UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

NONLINEAR, NONPARAMETRIC METHODS IN EMPIRICAL FINANCE

Thèse présentée à la Faculté des Sciences Economiques

Institut de Statistique

Université de Neuchâtel

pour l'obtention du grade de docteur en statistique

par

Donatien Tafin Djoko

Acceptée par le jury de thèse :

Prof Catalin Starica, Université de Neuchâtel, directeur de thèse

Prof Yves Tillé, Université de Neuchâtel, président du jury

Prof Tommaso Proietti, Università degli Studi di Roma (Italy), Rapporteur

Prof Eric Hjalmarsson University of Goteborg (Denmark), Rapporteur

Prof Jens-Peter Kreiß, Technische Universität Braunschweig (Germany)**,** Rapporteur

Soutenue le 25, Novembre 2013

Faculté des Sciences Economiques
Avenue du 1er-Mars 26
CH-2000 Neuchâtel
www.unine.ch/seco

LETTRES ET SCIENCES HUMAINES
SCIENCES
DROIT
SCIENCES ECONOMIQUES
THEOLOGIE

Université de Neuchâtel
Institut de Statistique

Thèse présentée à la Faculté des Sciences Economiques
pour l'obtention du grade de Docteur en Statistique

par

Donatien Tafin Djoko

Nonlinear, Nonparametric methods in empirical finance

Acceptée sur proposition du jury :

Prof. Catalin Starica	Université de Neuchâtel	Directeur de thèse
Prof. Yves Tillé	Université de Neuchâtel	Président du jury
Prof. Tommaso Proietti	Università degli Studi di Roma "Tor Vergata"	Rapporteur
Prof. Eric Hjalmarsson	University of Goteborg	Rapporteur
Prof. Jens-Peter Kreiß	Technische Universität Braunschweig	Rapporteur

Thèse soutenue le 25 Novembre 2013.

UNIVERSITÉ DE
NEUCHÂTEL

IMPRIMATUR POUR LA THÈSE

Nonlinear, Nonparametric Methods in Empirical Finance

Donatien Tafin Djoko

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Catalin Starica (directeur de thèse, Université de Neuchâtel)
Prof. Yves Tillé (président du jury, Université de Neuchâtel)
Prof. Kilian Stoffel (Université de Neuchâtel, remplaçant du prof. Jens-Peter Kreiss, Technische Universität Braunschweig)
Prof. Erik Hjalmarsson (University of Goteborg)
Prof. Tommaso Proietti (University of Roma Tor Vergata)

Autorise l'impression de la présente thèse.

Neuchâtel, le 17 décembre 2013

Le doyen

J-M Greter

Jean-Marie Grether

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

To the memories of my father Chretien
and my brother Jean-Pierre;

And to my newborn son, Lukas Eskil

Acknowledgements

First I would like to express my sincere gratitude to my supervisor, Professor Catalin Starica, for giving me the possibility to work towards a PhD degree in statistics under his supervision and for sharing his scientific knowledge with me. His experience in the area has help me during my research. I am very appreciative for the opportunity he gave me to follow my research path in total autonomy. Thank you for believing in me.

I thank all the member of the jury, for the commitment to be part of it.

I am specially grateful to Professor Yves Tillé, for his friendship and advice during my work at the Institute of statistics. Thank you for your encouragement and guideline during the sampling project.

Thanks to all the current and previous collaborators of the institute of statistics for their constant fruitful comments. A special gratitude to my colleagues, Alina Matei, Eric Graf, Caren Zürcher, Corine Diacon, and Jian Kang for their positive work spirit.

The authors of LaTeX deserves an enormous thank you for creating such an excellent editor program. They merit all the honors of the scientific community, especially as they do the work motivated by passion and are continuously ready to impart advice on how to get the most out of the program.

I would like to thank the person who made me believe in education, my Mother, Micheline. She gave me love and taught me to always try to do my best and that the only way to achieve something is to take some risk.

Finally, I would like to thank the person who enlightens my life, my beloved wife Ulrika. Thank for your daily words of encouragement. Nothing could have been possible without your positive energy.

Neuchâtel, January 20, 2014.

Abstract and keywords

Nonlinear, Nonparametric methods in empirical finance

Abstract During the recent global financial crisis, the world discovered how complex financial securities have become. Modelling and analyzing financial time series are increasingly challenging due to the complexity of modern financial markets. The aim of this thesis is to partially address these challenges using statistical methods. This manuscript focuses on both alternative (hedge fund) and traditional asset classes (equity). The methods we survey take into account the complexity of the data. They relax the parametric assumptions usually made on the data generating process. The first chapter is devoted to the description of the statistical properties of hedge fund time series returns. Chapter 2 is dedicated to a better understanding at a macro level, of drivers underlying hedge fund returns. A strategy to replicate aggregate hedge fund returns, based on specific market risk factors is discussed. Chapter 3 proposes a nonparametric portfolio allocation framework, aimed at implementing a performance-adaptive allocation of capital between individual hedge funds. Chapters 4 and 5 address the market indexing issue. A balanced sampling methodology is proposed to select a sub-set of assets that enables to track performance of the total of the population of interest. While the former chapter focuses on the stock market as a whole, the later is an attempt to track the dynamic of the hedge fund universe relative Net-Asset-Values.

Keywords Hedge fund, Traditional asset classes, Non-normality, Serial correlation, Market risk factor, Replication, Mean square error, Portfolio allocation, Kernel estimation, GO-GARCH, Distribution free, Market tracking, Survey sampling, Balanced sampling, Capitalization, Fund tracking.

Méthodes non-linéaires et non-paramétriques, en finance empirique

Résumé Au cours de la récente crise financière globale, la communauté scientifique a découvert la façon dont les titres financiers sont devenus complexes. La conséquence évidente d'un tel environnement est son impact sur la dynamique des prix des titres. La modélisation et analyse de séries temporelles financières sont de plus en plus remises en question par la complexité des marchés financiers modernes. L'objectif de cette thèse était de répondre partiellement à ces défis en utilisant des méthodes statistiques. Ce manuscrit se concentre à la fois sur les classes d'actifs alternatifs (fonds de couverture) et traditionnelles (actions). Les méthodes que nous enquêtons prennent en compte la complexité de ces données et découlent de ce qu'ils détendent les hypothèses paramétriques habituelles forcées sur le processus de génération des données. Le premier chapitre (1) est consacré à la description des propriétés statistiques des séries temporelles des rendements des fonds de couverture, par rapport aux instruments de placement plus traditionnels. Le chapitre 2 est dédié à une meilleure compréhension à un niveau macro, des facteurs sous-jacents les rendements des fonds de couverture. Une stratégie est alors mise en œuvre pour prédire les rendements agrégés de ces fonds, basée sur des facteurs de risque du marché. Le chapitre 3 proposes un cadre non-paramétrique d'allocation d'actifs, pour construire une affectation adaptative des capitaux entre fonds de couverture et obtenir un portefeuille optimal universel. Les chapitres 4 and 5 abordent eux la question de l'indexation du marché. L'auteur propose la méthode d'échantillonnage équilibré pour sélectionner un sous-ensemble d'actif permettant de reproduire la dynamique de l'ensemble de la population d'intérêt. Alors que le chapitre 4, met l'accent sur la capitalisation boursière totale, le second s'attaque au total des ratios des valeurs liquidatives des fonds alternatifs.

