between managers. Natural candidates are the $L1$ *Manhattan* and the $L2$ *Euclidian* distances that define the dissimilarity between managers i and i' as :

$$d_1(i,i') = \sum_{f=1}^{d} |x_{if} - x_{i'f}| \ ; \ d_2(i,i') = \sum_{f=1}^{d} (x_{if} - x_{i'f})^2 \ ; \ i,i' = 1,\ldots,n. \tag{1}$$

It must be emphasized that $L1$ is the less sensitive measure to outliers, compared to other Lq *Minkowsky* distances. Therefore, it is a robust and reliable distance when computing dissimilarities between managers[9].

Partitioning clustering methods, that divide the data according to natural classes present in it, have been used in a large variety of scientific disciplines and engineering applications (see Blatt, Wiseman and Domany (1997), Kaufman and Rousseeuw (1990) and references therein). A well known partitioning method is the K-means method. In the K-means algorithm, group membership is determined by calculating the multidimensional version of the mean for each group and assigning each object to the group with the closest mean. Means are calculated using least-squares and this makes K-means less resistent to outliers than the partitioning around medoids method. The algorithm PAM (partitioning around medoids), which we apply in this Section, is outlined underneath :

- random selection of k representative managers or cluster's centers (often called medoids),
- allocation of each remaining manager to the nearest medoid.

The optimal selection of k representative managers is the one that minimizes the cost function, which is defined as the average distance between each manager and its cluster center. This algorithm is a hard clustering technique since each object belongs to exactly one cluster by construction[10]. The algorithm output is a clustering vector $\vec{v}$ of size n, which component $v_i = j$ if manager i belongs to cluster j.

At time t we select all managers, $n(t)$ in number, who do possess a performance track record of at least 36 months, i.e. a window width of $d = 36$ monthly observations (see footnote 8). We choose to divide this population of managers into $k = 4$ distinct clusters in order to examine the consistency between our clustering method and the four FRM style indices.

We find that both the choice of the $L1$ distance and a normal standardization of the returns[11] leads to the best correspondence between FRM classification and our cluster division. It can be seen on Figure 2 that there is a strong one-to-one relationship between each cluster and each strategy : 93% (71%) of the managers who belong to cluster 1 (2) are running a *Trading* (*Long/Short*) strategy. Though clusters 3 and 4 are weakly defined clusters, their greatest proportion of managers (respectively 40% and 41%) are classified under *Event Driven* and *Relative Value* respectively.

The geometric structure of the clusters is displayed in Table 2. The diameter of the cluster is defined as the maximal dissimilarity between two objects of the same cluster and its separation is the minimal dissimilarity between one object of the cluster and one object of another one. The average silhouette width for each cluster, abbreviated silhouette in Table 2, measures the extent

[9] The distance between managers may be related to correlation through :

$$d(i,i') = \frac{1-\rho(i,i')}{2} = \begin{cases} \simeq 0 & \text{if managers are strongly positively correlated} \\ \simeq 1/2 & \text{if managers are not correlated} \\ \simeq 1 & \text{if managers are strongly negatively correlated.} \end{cases}$$

That choice requires the additional computation of a n by n correlation matrix using $L2$ distance which is sensitive to outliers. We focus on *Manhattan* and *Euclidian* distances only.

[10] A more detailed description of the algorithm may be found in Kaufman and Rousseeuw (1990) for instance.

[11] In order to account for leverage, we follow Fung and Hsieh (1997a) and standardize the returns for each manager so that they all have a mean 0 and a variance 1. This is a zero order approximation that tends to reduce differences in variances caused by heterogeneity in leverage levels.

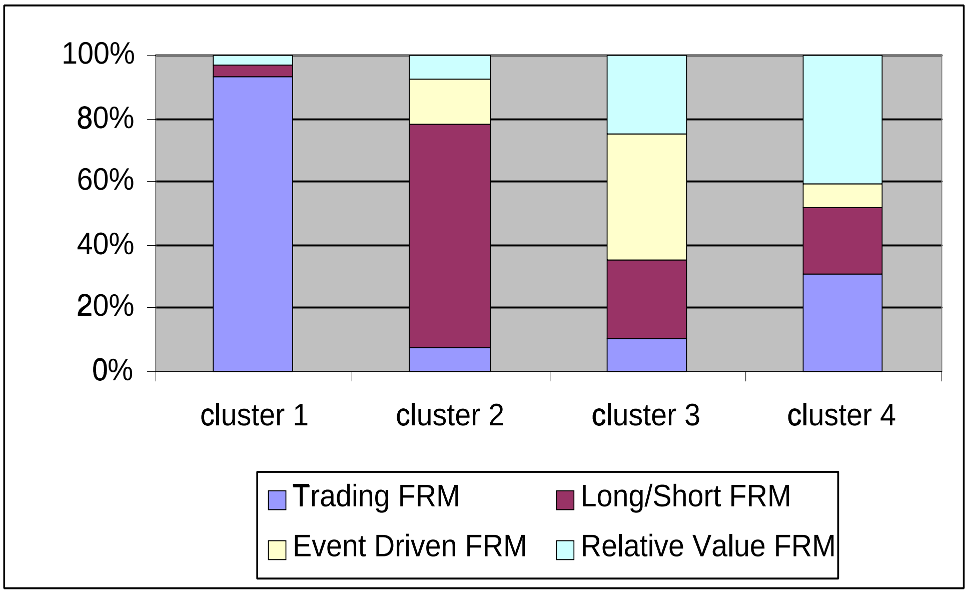


Figure 2: Proportion of managers who belong both to a specific FRM strategy and to a given cluster (for a 3 year period starting in May 1996 and ending in April 1999).

to which a cluster is clearly defined amongst the whole dataset[12]. Clusters 1 to 4 are ordered according to their average silhouette width. According to Kaufman and Rousseeuw (1990), the average silhouette width for the entire dataset $\overline{s}_k = \sum_{i=1}^{n} s(i)$, where $s(i)$ is defined in footnote (12), can be used for the selection of an ideal value of k the number of clusters, by choosing k for which $\overline{s}_k$ is as high as possible. We have checked that $\overline{s}_{k=4} > \overline{s}_{k>4}$, which further justifies our choice of dividing the managers into 4 different clusters.

Another way to look at the strong connection between clusters and strategies is to compute cluster based indices. For that purpose, we assign all managers who belong to the same cluster to an equally weighted[13] cluster based index (at time t) :

$$\vec{I}_j = \frac{1}{\sum_{i=1}^{n} \delta_{j,v_i(t)}} \sum_{i=1}^{n} \delta_{j,v_i(t)}\, \vec{x}_i \;;\; \delta_{j,v_i(t)} = \begin{cases} 1 & \text{if manager } i \text{ belongs to cluster } j \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\vec{I}_j$ and $\vec{x}_i = (x_{if})_{f=1,\dots,d}$ are vectors of length $d = 36$ and $\sum_{i=1}^{n} \delta_{j,v_i(t)}$ is the number of manager who belong to cluster j. The returns $\vec{x}_i$ are not standardized in this context.

The comparison between the FRM hedge funds style indices and the hard cluster based indices is very encouraging. It can be seen in Figure 3 that 3 of our indices have a strong linear relationship with those computed by FRM (For the first 3 clusters, a least-square regression gives $R^2 = 0.95,\ 0.94,\ 0.84$ respectively. On the other hand, cluster 4 and the strategy *Relative Value* have the weakest, yet still significant, linear relationship[14]. Besides, we see that only 3 out of 36 returns have negative values and that all non-normalized monthly returns are enclosed in a restricted interval $[-3.20\%, 2.33\%]$. Those are typical features of the returns achieved through a *Relative Value* strategy (see FRM (1998), Nicholas (1999)). In light of Figures 2 and 3 we can therefore match each hard cluster with a single style, namely : Cluster 1 $\longleftrightarrow$ *Trading*, Cluster 2 $\longleftrightarrow$ *Long/Short*, Cluster 3 $\longleftrightarrow$ *Event Driven* and Cluster 4 $\longleftrightarrow$ *Relative Value*.