Mots-clés Fonds alternatifs, Classes d'actifs traditionnelles, Non-normalité, Corrélation sérielle, Facteur de risque de marché, Réplication, Erreur quadratique moyenne, Allocation de portefeuille, Estimation à noyau, GO-GARCH, Non-paramétrique, Réplication des indices boursières, Sondage, Échantillonnage équilibré, Capitalisation, Réplication des indices de fonds alternatifs.

Contents

List of Figures

List of Tables

List of Algorithms

Introduction

The purpose of this dissertation is twofold. The first objective is to better understand important issues related to the risk structure of *hedge funds*, a relevant class of alternative investment instruments. The second part focuses on the *indexing* problem of both traditional equity asset class and alternative investment vehicles. The specificity of the present work consists in the use of advanced statistic methods for a non-linear, nonparametric, dynamic investigation of questions that the existing quantitative literature on statistical financial modeling treats mostly in a parametric, linear and static way.

The growing interest for the alternative investments sector goes back to the end of the 1970's, when a wave of deregulations considerably reshaped the financial industry. The liberalization of capital movements combined with broadly reaching innovations in information technology transformed the way in which the returns are reaped and the risks are shared by the financial actors. Those transformations led to the development of a globalized form of finance.

Under this new architecture, the last two decades have witnessed the emergence of a new class of risk-takers: the alternative investments sector. They constitute the fastest growing community in the global financial market and, probably, the least understood. Hedge funds are key players of this community. Since 2000 and the institutionalization of the industry, the hedge fund universe has grown exponentially. Driven by the demand for higher returns and diversification purposes, specially by pension funds, the industry asset under management is today more than four times bigger than a decade earlier[1].

Due to the unregulated status of their business model, hedge fund managers enjoy enormous flexibility and discretion in pursuing investment returns. The innovative thinking of the sector has provided important benefits for the financial community in terms of increase in trading volume and efficiency, risk sharing and diversification, liquidity, technical innovation on market microstructure as well as improvement in governance. There is no doubt that hedge funds managers are today key actors in the global financial industry. Given the poor returns of the conserva-

[1]The third quarter 2012 market micro-structure report on hedge fund industry released by the Chicago based Hedge Fund Research, pointed that the industry assets reached a record level of US $ 2.2 Trillion across over 7,867 single-managers funds.

tive equity-based strategies during the last part of the previous decade, various sources forecast they will play an increasing role in the future.

According to Rahl (2003): 'A hedge fund is a pooled investment vehicle that is privately organized, administrated by professional investment managers and not widely available to the public. Many of these funds share a number of characteristics: they hold long and short positions, employ leverage to enhance returns, pay a performance or incentive fee to their managers, have high minimum investment requirements, target absolute (rather than relative) returns, and may be organized offshore. Hedge funds strategies are relatively unconstrained in terms of short selling, leverage and investment universe'.

In contrast, traditional investment vehicles like mutual funds and/or pension funds, are driven by different culture and investment guidelines. Basically, a typical institutional investor or mutual fund operates in a highly state regulated environment. Transparency, liquidity, long term perspective, and relatively light fee structure are key characteristics governing the mutual fund industry. While hedge fund managers value their investment at low frequency during the year, mutual funds are forced to value and price their securities at high frequency (daily) based on market quotations or fair value.

The practice of indexing has its origin in the same effort of deregulation that characterized the 70's. The first example of this type of an investment strategy consisted in constructing a fund mimicking the dynamics of an equal-weighted index of all constituents of the New York Stock Exchange. The primary objective of an index fund is hence to track a specific financial market. Since then, index-based products have become increasingly popular. The current state of the art in index-tracking funds is the class of Exchange-traded funds (ETFs). According to The Economist, January 26 (2013), the ETFs industry has enjoyed an average annual growth rate of 34% over the last ten years. By May 2011, the magazine estimated that the global exchange-traded funds total assets under management reached almost US $1.5 trillion. Beyond the cost-benefits trade-off, indexing serves to construct efficient benchmarks for portfolio evaluation and risk management.

This monograph is structured in five chapters. Chapter 1 describes the statistical properties of hedge fund returns, compared to more traditional asset classes like equity and government bond. The particularities of this class of alternative strategy that are not shared by other better studied financial instruments are emphasized. Motivations regarding the specificity of the computational implementations are presented. This first chapter is conceived as an overview of the key characteristics of financial returns and serves as an introduction to the research questions explored in the following chapters.

Chapters 2 to 5 are articles encompassing different themes. Chapter 2 is dedicated to hedge fund replication. It provides an answer to the issue of whether it is possible to match hedge fund returns, without directly investing in them. It extends the purely linear factor-based replication strategy by incorporating the dynamic nature of hedge fund returns. Assuming that hedge funds are significantly exposed to common market risk factors that evolves through time, we implement the sequential investment theory as discussed by Györfi, Lugosi, and Udina (2006) to predict hedge index returns. Empirical results suggest that our method performs better than two established replication techniques, and that the explanatory power of some market risk factors are weak.

In Chapter 3, hedge fund portfolio allocation is investigated. This essay examines a dynamic, performance-adaptive asset allocation across individual hedge funds. The goal is to construct a time evolving allocation algorithm that optimally diversifies over multiple hedge fund styles. We relinquish the global stationary hypothesis which is the foundation of commonly used portfolio modeling and adopts a locally stationary paradigm. A nonparametric method is implemented from the perspective of the theory of sequential investment. This theory draws its origin from Kelly (1956)'s seminal paper on information rate and has been recently an active field of research. Specifically, the study shows that distribution-free portfolio of hedge funds yields sizeable economic value.

The second part of the thesis addresses the indexing problem. This part is divided in two chapters and provides the rationale behind a sampling market tracking strategy. In Chapter 4, we propose a survey sampling technique to track the performance of a specific financial market. We believe that the issue of replicating an index using a subset of its constituents is a natural sampling topic. The motivation of this chapter is to link finance and sampling through the method of balanced sampling. The notion of random selection and approximately balanced tracking portfolio are emphasized. A large number of portfolios are constructed, and results exhibit that balanced sampling designs allow the researcher to track efficiently the broad market index with a relatively small number of stocks.