[12] The silhouette width of each object i is defined as $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, where $a(i)$ is the average dissimilarity of i to all other objects of the same cluster and $b(i)$ is the minimum dissimilarity of i to all objects that do not belong to the same cluster. The value $s(i)$ may be interpreted as follows : (i) if $s(i) \simeq 1$ then object i is well classified, (ii) if $s(i) \simeq 0$ then object i lies between 2 clusters and (iii) if $s(i) \simeq -1$ then object i is badly classified.

[13] In contrast, Fung and Hsieh (1997a) compute the weights that maximize the linear correlation between the return index and its related principal component.

[14] A least-square regression gives $R^2 = 0.17$. The F-statistic ($= 7.13$ on 1 and 34 degrees of freedom) is significant at a 5% confidence level since the p-value equals 0.01.

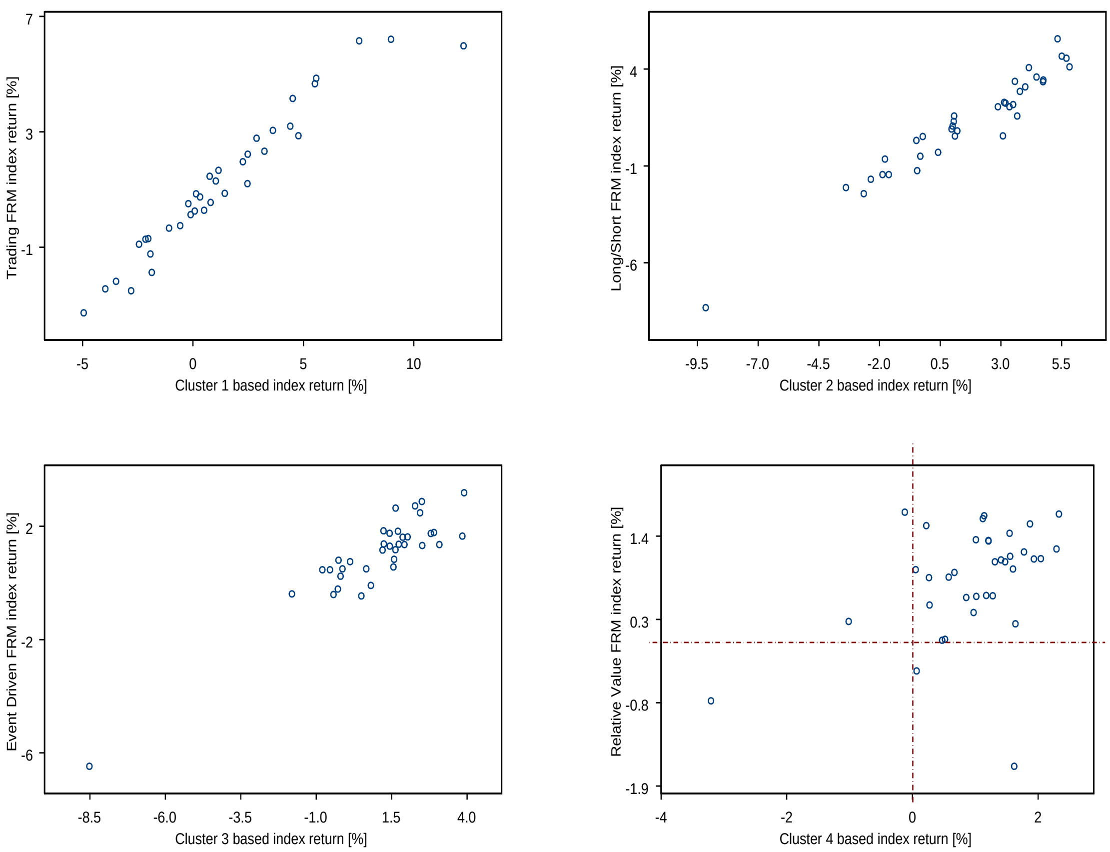


Figure 3: Comparison between monthly cluster based indices and FRM indices over a 3 year period starting in May 1996 and ending in April 1999.

In comparison, Fung and Hsieh (1997a) investigate the disclosure documents of the funds and associate their style factors with the commonly used qualitative style categories.

At this stage, we are able to accommodate managers in four different groups, each of them actually representing one of the FRM style categories. This procedure can prove useful to verify that a manager is classified correctly, to reclassify him if necessary and possibly to examine whether each manager's behaviour and strategy are transparent enough and consistent over time. It is important to contrast our index construction method from the one presented in Fung and Hsieh (1997a), since the former only assigns managers to distinct clusters, based on the distance between the managers' returns. Though the explanatory power of cluster analysis is reduced compared to principal component analysis, we have shown its benefit when constructing hedge funds indices. Another useful application is presented in the next Section where we investigate the leverage in the hedge funds' industry.

IV.A.1 Leverage

When a manager borrows to increase the amount of money invested in a position, he is using leverage. Leverage magnifies both potential gains and losses and therefore enhances the risk of the strategy[15]. It is nonetheless a considerable ingredient in some managers return generating process and demands careful examination by the investor or analyst, when either selecting hedge fund's managers or performing qualitative due diligence.

[15] As illustrated by Long Term Capital Management (LTCM) and others.

Assuming that managers are employing the same strategy to generate their returns, hard clustering may also prove useful to differentiate between managers who make use of different leverage levels.

In April 1999, we run the algorithm PAM on each of the four distinct clusters previously built. The $L1$ distance is selected, while the returns are not standardized in this case. We choose $k = 3$ to define the number of sub-clusters in each cluster, in order to differentiate across different leverage levels.

Within each group, we compute 3 distinct sub-indices, using the same technique as the one described in Section IV.A, and then regress them on their related FRM hedge funds style index. In each sub-group, the slope of the regression is computed. It is a measure of the perceived average leverage level amongst the managers. Those levels are labeled *low, medium* and *high*.

Results are displayed in Table 3, where the 12 linear regressions are computed[16]. All figures, except the two in brackets, are significant at a 99% confidence level.

From Table 3 we see that the large majority of *Event Driven* managers are very conservative toward leverage, as documented in the literature (Nicholas (1999) for instance). On the contrary, *Trading* exhibits the largest leverage range (4:1) of all styles[17].

IV.A.2 Dynamic index construction

In this Section, we account for the passage of time and build the time series of cluster indices. The dynamic construction of the cluster based indices is accomplished as follows. We slide the window a month at a time backwards and select $n(t')$ managers at time t'. We repeat this for all t' until $t' = t_0$ the inception date. The dynamic indices at time t are simply defined by $J_j(t) = I_j(t)\,; j = 1, \ldots, 4$, where $I_j(t)$ is the last component of $\vec{I_j}$ the static index built at time t. Two of the resulting dynamic indices, namely *Trading* and *Long/Short*, are displayed in Figure 4, from which one can infer the robustness of the procedure (Least-square regressions of 2 dynamic cluster based indices on *Trading* and *Long/Short* FRM hedge funds indices give $R^2 = 0.94,\ 0.73$ respectively).

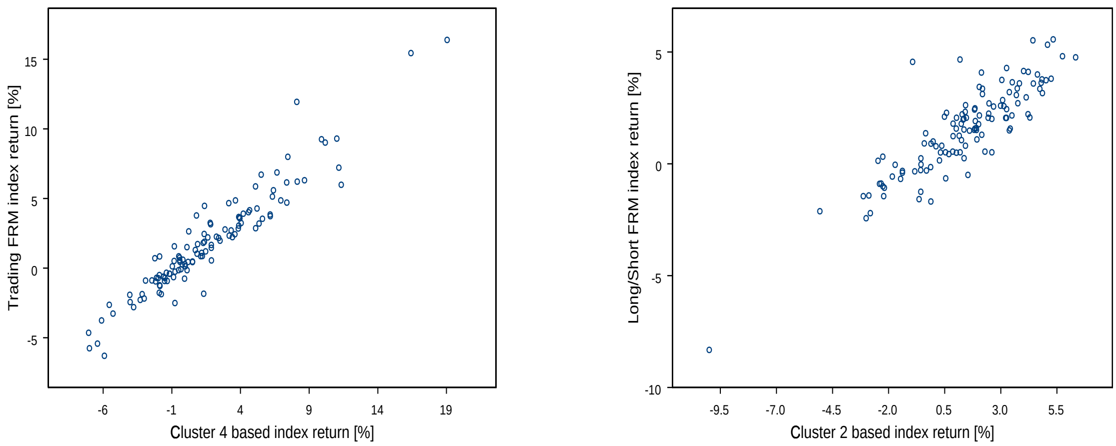


Figure 4: comparisons between cluster based indices and FRM indices for two particular styles *Trading* and *Long/Short* over a 10 year period starting in May 1989 and ending in April 1999.