While Chapter 4 is devoted to equity market indexing, Chapter 5 draws a new perspective on hedge fund tracking performance. In general, investors use an index as a gauge for a given market. However, unlike traditional equity asset class, hedge funds are by no means homogeneous. They represent a large range of investment profiles with various mandates. Therefore, indexing is an attempt to capture the broad universe of funds. Identifying a subset of managers that allow to track the dynamic changes of the total hedge fund universe is not trivial. This task is the objective of the fund of hedge funds strategies. In this chapter, the hedge fund indexing problem is solved by means of balanced sampling. The methodology constructs tracking portfolios that replicate the total relative Net-Asset-Value of a large basket of funds. Funds are selected randomly with unequal inclusion probabilities. Each tracking portfolio is decomposed to illustrate the automatic diversification characteristic of balanced sampling strategies.

The thesis ends with a general conclusion and some considerations on several natural extensions.

1 Characteristics of financial returns: Hedge funds Vs Traditional asset classes

Abstract

This chapter introduces key statistical characteristics of hedge funds. The difference to mutual funds and more traditional asset classes such as equities, and government bonds is emphasized. Section 1.1 describes the particularity of hedge fund business model. The static and dynamic properties of hedge fund returns are outlined in Sections 1.2 and 1.3, respectively. The analysis is both univariate and multivariate.Short definitions of the industry main investment categories are given in Appendix 1.5.

Keywords: Hedge fund, Traditional asset classes, Non-normality, Serial correlation

1.1 Hedge funds are different

Hedge funds are particular in many ways starting with how hedge fund data is gathered. The industry decides the timing and contents of the records delivered to commercial databases. Data frequency and inconsistencies are of particular concern in the statistical modelling of hedge fund data. Researchers have been aware for a while now of potential biases in hedge fund databases resulting from the nature of the mentioned data collection process. Ackermann, McEnally, and Ravenscraft (1999) pointed out that hedge fund databases have survivorship, liquidation, backfill, self reporting, and selection biases. For a more detailed overview of the database bias problems, the reader may refer to Liang (2000), Fung and Hsieh (2000), and Bollen and Pool (2009) among others. Therefore, when analyzing the industry performance characteristics, one should keep in mind the irreducible incomplete picture of the data.

Other particularity consists in the provision restrictions imposed by managers to investors. In addition to various managerial incentives, hedge funds have the discretion to place several restrictions on the ability of investors to redeem their allocation. Hedge fund share restrictions include, but are not limited to, lockup periods, notice periods, and redemption frequency periods. The industry observed that share restrictions are important to achieve their investment mandate, since they invest mainly in illiquid assets and generate a significant portion of their double-digit return bearing illiquidity risk. According to Agarwal, Daniel, and Naik (2009), funds with a higher degree of managerial discretion yield superior performance. The authors identify an illiquidity premium.

Hedge fund returns have statistical characteristics that are not shared by those of other better studied classes of financial investment instruments. For example, Asness, Krail, and Liew (2001) documented the presence of serial correlation in the hedge fund indexes, as well as the presence of correlation to past returns of equity benchmarks such as the Standards & Poors 500 (S&P 500) stock index. They explain this correlation as the result of either infrequent trading of illiquid securities in their portfolios or of manipulation of results by managers to smooth their returns. Although various other attempts were made to formally model these features (see for example Getmansky, Lo, and Makarov (2004) for a detailed discussion on the key concepts), it is still unclear which of the two possible causes of serial correlation in hedge fund returns prevails.

Risk, an important dimension of any investment, acquires in the case of hedge funds a particular centrality. The basic assumption in modern portfolio theory is the tradeoff between risk and return. High return profile in the hedge fund industry is undoubtedly associated with high risk. Understanding the risks of a hedge fund investment is important for all the actors involved: investors, counterparties, regulators. The investors need to know if the returns generated by the hedge fund are commensurate with the fees charged on a *risk-adjusted basis*. They also need to be aware of the risk they are adding to their portfolios by investing in a hedge fund. Hedge fund's counterparties need an accurate estimation of the risk to assess the capital at risk for engaging in financing the hedge fund business. From the regulators' perspective, the answer to these questions are fundamental in monitoring the systemic risk.

1.2 A static picture

The current section discusses the basic statistical properties of hedge fund returns. For the empirical analysis of this chapter, we focus on an aggregate level and use the Dow Jones Credit Suisse Hedge Fund Index[1] (DJ/CS). The DJ/CS indexes are asset-weighted and the universe comprised funds with A minimum of US $50 million assets under management (AUM), a minimum one year track record, and current audited financial statements. 14 indexes are computed from this universe of funds: one aggregate and 13 investment style sub-indexes (see the Appendix for more details on the descriptions of each DJ/CS index categories). The data are monthly returns and cover the period from January 1994 (except for the Multi-Strategy index which starts from April) through November 2012, for a total of 227 observations. Also included for purposes of comparison, are two equity indexes (S&P 500 and the small-cap Russell 2000 (RU2000)) and a bond market indicator[2] (yield on U.S. treasury securities at 20-year constant maturity at a monthly basis (20YUSCM)).

Table 1.1 reports the performance summary statistics of the 14 hedge fund styles and the 3 more traditional asset classes. The annual mean and standard deviation columns may explain the growing success of this class of alternative investment strategy over the last two decades. For most categories, while average volatility are lower, average returns are superior to the two equity indexes listed in Table 1.1, implying bigger annualized risk-adjusted performances. However, hedge funds are by no means an homogeneous industry. The worst performing category is Dedicated Short Bias, with an annual mean return of −2.93% and volatility of 16.89%, while the best performing is Global Macro index (11.57%, 9.60%, mean and standard deviation respectively).

Concerning the higher sample moments, we document a large heterogeneity in the magnitude of skewness and kurtosis across indexes. Indexes with large skewness values tend to exhibit large kurtosis. All indexes except Dedicated Short Bias, Global Macro, and Managed Futures have a significantly negative skewness, meaning that down market conditions occur more often than booms. Additionally, the higher level of kurtosis values are inconsistent with normality hypothesis. For the three more traditional indexes, the magnitude of those values belong to the lower bound of what reported in Table 1.1. We then test the time dependency in both returns and squared returns. A natural way of testing such time dependency is the Ljung-Box statistic (ρ). The last two-column entries of Table 1.1 report the 95% P-value of the 10 lags Ljung-Box statistics in testing time dependency in returns and serial correlation in volatility. Except Dedicated Short Bias and Equity Market Neutral indexes, there is strong evidence of time dependency in monthly hedge fund returns. This characteristic is not shared by the two traditional equity indexes listed in Table 1.1. As

[1]The sources of hedge fund data are multiples. For the empirical analysis in this dissertation, I use in addition to DJ/CS, two main sources: the Chicago-based Hedge Fund Research and Barclay commodities trading advisors databases.