The difference between the clustering at time t and clustering at time $t' = t - \Delta t$ is twofold. First, the number of selected managers entering the algorithm has changed because new ones have appeared and others have left the database (see Figure 1). In the second place, the performance record is shifted by Δt months. Hence, a manager might belong to cluster i at time t and to cluster $j \neq i$ at time $t' \neq t$. However, caution must be given to the fact that successive records of a manager's membership to a given cluster are not independent due to the window sliding construction.

[16]The intercept is fixed at 0 and # refers to the number of managers.

[17]Note that the lowest and highest slopes are not significant at a 99% confidence level and only concern a tiny number of managers (14 out of 629).

A change of cluster might be due, for instance, to a change in the manager's strategy implementation. It may also result from the potential difficulties he may encounter to meet his objectives, relative to the rest of the industry.

In light of the manager's commitment to a single strategy, the abrupt switching from one cluster to another is unlikely. We would rather expect a smooth transition from one cluster (style) to another. We therefore describe in the next Section a probabilistic approach to quantify the transition among clusters.

IV.B Fuzzy Clustering

We start with the assumption that a manager does not necessarily belong with probability one to a single cluster. Instead, one defines by p_{ij}, the probability for manager i to belong to cluster j, and uses a fuzzy clustering technique to compute the p_{ij}s' estimate. It is important to notice, that hard and fuzzy clustering techniques usually connect all managers in a complex and non-linear way. In particular, it can be seen in the Appendix Eq.(16), that the fractional membership of a single manager to a fuzzy cluster depends on the complete set of the selected hedge funds' managers returns. Clustering methods contrast, for instance, with regressions procedures where response variables are regressed one at a time on fixed predictors.

We follow Rose, Gurewitz and Fox (1990) and defer to the Appendix a thorough description of the main procedure. We use Shannon's maximum entropy criterion and apply it to our specific data clustering problem. The managers' selection process as well as the dynamic index construction methods remain as described in Section IV.A.

We focus again on the $L1$ distance when defining the cost function, whose robustness was already stated in Section IV.A. The $L1$ is more convenient than the $L2$ distance, because in the latter case one is more likely to obtain a solution of the type, $p(i \in C_j) \simeq 1$ and $p(i \in C_k) \simeq 0,\ k \neq j$. This feature is less significant using a linear cost function, because its growth is slower (see Eqs.(14 and (15) in Appendix VII).

In order to analyze the fuzzy clustering results displayed in Table 4, we shall first compare in April 1999 the fuzzy clustering with the FRM main strategies (top of Table 4) and second with the hard clusters identified in Section IV.A (bottom of Table 4). We count managers who are associated both with fuzzy cluster j[18] and are either reporting one of FRM main investment style categories (top) or are belonging to hard cluster j' (bottom). As an illustration, 108 managers belong at once to fuzzy cluster 1 and FRM style *Trading*, while 129 belong to fuzzy cluster 4 and hard cluster 4. In the first (second) case, managers are taken into account only if their probability to belong to a specific fuzzy cluster is greater than 0.5 (0.4). We see that the majority of the managers are located on the diagonals of Table 4[19]. We are thus in a position to see the one-to-one correspondence between the dominant fuzzy clusters, the hard clusters and the FRM investment style categories.

Our contribution to the statistical study of the manager's strategy is therefore the following : we are able to split the manager's dynamic trading style into four distinct categories. The probability $p_{ij}(t)$ is a natural measure of manager's i inclination to follow style j at time t. A significant issue, which is answered in the next Section, is to determine whether or not a manager consistently follows over time the strategy he claims to run.

IV.C Strategies' style consistency

We conclude our application of clustering techniques to the hedge funds industry with the formulation of two consistency measures, that allow us to examine whether a manager displays a consistent style over time. First, we define the *mean time consistency* as :

$$\varrho_1(i) = \max_{j=1,\dots,4}\left(\frac{1}{T}\sum_t p_{ij}(t)\right) = \max_{j=1,\dots,4}\overline{p}_{ij}. \tag{3}$$

[18]We only select managers with a probability $p_{\cdot j} \geq 0.5$ (left figures) or $p_{\cdot j} \geq 0.4$(right figures).

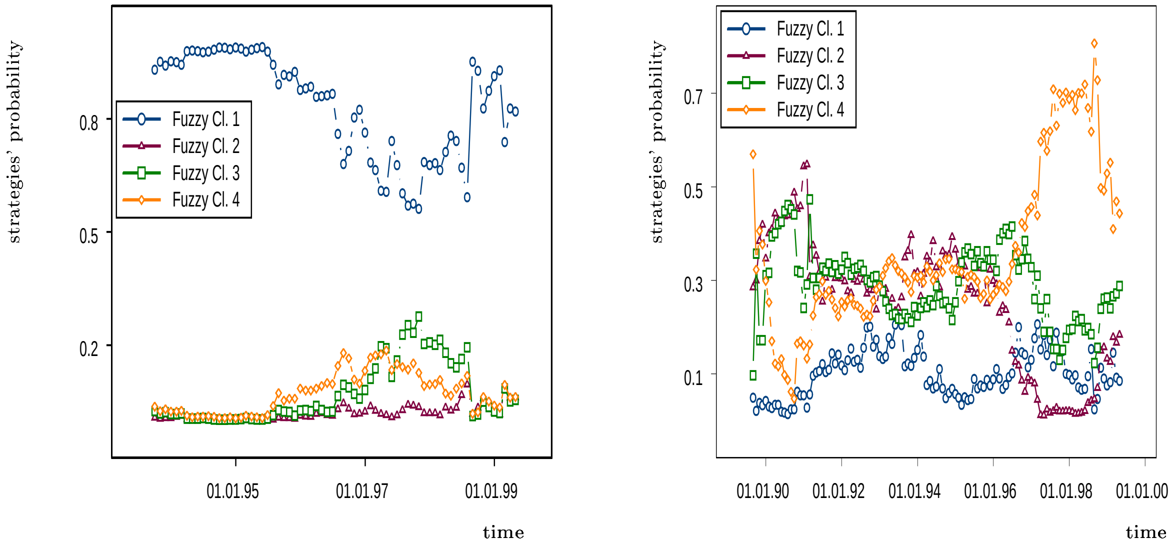
[19]The unavoidable discrepancies (in the outer diagonals) result from a single factor : algorithms leading to hard and fuzzy clustering are essentially different. Actually, the only managers who will belong with almost certainty to fuzzy cluster j and also belong to hard cluster $j' = j$ are those lying in the vicinity of the cluster's center of mass.

It measures the extent to which the managers' dominant strategy probability, say $j' \in \{1,2,3,4\}$ matches its ideal value 1, where T is the total number of probability observations for manager i. The second measure, we label *consistency standard deviation* or *volatility*, estimates its variability over time :

$$\varrho_2(i) = \sqrt{\frac{1}{T}\sum_t (p_{ij'}(t) - \varrho_1(i))^2} = \sigma(p_{ij'}). \tag{4}$$

Note that $\varrho_2(i)$ is upper-bounded and that perfect uniformity is equivalent to $\varrho_2(i) = 0$. We may think of $\varrho_1(i)/\varrho_2(i)$ as an appraisal ratio. The larger this ratio, the more coherent is manager's i dominant strategy.