[2]Equities indexes are from Datastream and the yield on 20-year U.S. treasury securities are downloaded from the FED data library.

aforementioned in Section 1.1, various studies associated this feature to the illiquid nature of theses investment vehicles. However, regarding serial correlation in volatility, we notice that the evidence is mixed. While for the global Hedge Fund index we cannot reject the null hypothesis that the first 10 auto-correlations of the squared returns are jointly 0, this is not the case for 6 other categories.

According to various hedge fund managers, their performances are relatively uncorrelated with broad market benchmarks. This property is closely related to the term 'hedge', since investing in those funds my provide diversification. Table 1.2 reports the correlation matrix of the DJ/CS hedge fund index returns. The table also reports the correlations of hedge fund indexes with the returns of three traditional investment vehicles. We notice a rather large dispersion in the magnitude of correlations within hedge funds. If we focus to the global Hedge Fund index, we observe a low and positive association to the long term government bond and relatively high correlation to the equity benchmarks. The diversification benefit to the equity markets is justified only for Dedicated Short Bias index (correlation of -76.4% , and -82.7% with S&P500 and RU2000 respectively). It is also very important to mention the non negligible cross correlation across hedge fund styles. However, as we will see in a sequel, correlation change over time, reflecting the dynamic nature of hedge fund investing.

Table 1.1 – *This table reports summary statistics for hedge fund, equity index returns and for the 20-year US government yield: the annualized average return and standard deviation, the standardized skewness and kurtosis, the t-statistics of mean, skewness and kurtosis, and the P-values of jarque-Bera normality test, and Ljung-Box test of order* 10 *for no serial correlation of returns* ($\rho(r)$) *and squared returns* $\rho(r^2)$.

Index	Annual Mean(%)	Mean t-statistic	Annual SD(%)	Skewness	Skewness t-statistic	Kurtosis	Kurtosis t-statistic	Jarque-Bera P-value	$\rho(r)$ P-value	$\rho(r^2)$ P-value
Hedge Fund	8.57	-692.93	7.48	-0.18	-1.10	5.50	7.68	0.00	0.07	0.00
Convertible Arbitrage	7.45	-757.71	6.85	-2.69	-16.52	19.15	49.67	0.00	0.00	0.00
Dedicated Short Bias	-2.93	-309.85	16.89	0.68	4.16	4.39	4.27	0.00	0.60	0.19
Emerging Markets	8.24	-352.40	14.71	-0.76	-4.68	8.13	15.79	0.00	0.00	0.43
Equity Market Neutral	5.39	-509.08	10.21	-11.89	-73.12	165.19	498.81	0.00	0.52	
Event Driven	9.04	-826.15	6.27	-2.23	-13.71	13.60	32.59	0.00	0.00	0.99
Distressed	10.02	-788.93	6.56	-2.15	-13.23	13.77	33.13	0.00	0.00	0.99
E.D. Multi-Strategy	8.60	-763.84	6.78	-1.75	-10.73	10.49	23.04	0.00	0.66	0.00
Risk Arbitrage	6.37	-1254.05	4.14	-0.96	-5.88	7.46	13.72	0.00	0.00	0.00
Fixed Income Arbitrage	5.54	-918.36	5.66	-4.47	-27.49	34.16	95.83	0.00	0.00	0.00
Global Macro	11.57	-538.26	9.60	0.03	0.18	6.89	11.97	0.00	0.00	0.00
Long/Short Equity	9.21	-525.75	9.85	-0.01	-0.06	6.20	9.85	0.00	0.01	0.00
Managed Futures	6.02	-444.80	11.67	0.04	0.22	2.94	-0.18	0.96	0.04	0.59
Multi-Strategy	7.85	-965.45	5.33	-1.70	-10.40	8.89	18.00	0.00	0.00	0.00
20YUSCM	2.24	-34234.27	0.15	-0.15	-0.92	2.80	-0.60	0.54	0.00	0.00
S&P500	7.10	-335.20	15.48	-0.66	-4.05	3.94	2.90	0.00	0.59	0.00
RU2000	8.16	-259.19	20.00	-0.49	-3.03	3.99	3.05	0.00	0.47	0.01

Table 1.2 – *Correlation matrix of DJ/CS hedge fund, equity index returns, and 20-year US government yield. All values are percentages and computed on monthly returns. The data covers the period from January 1994 to November 2011, except for the Multi-Strategy index starting at April 1994.*

	Hedge Fund	Convertible Arbitrage	Dedicated Short Bias	Emerging Markets	Equity Market Neutral	Event Driven	Distressed	Event Driven Multi-Strategy
Hedge Fund	100.0							
Convertible Arbitrage	49.8	100.0						
Dedicated Short Bias	-53.2	-31.4	100.0					
Emerging Markets	71.8	36.5	-52.6	100.0				
Equity Market Neutral	46.7	45.2	-33.3	40.4	100.0			
Event Driven	79.5	62.9	-60.6	67.1	52.0	100.0		
Distressed	69.4	55.4	-56.6	58.0	45.7	91.1	100.0	
E.D. Multi-Strategy	78.7	62.7	-54.2	67.3	50.2	93.4	73.1	100.0
Risk.Arbitrage	50.0	40.7	-44.3	46.0	45.6	56.7	45.8	54.9
Fixed Income Arbitrage	49.3	58.0	-28.7	37.5	32.0	52.9	47.7	52.8
Global Macro	71.0	22.6	-10.9	42.3	27.0	38.4	31.2	40.5
Long/Short Equity	83.5	42.2	-71.9	65.6	50.9	76.5	66.5	72.9
Managed Futures	32.8	-0.2	-1.3	9.6	11.7	15.4	13.2	13.5
Multi-Strategy	51.5	57.7	-37.2	38.7	44.7	58.2	46.4	60.9
20YUSCM	9.3	12.3	9.8	2.2	23.3	10.0	14.4	6.4
S&P500	61.2	31.4	-76.4	54.7	43.0	60.9	56.7	56.6
RU2000	60.5	31.0	-82.7	56.9	34.7	64.5	58.2	59.7

Table 1.2: (continued)