We illustrate in Figure 5 the change in probability for two distinct managers. The consistency



Trading manager | *Relative Value* manager

Figure 5: Strategy decomposition in 4 fuzzy clusters for two different managers.

measures give respectively $\varrho_1 = 0.84$ (0.36) and $\varrho_2 = 0.14$ (0.17) for a manager in the left (right) panel. Both of them run the strategy they claim to have adopted (*Trading* and *Relative Value* respectively). However, the *Trading* manager displayed on the left panel in Figure 5 shows much more style consistency than the *Relative Value* manager.

In Table 5, we illustrate two applications based on the consistency measures previously defined. In the first one, we select the dominant strategy for all managers, determine the FRM style they claim to follow and count the number of occurrences. In the second one, we compute the mean of both consistency measures for each dominant strategy. As previously indicated (see Section IV.B), each dominant strategy can be associated, on a bijective relationship, with one of FRM four major style categories. In the following, we therefore also use FRM's styles designation for fuzzy clusters. To avoid any misinterpretation, we relate to FRM when necessary.

A general feature emerges from the results displayed in Table 5. We see that managers proned to *Trading* or *Long/Short* strategy are generally more consistent, though more volatile (according to their ϱ_2), than *Event Driven* or *Relative Value* managers[20]. Besides, in the long run, a significant amount of managers have a pronounced tendency not to fulfill their strategy commitment (see Table 5). We see that about 20% (64/317) of the managers classified by FRM under *Long/Short* exhibit return profiles that are nearing those obtained by a *Relative Value* approach. *Relative Value* and *Long/Short* managers with no bias towards a net market exposure are indeed both targeting market

[20]This characteristic is independent of the definition in Eq.(4) for the dispersion of consistency. We obtain comparable results if we take $\varrho_2 = \max_t p_{ij'}(t) - \min_t p_{ij'}(t)$ instead of the standard deviation for instance.

neutrality. Therefore their styles have a tendency to merge together. The proportion of managers classified by FRM under *Trading*, who do not display evidence of their specific skills, reaches a high 48% (133/279). They are, for the most, immersed in the *Relative Value* style. Finally, the boundary between *Event Driven* and *Relative Value* is not clearly defined, since numerous managers are classified under one of them while most of the time following the other one. Notice that *Merger Arbitrage* is an example of a lower level style category that is ambiguous to classify under *Event Driven* or *Relative Trading*.

We also notice[21] that none of the selected managers whose approach is defined to be *Long/Short* with a short bias (i.e. with a net short market exposure) display a significant *Long/Short* feature. Instead, the *Relative Value* style dominates. Though they are not necessarily bearish about the market, the very few number of managers with a short bias had certainly to reduce their net short market exposure to survive in the long run. This has important implication as to their relative performance with respect to a peer group.

The next Section measures conditional survival probabilities, criteria that may prove useful for the managers' selection process. As we have already said in the introduction, style consistency, in particular, will be used as a conditioning variable when studying the survival probabilities of hedge fund managers.

V Survival analysis

As mentioned in Section II, hedge fund managers may experience distressing situations that compel them to stop their activity. The consequences for investors are generally harmful, since the poor or negative performance of such managers' funds affects their own portfolios. In this Section, we present a method that accounts for the right-censorship of the data. A manager is said to be right-censored either if he is alive at the end of the survey (April 1999) or if he left the database while continuing to operate.

The same sample of managers, that met the selection criterion described in Section I is considered for this analysis. Remember that they have a performance track record of at least 36 months. While this condition leaves unexperienced managers out of the survival study, it certainly reduces backfilling bias[22] and allows the conditioning of the results on our previous statistical classification of the managers' styles and styles consistency.

In Section V.A we first define the survival and hazard rate functions. Section V.B provides some empirical evidence about the relationship between managers' tendency to disappear and some attributes of interest, such as their style, the size of their assets under management, their implicit direct exposure as well as their style's consistency.

Finally, in order to preserve comparison with other hedge funds databases[23] survivorship bias, we compute the survivorship bias of the hedge funds' industry associated with the FRM database.

V.A Survival probabilities

We now present the Kaplan-Meier estimator of the survival probability. A survival function $S(t)$ defined over time t is the probability that an individual[24] survives at least until time t. More formally, let T be a positive random variable with distribution function $F(t)$ and density $f(t)$:

$$S(t) = 1 - F(t) = P(T > t).$$

For practical purposes, we also define the hazard rate function as :

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t), \tag{5}$$

[21] Figures on the short bias traders are not reported for the sake of brevity but can be obtained form the authors upon request.

[22] Certain managers enter the database along with an hypothetical historical performance, see Ackermann, McEnally and Ravenscraft (1999), Fung and Hsieh (2000a) for a description of the different sources of bias.

[23] The two major data suppliers are HFR (Hedge Fund Research Inc.) and Tass (TASS Management Limited).

[24] By individual we mean an object or a person which may experience a sudden disappearance.

which gives the probability that an individual dies during the next small interval of time, given that he has survived until time t. A common estimate of the survival distribution is the Kaplan-Meier estimate (Elandt-Johnson and Johnson (1980)), defined as :

$$\hat{S}(t) = \prod_{t_i<t} \frac{r(t_i) - d(t_i)}{r(t_i)} \;;\; \hat{S}(0) = 1, \tag{6}$$

where $t_1 < t_2 < \cdots < t_m$ denote m distinct death times, $r(t_i)$ and $d(t_i)$ are the number at risk, i.e. the number of individuals present in the survey, and the number of deaths at time t_i respectively. A better understanding of Eq.(6) is achieved through conditional probabilities

$$\begin{aligned} P(T > t_i) &= P(T > t_{i-1}) \cdot P(T > t_i \,|\, T > t_{i-1}) \\ P(T > t_i \,|\, T > t_{i-1}) &= \frac{r(t_{i-1}) - d(t_{i-1})}{r(t_{i-1})}. \end{aligned}$$

V.B Applications

We are first interested in computing the survival probability of the hedge funds managers using the framework described in the previous Section. In a second step, we condition survival analysis on several variables of interest.

Unfortunately, the FRM database does not provide information on the reasons why reporting ends (see Section II). According to Fung and Hsieh (2000a), funds that are liquidated performed substantially worse than funds that became defunct for other reasons. Therefore, to differentiate between them we decided to compute their monthly average Sharpe ratio[25] over respectively the last and the last two years of activity. If either one or both of these ratios are inferior to 0.2, we assume that the manager has ceased his activity due to a lack of performance. We find that 629 managers are alive in April 1999, 89 voluntarily stopped reporting and 162 disappeared due to poor performance[26].

In the first place, we compute the survival function for the totality of the selected managers. It is displayed in Figure 6. The survival function for a survival time smaller than 80 months is almost linear, with a constant attrition rate of about 5% per year. The hazard rate function tells us (see Eq.(5)) that given two managers alive at time t, the older one is more likely to disappear due to inferior performance[27]. This has important implication when selecting managers and minimizing the survival risk of a portfolio of hedge funds. Since 13 managers only left the business with at least 80 months of survival, the right part of the curve in Figure 6 is not dense enough to be statistically significant.

The result is not inconsistent with Agarwal and Naik (2000b) statement that the younger the fund the more likely it is to disappear. It is of different nature, since we are dealing with conditional probabilities and do constrain the managers to enter the study with at least 36 month of reported performance.

Survival functions can be conditioned on different variables of interest. The first variable under examination is the type of dominant strategy followed by a manager, as computed using Eq.(3), while the second, third and fourth variables discriminate between managers who respectively have small/large assets under management, implicit directional or non-directional exposure and finally those who have small/large style "inconsistency" with respect to their dominant strategy. The directional exposure is measured as the linear coefficient (or beta) of the regression of the manager's net monthly returns on the Russel 2000 index return over the last 12 months of activity[28]. The inconsistency of the dominant strategy is measured by the consistency standard deviation measure ϱ_2 defined in Eq.(4). The medians for these variables are computed and each manager is classified, depending upon whether he is above or under the median, in two different groups, big or small respectively.

[25] The risk free interest rate is set at 0.

[26] If we select 0.1 or 0.3 as the monthly average Sharpe ratio threshold, neither of the forthcoming results nor conclusions are affected.