	Risk Arbitrage	Fixed Income Arbitrage	Global Macro	Long/Short Equity	Managed Futures	Multi-Strategy	20YUSCM	S&P500	RU2000
Hedge Fund									
Convertible Arbitrage									
Dedicated Short Bias									
Emerging Markets									
Equity Market Neutral									
Event Driven									
Distressed									
E.D. Multi-Strategy									
Risk.Arbitrage	100.0								
Fixed Income Arbitrage	29.7	100.0							
Global Macro	24.3	28.2	100.0						
Long/Short Equity	58.6	39.8	37.9	100.0					
Managed Futures	4.1	5.7	47.0	19.3	100.0				
Multi-Strategy	35.1	39.2	18.8	49.2	11.2	100.0			
20YUSCM	19.5	4.3	11.4	5.7	-0.5	6.6	100.0		
S&P500	46.3	32.4	21.1	72.2	2.6	40.2	5.2	100.0	
RU2000	51.1	25.6	17.3	76.5	2.1	39.4	-1.8	78.3	100.0

1.3 Time-varying analysis

The existence of different market regimes have important consequences in the financial arena, with direct implications in investor's risk-return structure. For unregulated hedge fund managers, changes in regime (transitory or persistent) mean most likely adjusting the investment strategies. A realistic example of regime switching allocation is going long in equity during business cycle expansions and short equity during recessions. Switching strategies imply regime dependent exposures to broad-based market indicators like the S&P500.

To investigate the time variability in hedge fund exposures to the market, we first consider the rolling 1-year correlation between the DJ/CS hedge fund indexes and the S&P 500, plotted in Figure 1.1. The evolution of correlations is highly dynamic. While most of the changes in rolling correlation estimates are transitory, we emphasize the existence of persistence changes that have a long term effect on manager behaviors and return distributions. For example, observing the sub-figure in Figure 1.1, we see a persistent change in the correlation dynamic between the Multi-Strategy index and the S%P 500. After years of negative and low dependence structure, the nature of the relation has changed from the start of 1999 and has been high up ever since. Except for the Dedicated Short Bias index which exhibits long term negative correlation to the equity market, it is hard to believe that hedge funds provide diversification benefits. Varied evidences suggest they trade in the same assets as the rest of the financial community, with certainly a different philosophy (dynamic trading strategies).

1.4 Conclusion

Assessing and understanding the statistical properties of financial time series is an important step for a quantitative analysis. The rapid growth of hedge fund investing requires a solid investigation of the nature of their performance. This chapter shows how the characteristics of hedge fund returns are different from traditional asset classes. Hedge fund return characteristics are time dependent, and their distribution is asymmetric and leptokurtic.

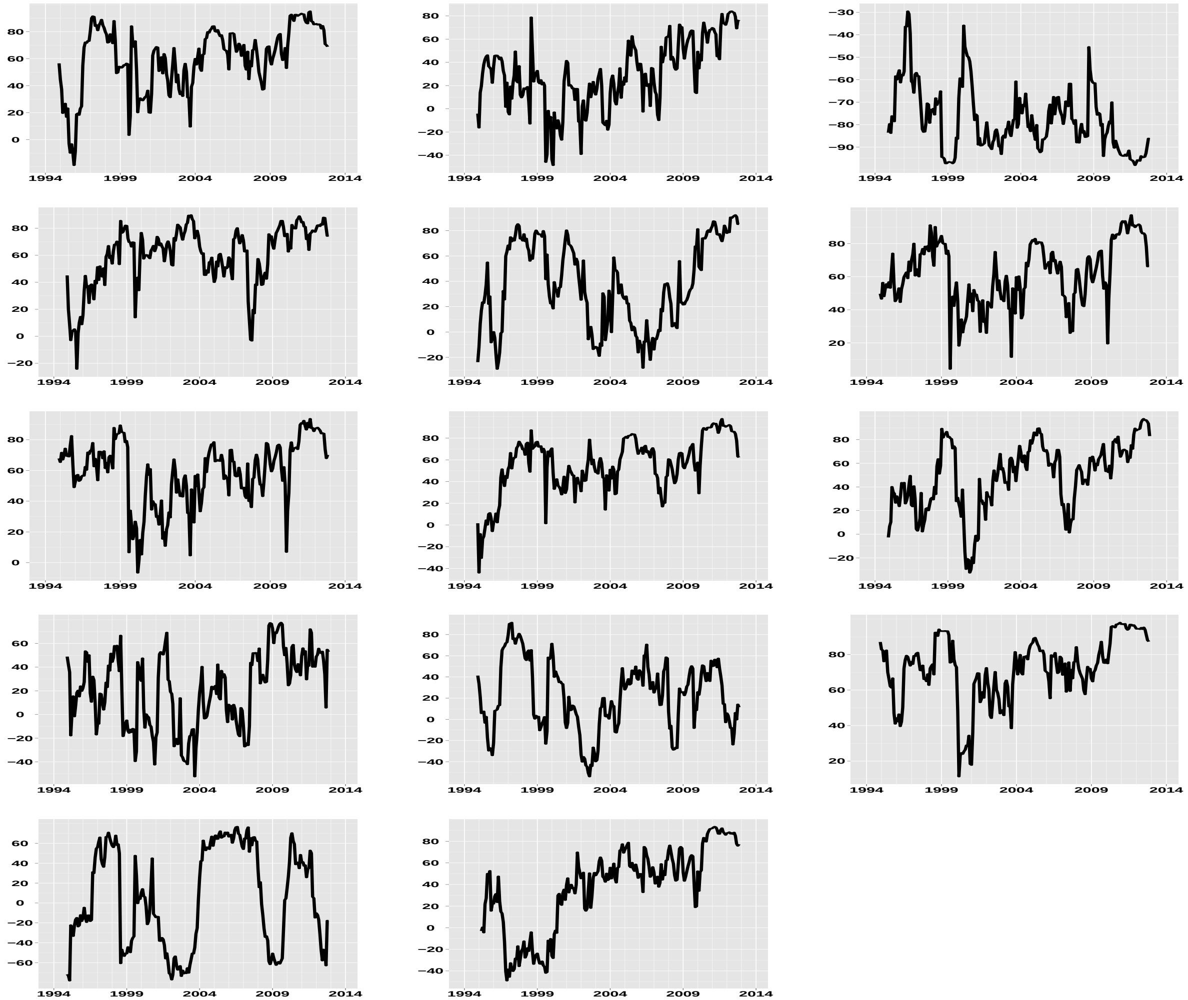

Figure 1.1 – 1-*year rolling correlations (%) between DJ/CS hedge fund index returns and the S&P 500 returns. The order from top-left to bottom-right corresponds to hedge fund category listed in Table 1.1.*

1.5 Appendix

This appendix contains the Credit Suisse hedge fund index category definitions used in this chapter. The definitions are taken from the data provider hedgeindex (2012):

" **Hedge Fund Index**: *It is an asset-weighted hedge fund index and includes only funds, as opposed to separate accounts. The index uses the Credit Suisse Hedge Fund Database (formerly known as the "Credit Suisse/Tremont Hedge Fund Database"), which tracks approximately 8000 funds and consists only of funds with a minimum of US$50 million under management, a 12-month track record, and audited financial statements. The index is calculated and rebalanced on a monthly basis, and reflects performance net of all hedge fund component performance fees and expenses.*

Convertible Arbitrage: *Convertible arbitrage funds typically aim to profit from the purchase of convertible securities and the subsequent shorting of the corresponding stock when there is a pricing error made in the conversion factor of the security. Managers of convertible arbitrage funds typically build long positions of convertible and other equity hybrid securities and then hedge the equity component of the long securities positions by shorting the underlying stock or options. The number of shares sold short usually reflects a delta neutral or market neutral ratio. As a result, under normal market conditions, the arbitrageur generally expects the combined position to be insensitive to fluctuations in the price of the underlying stock.*

Dedicated Short Bias: *Dedicated short bias funds typically take more short positions than long positions and earn returns by maintaining net short exposure in long and short equities. Detailed individual company research typically forms the core alpha generation driver of dedicated short bias managers, and a focus on companies with weak cash flow generation is common. To affect the short sale, the manager typically borrows the stock from a counterparty and sells it in the market. Short positions are sometimes implemented by selling forward. Risk management often consists of offsetting long positions and stop-loss strategies.*

Emerging Markets Hedge Fund: *Emerging markets funds typically invest in currencies, debt instruments, equities and other instruments of countries with "emerging" or developing markets (typically measured by GDP per capita). Such countries are considered to be in a transitional phase between developing and developed status. Examples of emerging markets include China, India, Latin America, much of Southeast Asia, parts of Eastern Europe, and parts of Africa. The index has a number of subsectors, including arbitrage, credit and event driven, fixed income bias, and equity bias.*

Equity Market Neutral: *Equity market neutral funds typically take both long and short positions in stocks while seeking to reduce exposure to the systematic risk of the market (i.e., a beta of zero is desired). Equity market neutral funds typically seek to exploit investment opportunities unique to a specific group of stocks, while maintaining a neutral exposure to broad groups of stocks defined for example by sector, industry, market capitalization, country, or region. The index has a number of subsectors including statistical arbitrage, quantitative long/short, fundamental long/short and index arbitrage. Managers often apply leverage to enhance returns.*

Event Driven: *Event driven funds typically invest in various asset classes and seek to profit from potential mispricing of securities related to a specific corporate or market event. Such events can include: mergers, bankruptcies, financial or operational stress, restructuring, asset sales, recapitalization, spin-offs, litigation, regulatory and legislative changes as well as other types of corporate events. Event driven funds can invest in equities, fixed income instruments (investment grade, high yield, bank debt, convertible debt and distressed), options and various other derivatives. Many event driven fund managers use a combination of strategies and adjust exposures based on the opportunity sets in each subsector.*

Event Driven Distressed: *Measures the aggregate performance of event driven funds that focus on distressed situations. These funds typically invest across the capital structure of companies subject to financial or operational distress or bankruptcy proceedings. Such securities often trade at discounts to intrinsic value due to difficulties in assessing their proper value, lack of research coverage, or an inability of traditional investors to continue holding them. This strategy is generally long-biased in nature, but managers may take outright long, hedged or outright short positions. Distressed managers typically attempt to profit on the issuer's ability to improve its operation or the success of the bankruptcy process that ultimately leads to an exit strategy.*

Event Driven Multi-Strategy: *Multi-strategy event driven managers typically invest in a combination of event driven equities and credit. Within the equity space, sub-strategies may include risk arbitrage, holding company arbitrage, equity special situations, and value equities with a hard or soft catalyst. Within the credit-oriented portion, sub-strategies may include long/short high yield credit (sub-investment grade corporate bonds), leveraged loans (bank debt, mezzanine, or self-originated loans), capital structure arbitrage (debt vs. debt or debt vs. equity), and distressed debt (workout situations or bankruptcies) including post-reorganization equity. Multi-strategy event driven managers typically have the flexibility to pursue event investing across different asset classes and take advantage of shifts in economic cycles.*

Event Driven Risk Arbitrage: *Risk arbitrage event driven hedge funds typically attempt to capture the spreads in merger or acquisition transactions involving public companies after the terms of the transaction have been announced. The spread is the difference between the transaction bid and the trading price. Typically, the target stock trades at a discount to the bid in order to account for the risk of the transaction failing to close. In a cash deal, the manager will typically purchase the stock of the target and tender it for the offer price at closing. In a fixed exchange ratio stock merger, one would go long the target stock and short the acquired's stock according to the merger ratio, in order to isolate the spread and hedge out market risk. The principal risk is usually deal risk, should the deal fail to close.*

Fixed Income Arbitrage: *Fixed income arbitrage funds typically attempt to generate profits by exploiting inefficiencies and price anomalies between related fixed income securities. Funds often seek to limit volatility by hedging out exposure to the market and interest rate risk. Strategies may include leveraging long and short positions in similar fixed income securities that are related either mathematically or economically. The sector includes credit yield curve relative value trading involving interest rate swaps, government securities and futures; volatility trading involving options; and mortgage-backed securities arbitrage (the mortgage-backed market is primarily U.S.-based and over-the-counter).*

Global Macro: *Global macro funds typically focus on identifying extreme price valuations and leverage is often applied on the anticipated price movements in equity, currency, interest rate and commodity markets. Managers typically employ a top-down global approach to concentrate on forecasting how political trends and global macroeconomic events affect the valuation of financial instruments. Profits can be made by correctly anticipating price movements in global markets and having the flexibility to use a broad investment mandate, with the ability to hold positions in practically any market with any instrument. These approaches may be systematic trend following models, or discretionary.*

Long/Short Equity: *Long/short equity funds typically invest in both long and short sides of equity markets, generally focusing on diversifying or hedging across particular sectors, regions or market capitalizations. Managers typically have the flexibility to shift from value to growth; small to medium to large capitalization stocks; and net long to net short. Managers can also trade equity futures and options as well as equity related securities and debt or build portfolios that are more concentrated than traditional long-only equity funds.*