[27] If $S(t) = 1 - at$ then $\lambda(t) = \frac{a}{1-at}$ for $t \in [0, 1/a]$ is an increasing function of t.

[28] A longer regression period (24 months) or/and another market index (MSCI World) do not alter our forthcoming conclusions.

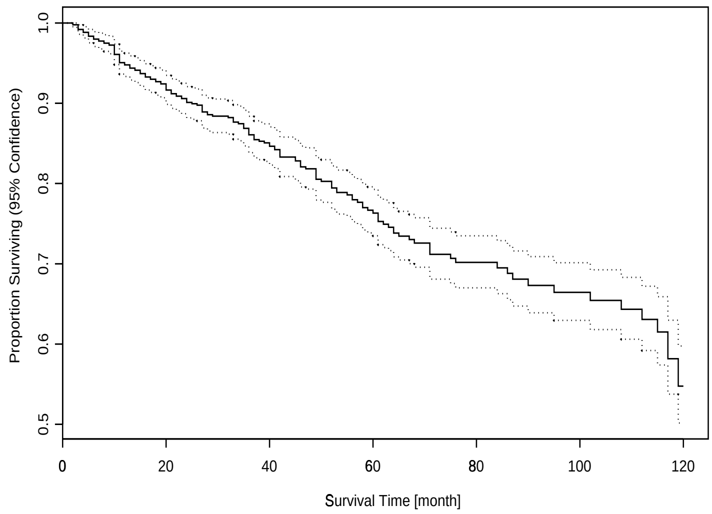


Figure 6: Proportion of surviving managers with a 95% confidence interval (dotted lines), using Kaplan-Meier estimate defined in Eq.(6).

The four sets of empirical results for the four partitionings are reported in Figure 7. The top left panel displays each of the four dominant fuzzy cluster based styles' tendency to disappear. We note that *Trading* and *Relative Value* define the upper and lower boundary respectively, while *Long/Short* and *Event Driven* are mostly comprised in between. There is slight evidence (at a 95 % confidence level) that the strategies differ in their survival probabilities[29]. Indeed *Relative Value* has the highest attrition rate whereas *Trading* displays the lowest. The results in the top-right panel suggest that managers with a small amount of assets under their management[30] are more likely to cease their activity. The same statement holds for managers with a large implicit directional exposure as measured by their β (see bottom left panel). The survival functions are in both cases significantly different for any confidence level[31].

We conclude this Section by looking at the extent to which inconsistency in the dominant strategy affects a manager's survival probability. The bottom right panel in Figure 7 shows that managers with a consistency volatility measure ϱ_2 above the median are expected to disappear less rapidly than managers below the median[32]. This demonstrates that managers who are more flexible and who rely more heavily on style drift when creating, managing and rebalancing their portfolios are less inclined to disappear due to inferior performance.

From the four criteria that significantly act on the attrition rate, the style inconsistency is the most prominent one. The implicit directional exposure and the assets under management are also discriminating variables though they exhibit less pronounced attrition characteristics. The dominant strategy is the least important attribute.

V.C Survivorship bias for the FRM database

The last issue we would like to discuss is the survivorship bias encountered in the FRM database. We follow Brown, Goetzmann and Ibbotson (1999), Fung and Hsieh (1997b) and define two equally weighted portfolios. One which contains only managers who are alive at the termination of the study and the other with both alive and defunct managers. We use Fung and Hsieh (2000a) terminology and label them *surviving* and *observable* portfolios respectively. The survivorship bias is defined as the average monthly difference return between each other.

From January 1996 to April 1999[33], the *observable* portfolio returned on average 1.23% per

[29] A Chi-squared test gives : 7.8 on 3 degrees of freedom and a p-value of 0.049. The hypothesis H_0 (the distributions are different) is accepted at a 95% confidence level.

[30] Note that 135 out of 880 managers do not report their assets under management.

[31] A Chi-squared test gives respectively : 7.8 and 13.2 on 3 degrees of freedom and a p-value of 0.00.

[32] A Chi-squared test gives : 76.6 on 3 degrees of freedom and a p-value of 0.00.

[33] Most of the disappearing funds ended during this 3 year period.

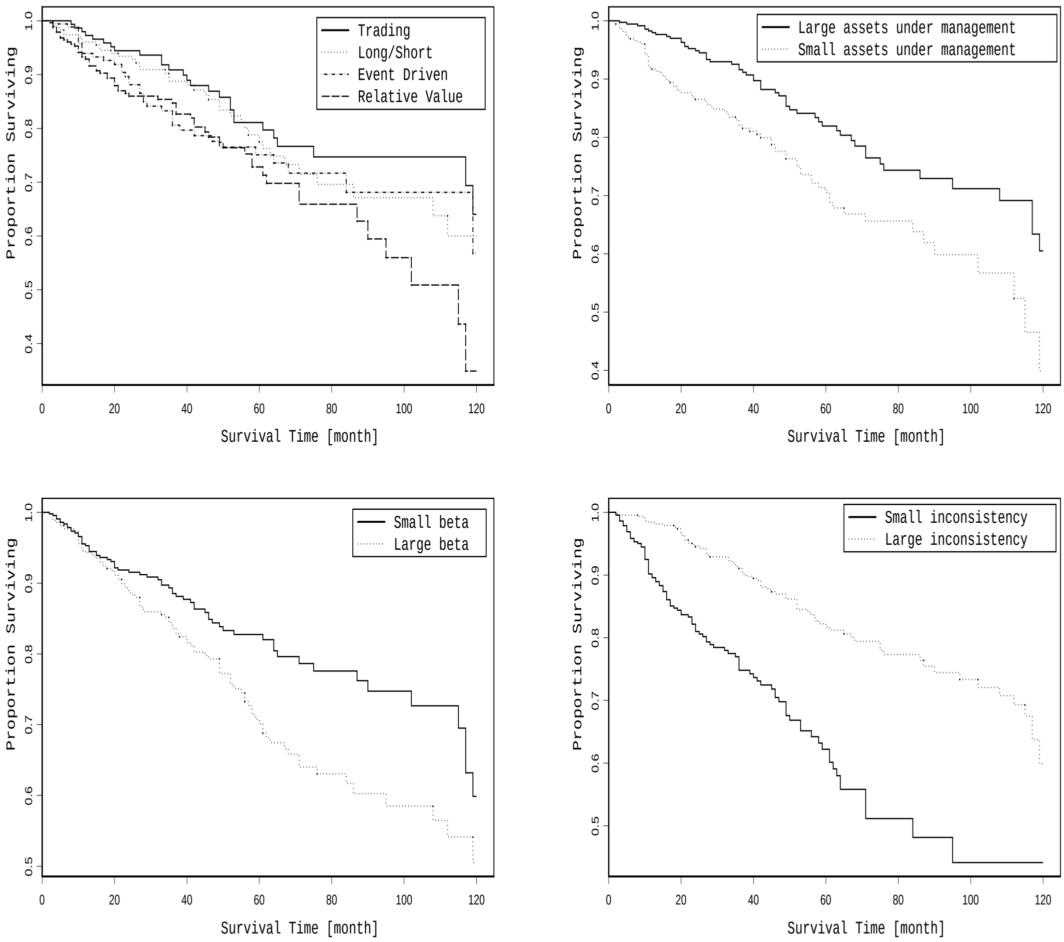


Figure 7: Proportion of surviving managers conditioned on : (i) the four major strategies, (ii) the assets under management, (iii) the beta and (iv) the strategy inconsistency measured with ϱ_2. Managers are classified in 2 different groups (big/small) depending upon whether they are above or under the median of the variables of interest.

month, while the *surviving* portfolio yielded 1.34%. The annual survivorship bias is hence 1.30%.

To study the incidence of the selection criteria, described in section III, on this result, we first narrow the window width to one year and second enable all style categories to enter the analysis, namely : *Trading*, *Long and Short Market Hedged*, *Event Driven*, *Relative Value*, *Market Directional* and *Multi-Strategy*. For the same period, the annual survivorship biases are slightly increased to 1.67% and 1.50% respectively .