Managed Futures: *Managed futures funds (often referred to as CTAs or Commodity Trading Advisors) typically focus on investing in listed bond, equity, commodity futures and currency markets, globally. Managers tend to employ systematic trading programs that largely rely upon historical price data and market trends. A significant amount of leverage may be employed since the strategy involves the use of futures contracts. CTAs tend not to have a particular bias towards being net long or net short any particular market.*

Multi-Strategy: *Multi-strategy funds typically are characterized by their* *ability to allocate capital based on perceived opportunities among several hedge fund strategies. Through the diversification of capital, managers seek to deliver consistently positive returns regardless of the directional movement in equity, interest rate or currency markets. The added diversification benefits may reduce the risk profile and help to smooth returns, reduce volatility and decrease asset-class and single-strategy risks. Strategies adopted in a multi-strategy fund may include, but are not limited to, convertible bond arbitrage, equity long/short, statistical arbitrage and merger arbitrage. "*

2 Hedge Fund Replication: A Dynamic Performance-Adaptive Local Linear Regression Approach

Abstract

This article implements the sequential investment methodology of Györfi, Lugosi, and Udina (2006) for the purpose of hedge fund replication. Using a top-down replication approach and style-specific sets of observable market risk factors, we construct a month-to-month, *performance-adaptive replication* strategy of hedge fund indexes. The proposed methodology extends the purely linear, multi-factor technique by incorporating a flexible historical data selection process to determine the loadings of the factors used in the replication of the index under scrutiny.

The empirical results suggest that the sequential investment approach performs better - in the sense of mean square replication error, cumulative square replication error, percentage of the direction change, and correlation between the estimated series and the realized series - than the linear modeling. Additionally, the return series constructed through replication using this methodology share several statistical properties with the hedge fund index returns. The experiment highlights the power of a performance-adaptive, factor-based replication approach in the context of time series hedge fund returns forecasting.[1]

Keywords: Hedge fund, Market risk factor, Replication, Mean square error

2.1 Introduction

The beginning of the 21st century marked the end of the Internet bubble and the emergence of hedge fund managers as one of the key players in

[1]This chapter is a reprint of: Tafin Djoko and Starica (2013) Hedge Fund Replication: A Dynamic Performance-Adaptive Local Linear Regression Approach. Submitted for publication.

the financial industry. During the last decade, the hedge fund sector has been the fastest growing sector of the financial services. Despite the severity of the 2007 global financial crisis, the hedge fund industry manages more than $2 trillion[2] of assets worldwide. Moreover, one can notice a strong tendency towards 'main-streaming' of this type of investments. A growing proportion of the assets under management is currently represented by institutional investors[3]. As a result of sector's innovative thinking, hedge fund investments have provided important advantages to the financial community in terms of increase trading volume and efficiency, risk sharing and diversification, liquidity, technical innovation on investment strategies as well as improved governance.

The immense success of the hedge fund class in the financial arena is explained by its unique combination of sought-after investment qualities : double-digit returns, low correlation to broad asset classes and relatively low volatility. Achieving such an unique structure of returns is possible due to a large extent to one of the peculiar features of the hedge fund business model: its relatively unregulated status. Hedge fund managers benefit of large flexibility and discretion in pursuing investment returns.

Other features of the hedge fund industry business model, like the opacity and the hefty fee structures, compared to more traditional investment vehicles, are potential impediments for both the further growth and the 'main-streaming' of the industry[4]. To address the lack of transparency and the prohibitive fee structure issues, a significant part of the quantitative research focusing on the class of alternative investment tries to answer the question whether it is possible to devise investment strategies based on liquid and trad-able underlying assets that yield hedge fund-like returns or at least the systematic exposure of hedge fund returns to known risk factors. Such strategies would render the hedge fund investment practices less opaque, highlighting the exposure of the hedge fund investments to various risk factors and would offer cheaper alternatives to the expensive hedge fund fee structure. Moreover, understanding the risk structure of hedge fund investments is essential to precisely quantify the risk of any diversified investor that holds hedge fund instruments in her portfolio. Hence, successful replication of hedge funds returns is extremely useful for both investment and risk analysis perspective.

[2]According to data from Hedge Fund Research, the industry's assets under management fell to $1.4 trillion at the end of 2008 from a record $1.9 trillion in 2007. The September 2011 release reports a record level for the total hedge fund industry capital of $2.04 trillion.

[3]The *Rothstein Kass* fifth annual report on hedge fund industry trends in 2011, pointed out that 70% of survey managers expected institutional money to dominate the sector. This number was just 20% in 2007 survey.

[4]Paul Schaeffer, managing director of strategy and innovation for SEI's investment manager services division says, "To maintain that growth trajectory, the hedge fund industry will need to branch out from its traditional high-net-worth, foundation and endowment clientele to serve the broader institutional market. But to compete for those assets, the industry must recognize that large institutions have a distinct set of demands."

Two main different empirical approaches have emerged in the stream of literature dedicated to hedge fund replication. In a *bottom-up* approach, also called *rule-based*, a number of papers have focused on identifying the set of investment instruments and techniques specific to the different hedge fund styles. The objective is hence to reproduce as closely as possible the trading of hedge fund managers and so reconstruct the fund return series. By mimicking fund's systematic strategies, Fung and Hsieh (2001) established a strong connection between the returns of managed futures (trend followers) and option straddles. Mitchell and Pulvino (2001) tracked the returns of merger arbitrage funds by passively invest in announced corporate mergers. Duarte, Longstaff, and Fan (2007) replicated the returns of fixed-income strategies based on observable prices of fixed-income securities and their derivatives. Furthermore, Agarwal, Fung, Loon, and Naik (2009) reconstructed convertible arbitrage (CA) returns by 'cloning' a family of CA strategies using data from the underlying securities.

The second approach advocates a *top-down* perspective by trying to determine the exposure of hedge fund returns to known systematic risk factors. This *factor-based approach* originates from the seminal paper of (Fung and Hsieh, 2004a) which proposed a risk-factor linear model based on seven factors to account for the risk of diversified portfolios of hedge funds. The risk factors were selected based on their empirically proven relevance to the hedge fund styles under discussion. In an extension beyond the linear case, De Los Rios and Garcia (2008) advocates the introduction of option-like factors with non-linear exposure to traditional asset classes, to extract the non-linear dynamic of hedge fund returns with respect to underlying market risk factors. In the same attempt of capturing the non-linear structure inherent in hedge fund returns stream, Roncalli and Teïletche (2008) discussed the implementation of Kalman filtering approach on various hedge fund indexes and found that, their estimates were better than the rolling-windows replication, while Amenc, Martellini, Meyfredi, and Ziemann (2009) test the performance of a Markov regime-switching model. Both statistical techniques try to explicitly account, through state-space modeling, for the dynamic dependency between hedge fund returns and market risk factors. As an alternative approach to hedge fund replication, Roncalli and Weisang (2011) uses a particle filter algorithm to capture the non-linear, non-normal features exhibit by hedge fund returns. While the particle filter tracking technique seems to be able to reproduce higher moments of the hedge fund returns, such as negative skewness and excess kurtosis, it fails to replicate a sizable proportion of the original hedge fund return series. In a related approach, Kat and Palaro (2006) generated return series that display statistical properties close to those of a hedge fund portfolio, by assessing the beta exposures to a basket of more than 70 liquid futures contracts. Despite its flexibility, the proposed methodology is highly complex and difficult to implement.