In comparison, Brown, Goetzmann and Ibbotson (1999), Fung and Hsieh (1997b) estimate the average survivorship bias to be 3.0% per year over periods running from 1994 to 1998 and 1989 to 1995 respectively. Besides, Brown, Goetzmann and Ibbotson (1999) found no evidence of inferior performance in hedge funds that left the database due to poor performance compared to others that left for other reasons, in contrast to Fung and Hsieh (1997b). In our case, managers with an inferior performance, as measured by a Sharpe ratio lower than 0.2 (0.3, 0.1 respectively) average a net monthly return of 0.351% (0.204%, 0.032%), while those who intentionally choose not to report anymore average a net monthly return of 2.120% (1.980%, 1.879%). This partially explains the reduced difference between our *observable* and *surviving* portfolios : the latter seems to be amputated from above average performers who stopped reporting even though they are still active.

VI Conclusion

In this article, we investigate two problems that affect hedge funds asset allocation strategy, namely their style consistency and their survival probability.

First, we use a hard clustering method to compute hedge funds indices and the managers' leverage level. We find that part of the managers do not consistently follow a pre-specified investment style. We then take advantage of fuzzy clustering to assess the extent of each manager's commitment to a specific strategy over time. Two consistency measures are introduced to quantify the coherence of a manager's style. We show, for instance, that a significant part of *Trading* managers actually shift to the *Relative Value* category, based on a return profile that does not correspond to the most representative peer managers.

In a second step, we use a Kaplan-Meier based estimator in order to conduct a survival analysis of hedge funds' managers. We show that of the four major management styles, *Relative Value* exhibits a slightly stronger tendency to vanish sooner than the others. In addition, style's inconsistency is a strongly discriminating variable : managers who are more flexible and change their style, are less likely to cease their activity. We also find that a smaller size of assets under management as well as a higher beta lead to a significant larger probability for a manager to disappear.

Conditional survival rate measurement and fuzzy clusters based style analysis shall hopefully contribute to enlarge the set of quantitative tools with which one can analyze the hedge funds' industry and implement asset allocation strategies for hedge funds portfolios.

VII Appendix

In this Appendix, we detail the fuzzy clustering approach introduced in Section IV.B. We first recall the basic concept of Shannon's entropy, following Jaynes (1993). According to Shannon's entropy principle and based on Rose, Gurewitz and Fox (1990), we then present the fuzzy clustering model. It leads to an analytical expression for the most probable set of cluster probabilities association and we derive the most likely set of clusters for the $L1$ distance. Finally, we exhibit in Figure 8 the clustering phase diagram of 629 managers in April 1999.

VII.A Information entropy

Every probability distribution has some "uncertainty" associated with it. The concept of information entropy or Shannon's entropy is introduced to provide a quantitative measure H of this uncertainty.

Let the probabilities of n possible outcomes $A_1, \ldots, A_n$ of an experiment be respectively $p_1, \ldots, p_n$. A consistent measure of the amount of uncertainty carried by a probability distribution has to satisfy some well established axioms (see Jaynes (1993)) and is given by :

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i. \tag{7}$$

The function H is Shannon's measure of uncertainty and is generally called the information entropy.

The distribution $(p_1, \ldots, p_n)$ that maximizes Eq.(7), taking into account all the information we have (constraints for instance), represents the most honest description of the probability distribution. It does indeed contain as much uncertainty as possible while being consistent with all the available information.

VII.B Model description

Clustering methods are major tools for the analysis of data with poor distribution's informations. In fuzzy clustering, the main principle is that each point of the data set is associated in probability with a cluster. In this study, the only measurable quantity, for a given clustering configuration, is its associated cost. The objective within this probabilistic framework is to find an optimal probability distribution of association. We use the Shannon's entropy introduced in Section VII.A.

We denote $\vec{x} \in \mathbb{R}^d$ a variable which can take n different discrete values $(\vec{x}_1, \ldots, \vec{x}_n)$ and $(\vec{y}_1, \ldots, \vec{y}_k)$, $\vec{y}_j \in \mathbb{R}^d$ a set of k parameters.

We assume that the association between a point $\vec{x}_i$ and a cluster $\vec{y}_j$ is demanding : the cost is denoted by $E(\vec{x}_i, \vec{y}_j) \geq 0$. Hence, the average total cost for a given configuration of clusters $\{Y\} = (\vec{y}_1, \ldots, \vec{y}_k)$ is given by :

$$\langle E\rangle(\{Y\}) = \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\, E(\vec{x}_i, \vec{y}_j), \tag{8}$$

where p_{ij} is the probability for point $\vec{x}_i$ to belong to cluster j (defined by $\vec{y}_j$). Next we require each point to be distributed among every cluster with certainty. Therefore n additional constraints follow:

$$\sum_{j=1}^{k} p_{ij} = 1, \ \forall i = 1, \ldots, n. \tag{9}$$

Subject to constrains (8) for the average total cost and (9) for normalization, the variational problem for p_{ij} is :

$$\delta \left[H(p) - \sum_{i=1}^{n} (\lambda_i - 1) \left(\sum_{j=1}^{k} p_{ij} - 1 \right) - \gamma \left(\sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\, E(\vec{x}_i, \vec{y}_j) - \langle E\rangle_{\{Y\}} \right) \right] = 0$$

or

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left[\frac{\partial H}{\partial p_{ij}} - (\lambda_i - 1) - \gamma E(\vec{x}_i, \vec{y}_j)\right]\delta p_{ij} = 0,$$

where $H(p) = -\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij}\,\log p_{ij}$ is the information entropy. From Eq.(7) :

$$\frac{\partial H}{\partial p_{ij}} = -\log p_{ij} - 1$$

and the solution is thus described by the decreasing exponential :

$$p_{ij} = \exp\left(-\lambda_i - \gamma\, E(\vec{x}_i, \vec{y}_j)\right) = \frac{\exp\left(-\gamma\, E(\vec{x}_i, \vec{y}_j)\right)}{Z_{\vec{x}_i}}, \tag{10}$$

where $Z_{\vec{x}_i} = \sum_{j=1}^{k} \exp\left(-\gamma\, E(\vec{x}_i, \vec{y}_j)\right)$ is a consequence of the normalization constraints (9). Notice that the Lagrange multiplier γ is intimately connected with $\langle E\rangle_{\{Y\}}$ through Eqs.(8) and (10) (i.e. γ is determined once $\langle E\rangle_{\{Y\}}$ is fixed).

We have computed the association probability between each point and each cluster for a fixed set of clusters $\{Y\}$. In the following, the most likely set of clusters is derived.

For that purpose, notice that fuzzy clustering is a generalization of hard clustering. In hard clustering, each point is associated with a single cluster :

$$q_{ij} = \begin{cases} 1 & \text{if point } \vec{x}_i \text{ belongs to cluster } \vec{y}_j \\ 0 & \text{otherwise.} \end{cases}$$

The cost of such a hard clustering configuration is :

$$E(\{Y\}, \{Q\}) = \sum_{i=1}^{n}\sum_{j=1}^{k} q_{ij}\, E(\vec{x}_i, \vec{y}_j). \tag{11}$$

Fuzzy clustering enters the procedure at this stage, where we aim at computing the most likely set of clusters, and the marginal probability $P(\{Y\}) = \sum_{\{Q\}} p(\{Y\}, \{Q\})$ is considered. Since we avoid any assumptions on the distribution of the data, we apply the maximum Shannon's entropy criterion to state that :

$$p(\{Y\}, \{Q\}) = \frac{\exp\left(-\gamma\, E(\{Y\}, \{Q\})\right)}{\sum_{\{Y'\},\{Q'\}} \exp\left(-\gamma\, E(\{Y'\}, \{Q'\})\right)}, \tag{12}$$

which is the probability of a configuration whose cost is given by $E(\{Y\}, \{Q\})$. The combination of Eqs.(11) and (12) leads to :

$$P(\{Y\}) \propto \sum_{\{Q\}} \exp\left(-\gamma\, E(\{Y\}, \{Q\})\right) = \prod_{i=1}^{n}\sum_{j=1}^{k} \exp\left(-\gamma\, E(\vec{x}_i, \vec{y}_j)\right).$$

Maximum likelihood of the marginal probability leads to :

$$\frac{\partial \log P(\{Y\})}{\partial \vec{y}_j} = 0$$

or

$$\frac{\partial\left[\sum_{i=1}^{n} \log \sum_{l=1}^{k} \exp\left(-\gamma\, E(\vec{x}_i, \vec{y}_l)\right)\right]}{\partial \vec{y}_j} = 0. \tag{13}$$

Eqs.(10) and (13) determine the most probable set of cluster association as will be seen in section VII.D, where the solutions are computed for two different kind of cost functions.

In the next Section we discuss the role of the Lagrange multiplier γ.

VII.C Qualitative discussion

The Lagrange multiplier γ may be seen as the free parameter of the system. It is indeed determined by the value of the average total cost given in Eq.(8). For $\gamma = 0$, all $\vec{y}_j$ collapse to $\vec{y}_{CM}$ the global center of mass of the data set and therefore each point belongs to each cluster with probability $1/k$. At some positive γ, the original cluster splits into smaller ones and thus undergoes a phase transition. The new clusters will then split at higher γ so that the process may be viewed as a sequence of phase transitions. Finally, for $\gamma \to +\infty$ each point is its own center of mass and belongs with certainty to exactly one cluster with probability 1. In this asymptotic case, fuzzyness is completely removed of the system.

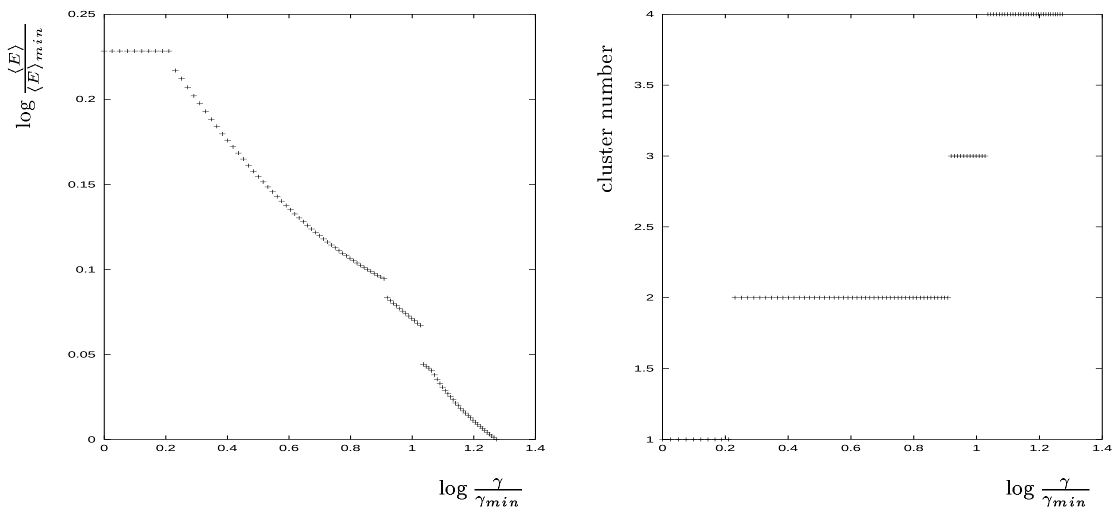


Figure 8: Phase diagram (up to four distinct clusters) for 629 managers on April 1999.

VII.D Solutions

In order to proceed and find the most likely set of clusters, we have to specify $E(\vec{x}_i, \vec{y}_j)$ the cost for associating a point with a given cluster. The quadratic nature of the squared euclidean distance is appealing. It is defined as :

$$E_2(\vec{x}_i, \vec{y}_j) = |\vec{x}_i - \vec{y}_j|^2 = \sum_{f=1}^{d} |x_i(f) - y_j(f)|^2 \tag{14}$$

and leads to a coupled system of equations (see Rose, Gurewitz and Fox (1990)) :

$$\begin{aligned} \vec{y}_j &= \frac{\sum_{i=1}^{n} \vec{x}_i \, p_{ij}}{\sum_{i=1}^{n} p_{ij}} \\ p_{ij} &= \frac{\exp\left(-\gamma \, E_2(\vec{x}_i, \vec{y}_j)\right)}{\sum_{p=1}^{k} \exp\left(-\gamma \, E_2(\vec{x}_i, \vec{y}_p)\right)}. \end{aligned} \tag{15}$$

We detail the calculations in a less straightforward case, namely :

$$E_1(\vec{x}_i, \vec{y}_j) = \sum_{f=1}^{d} |x_i(f) - y_j(f)|.$$

Eq.(13) leads to :

$$\frac{\partial F}{\partial y_p(l)} = \sum_{i=1}^{n} \frac{\exp\left(-\gamma \, E_1(\vec{x}_i, \vec{y}_p)\right)(2\theta_{x_i(l)<y_p(l)} - 1)}{\sum_{j=1}^{k} \exp\left(-\gamma \, E_1(\vec{x}_i, \vec{y}_j)\right)} = 0 \, ; \; \forall p, l,$$

where θ is the Heaviside function and $F = \sum_{i=1}^{n} \log \sum_{l=1}^{k} \exp\left(-\gamma\, E(\vec{x}_i, \vec{y}_l)\right)$.

In vectorial form :

$$\frac{\partial F}{\partial \vec{y_p}} = \sum_{i=1}^{n} \frac{\exp\left(-\gamma\, E_1(\vec{x}_i, \vec{y}_p)\right) \left[(2\vec{\theta}_{\vec{x}_i < \vec{y}_p} - \vec{1} + \vec{y_p}) - \vec{y_p} \right]}{\sum_{j=1}^{k} \exp\left(-\gamma\, E_1(\vec{x}_i, \vec{y}_j)\right)} = 0\,;\ \forall p.$$

We add and subtract $\vec{y_p}$ to reach the same structure as the one displayed in Eq.(15) :

$$\begin{aligned} \vec{y}_j &= \frac{\sum_{i=1}^{n} (2\vec{\theta}_{\vec{x}_i < \vec{y}_p} - \vec{1} + \vec{y_p})\, p_{ij}}{\sum_{i=1}^{n} p_{ij}} \\ p_{ij} &= \frac{\exp\left(-\gamma\, E_1(\vec{x}_i, \vec{y}_j)\right)}{\sum_{p=1}^{k} \exp\left(-\gamma\, E_1(\vec{x}_i, \vec{y}_p)\right)}. \end{aligned} \tag{16}$$

Notice that coupled system of the form (15) or (16) are generally solved using fixed point iteration procedures. Other choices for the cost function generally lead to more complex sets of equations.

References

Ackermann C., R. McEnally and D. Ravenscraft (1999), "The Performance of Hedge Funds : Risk, Return and Incentives", *Journal of Finance*, **54**, 833-874.

Agarwal V. and N.Y. Naik (2000a), "On Taking the 'Alternative' Route : Risks, Reward, Style and Performance Persistance of Hedge Funds", *Journal of Alternative Investments*, **2**, 6-23.

Agarval V. and N.Y. Naik (2000b), "Multi-Period Performance Persistance Analysis of Hedge Funds", forthcoming in *Journal of Financial and Quantitative Analysis.*

Agarval V. and N.Y. Naik (2000c), "Performance Evaluation of Hedge Funds with Option-based and Buy-and-Hold Strategies", *Working paper.*

Blatt M., S. Wiseman and E. Domany (1997), "Data Clustering Using Model Granular Magnet", *Neural Computation*, **9**, 1805-1842.

Brown S.J., W.N. Goetzmann and J. Park (1997), "Conditions for Survival : Changing Risk and the Performance of Hedge Fund Managers and CTAs" , *Review of Financial Studies*, **5**, 553-580.

Brown S.J., W.N. Goetzmann and R.G. Ibbotson (1999), "Offshore Hedge Funds : Survival and Performance 1989-1995", *Journal of Business*, **72**, 91-117.

Financial Risk Management Ltd. and Goldman Sachs and Co. (1998), "Pension and Endowment Forum : Hedge Funds Demystified".

Fung W. and D.A. Hsieh (1997a), "Empirical Characteristics of Dynamic Trading Strategies : The Case of Hedge Funds", *Review of Financial Studies*, **10** 275-302.

Fung W. and D.A. Hsieh (1997b), "The Information Content of Performance Track Records : Investment Style and Survivorship Bias in Historical Returns of Commodity Trading Advisors", *The Journal of Portfolio Management*, **24**, 30-41.

Fung W. and D.A. Hsieh (1999), "A Primer on Hedge Funds", *Journal of Empirical Finance*, **6**, 309-331.

Fung W. and D.A. Hsieh (2000a), "Performance Characteristics of Hedge Funds and Commodity Funds : Natural Versus Spurious Biases", *Journal of Quantitative and Financial Analysis*, **35**, 291-307.

Fung W. and D.A. Hsieh (2000b), "The Risk in Hedge Fund Strategies : Theory and Evidence from Trend Followers", forthcoming in *Review of Financial Studies.*

Jaynes E.T. (1993), "Probability Theory : The Logic of Science", *http:// omega.albany.edu:8008/JaynesBook.html.*

Elandt-Johnson R.C. and N.L. Johnson (1980), "Survival Models and Data Analysis", *Wiley.*

Kaufman L. and P.J. Rousseeuw (1990), "Finding Groups in Data : An Introduction to Cluster Analysis", *Wiley.*

Liang B. (1999), "On the Performance of Hedge Funds", *Financial Analysts Journal*, **55**,72-85.

Lunde A., A. Timmermann and D. Blake (1999), "The Hazards of underperformance : A Cox Regression Analysis", *Journal of Empirical Finance*, *6*, 121-152.

Nicholas J.G. (1999), "Investing in Hedge Funds", *Bloomberg Press.*

Rose K., E. Gurewitz and G.C. Fox (1990), "Statistical Mechanics and Phase Transitions in Clustering", *Physical Review Letters*, **65**, 945-948.

Schneeweis T. and R. Spurgin (1997), "Comparisons of Commodity and Managed Futures Benchmark Indices", *Journal of Derivatives*, 33-50.

Panel A

Category	# managers		Assets ($\$10^6$)	
	until April 99	in April 99	mean	median
Trading	468	297	202	179 (370)
Market Directional	331	172	186	40 (222)
Long/Short Market Hedged	660	398	153	35 (493)
Event Driven	273	195	116	51 (191)
Relative Value	377	252	140	49 (268)
Multi-Strategy	199	147	101	21 (137)

Panel B

Category	Incentive Fee [%]		Management Fee [%]	
	mean	median	mean	median
Trading	20.4	20.0 (390)	2.1	2.0 (395)
Market Directional	14.5	20.0 (245)	1.3	1.3 (271)
Long/Short Market Hedged	18.9	20.0 (595)	1.2	1.0 (595)
Event Driven	19.2	20.0 (255)	1.4	1.5 (255)
Relative Value	19.3	20.0 (334)	1.3	1.0 (332)
Multi-Strategy	10.0	10.0 (156)	1.5	1.5 (173)

Panel C

Category	Net monthly returns summary statistics [%] 05.1989–04.1990				
	mean	median	stdev	min	max
Trading	1.65	1.17 (463)	2.82	-5.10	15.59
Market Directional	1.31	1.67 (331)	3.62	-17.07	10.32
Long/Short Market Hedged	1.59	1.85 (656)	1.99	-8.52	5.40
Event Driven	1.35	1.42 (271)	1.47	-5.45	5.69
Relative Value	1.06	1.24 (374)	0.78	-2.99	2.81
Multi-Strategy	1.04	0.97 (198)	2.44	-7.76	16.47

Table 1: This table provides a summary statistic for the FRM database. It displays in Panel A the number of managers and the assets under their management; in Panel B, the incentive and management fees; in Panel C, the mean returns, the medians, the standard deviations, the minimum and maximum monthly returns for each of the 6 major FRM investment styles from 05.1990 to 04.1999. In brackets the number of managers who enter the calculations. Monthly returns are provided by FRM or computed from the net asset value (NAV) of the funds. Financial Risk Management is an independent research-based investment services company, specializing in constructing portfolios of hedge funds to achieve absolute return investment objectives (http://www.frmhedge.com). FRM has developed proprietary databases, processes and systems for fund identification and evaluation, for portfolio construction, and for monitoring funds and portfolios. As an investor, FRM gains information and performance data from many funds that do not provide this type of information to many of the publicly available databases. FRM's hedge fund database currently contains data on over 3500 funds managed by more than 1500 management groups. FRM's hedge fund database stores extensive information for each manager, including detailed strategy description, historical performance and due diligence reports.

	Size (# of managers)	Diameter	Separation	Silhouette
Cluster 1	119	49.49	5.82	0.20
Cluster 2	190	47.47	10.30	0.09
Cluster 3	153	50.00	9.29	0.07
Cluster 4	167	51.41	5.82	-0.02

Table 2: Geometric properties of the clusters built on a 3 year period starting in May 1996 and ending in April 1999 and comprising 629 managers. The diameter, separation and silhouette are dimensionless quantities.

Category	Low			Medium			High		
	#	slope	R^2	#	slope	R^2	#	slope	R^2
Trading	33	0.51	0.63	46	1.24	0.97	40	2.26	0.82
Long/Short	104	0.81	0.89	54	1.49	0.93	32	2.00	0.79
Event Driven	140	0.95	0.88	12	3.44	0.57	1	5.30	(0.14)
Relative Value	13	-0.44	(0.00)	83	0.79	0.53	71	1.29	0.50

Table 3: Linear regression analysis of each of the 12 cluster based sub-indices on the predictors given by FRM four major hedge funds style indices over a 3 year period starting in May 1996 and ending in April 1999. The number of managers (#), the slope of the regression as well as the goodness of fit (R^2) are displayed. The slope is a measure of the leverage level. Managers are sorted according to their different leverage levels, labeled as *low, medium* and *high*, based on a second hard partitioning of the managers. The figures that are not significant at a 99% confidence level are in brackets.

Number of managers ↘	Fuzzy Cl. 1	Fuzzy Cl. 2	Fuzzy Cl. 3	Fuzzy Cl. 4
Trading FRM	**108 / 117**	7 / 11	0 / 9	16 / 38
Long/Short FRM	15 / 18	**91 / 116**	5 / 34	11 / 26
Event Driven FRM	0 / 0	10 / 23	**11 / 45**	11 / 24
Relative Value FRM	2 / 6	5 / 9	0 / 33	**42 / 74**
Number of managers ↘	Fuzzy Cl. 1	Fuzzy Cl. 2	Fuzzy Cl. 3	Fuzzy Cl. 4
Hard Cl. 1	**110 / 116**	0 / 0	0 / 0	0 / 0
Hard Cl. 2	0 / 0	**108 / 141**	1 / 33	0 / 2
Hard Cl. 3	0 / 0	4 / 18	**15 / 83**	7 / 30
Hard Cl. 4	15 / 25	0 / 0	0 / 5	**73 / 129**

Table 4: Comparison between fuzzy clusters and the FRM major strategies on top and hard clusters underneath, over a 3 year period starting in May 1996 and ending on April 1999 (629 selected managers). The probability cut-off is set at 0.5 and 0.4 respectively for left / right figures.

	Fuzzy cluster			
	1 (Trading)	2 (Long/Short)	3 (Event Driven)	4 (Relative Value)
Trading FRM	146	23	14	96
Long/Short FRM	13	175	65	64
Event Driven FRM	0	31	54	30
Relative Value FRM	3	14	52	100
Mean value $\overline{\varrho_1(i)}$	0.71	0.59	0.43	0.48
Mean value $\overline{\varrho_2(i)}$	0.13	0.11	0.08	0.10

Table 5: Number of managers classified under a certain investment style (with fuzzy clustering method and FRM's specification) and average values of the consistency measures.