A close reading of these studies provides the researcher with a relatively precise guidance on the choice of the drivers (market risk factors) that are susceptible to explain a sizable fraction of the time varying evolution of hedge fund returns. Our approach intends to capitalize on this knowledge to draw a more nuanced picture of the relation between hedge fund returns and risk factors through the use of a relatively simple extension of the factor-based methodology. More concretely, the paper illustrates the implementation of sequential investment theory as discussed in Györfi, Lugosi, and Udina (2006) for the purpose of hedge fund index replication.

Commonly, the decision on the structure of a portfolio makes use of the past price evolution of the concerned investment instruments over an entire, contiguous time interval prior to the moment of the investment. In a top-down replication context, one builds a replicating portfolio from a given, well-defined set of observable market risk factors based on the historical values of the risk factors and hedge fund returns *covering a moving window of two, five or ten years of past data*. In contrast, a sequential investment strategy starts with the contiguous time interval prior to the moment of the construction of the replicating portfolio and identifies first the instances within the chosen interval where the state of the market, i.e. the risk factor structure, was 'similar' to the current conditions. Only the risk factor values and the hedge fund price information from these instances are then used to construct the current replicating portfolio. To summarize, the novelty of the sequential method consists in inspecting the price and factor history to collect a string of returns and risk factors corresponding to time-instances when the market conditions were 'close' to the present ones. These data only constitute the input for the estimation routine that yields the replicating portfolio.

The definition of the 'distance' between past and current market conditions is, of course, a crucial aspect in applying the sequential approach. Since we do not have *a priori* knowledge of the variables (risk factors, prices) that are instrumental in defining the state of the market, a refinement of the methodology puts in competition a number of elementary sequential strategies. Each one of them is characterized by a proxy for the market conditions, i.e. a given set of relevant variables (and a depth of the history of these variables) whose values are used to define 'similar' or 'close' market conditions. For each one of the market condition proxy, a replication portfolio is constructed, based on the time instances identified by the proxy as 'similar' to those of the replication moment. The final replication decision follows then an adaptive, time varying allocation of resources between different proxy replicating portfolios based on their recent past performance. For more information on sequential investments see Györfi, Lugosi, and Udina (2006) as well as Section 2.2.

The proposed methodology extends the purely linear, multi-factor technique by incorporating a more flexible data selection process to determine the loadings of the factors used in the replication of the index under scrutiny. Although the methodology is the same for all styles of hedge funds, the implementation is style-specific being determined by the choice of risk factors. The technique allows a style-specific detection of dynamic pattern between index returns and market risk factors.

Our empirical results suggest that the method performs better - in the sense of mean square replication error, cumulative square replication error, percentage of the direction change, and correlation between the predicted series and the realized series - than two established replication linear techniques: the classical and the rolling-windows linear approach. They highlight the power of performance-adaptive, factor-based prediction/explaining rules in the context of time series hedge fund returns.

Before we begin the in-detail discussion of our approach, we would like to make the following clarification about the terminology used in the paper. By *prediction* we mean guessing ex-ante, in the beginning of a month, the next future value of the monthly index return based on the values of the risk factors available at that moment. By *explaining* we mean establishing the relationship, at the end of a month, between the index return value of the past month and the values of the risk factors known at that moment. In other words, by explaining, we intend to assess the usefulness of market risk factors in tracking hedge fund index returns.

The remainder of the paper is structured as follows. We first illustrate the methodology behinds the local linear, *performance-adaptive replication* design in Section 2.2. Then, we describe alternative replication strategies. We present the database and the choice of common market risk factors in Section 2.3. We discuss some implementation issues and a detailed characterization of the proposed replication techniques in Section 2.4. Empirical results are presented in Section 2.5. Furthermore, a robustness check our empirical findings is performed. A short conclusion is drawn in Section 2.6. Tables and Plots are displayed in Exhibit 1 and 2 respectively.

2.2 Methodology

This section describes the implementation of sequential investment theory (as discussed in Györfi, Lugosi, and Udina (2006)) for the purpose of hedge fund index replication. The approach extends the linear, factor-based model which constructs a replicating portfolio of hedge fund returns from market risk factors. In the classical linear approach, the proportion of the wealth to be invested in each risk factor is obtained by regressing the vector of past index returns on the corresponding past risk factor values. In contrast to the classical approach, a sequential investment replication strategy identifies first those months in the past where the state of the market as described by a set of the risk factors, was 'similar' to the current conditions. Only the risk factor values and the hedge fund price information from these time instances are then used in the linear regression that determines the weights of the market risk factors in the current replicating portfolio.

Since we do not have *a priori* knowledge of the variables (risk factors, prices) that are instrumental in defining the state of the market, a refinement of the described methodology puts in competition a number of elementary sequential strategies called 'experts'. Each of these elementary sequential strategies is characterized by a *given set of relevant variables* (and a depth of the history of these variables) whose values are used to define 'similar' market conditions. Each 'expert' hence proposes its own replication portfolio, based on the time instances that it identified as 'similar' to the replication moment. The final replication decision that determines the contributions of each relevant risk factor to the replicating portfolio follows then an adaptive, time varying allocation of resources between these 'experts'. The allocation is based on their recent past performance (see Györfi, Lugosi, and Udina (2006)). What follows provides a detailed description of the methodology that we implement in the sequel.

The basic underlying assumption of the a linear replication strategy is the existence of a set of market risk factors, which dynamically explain the returns on a hedge fund index. Formally, we assume that the hedge fund index returns r_t can be linearly decomposed into the sum of several components as follows:

$$r_t = \beta_0 + \sum_{j=1}^{d} \beta^{(j)} MarketRiskFactor_t^{(j)} + \epsilon_t. \tag{2.1}$$

The set of market risk factors is fixed and specific to each style of investment. There are d market risk factors; $\beta^{(j)}$, $j = 1, \ldots, d$ are the factor loadings; and ϵ is the idiosyncratic part of the return. At each time $t = 1, 2, \ldots,$ knowing the past sequence of hedge fund index returns $\boldsymbol{r}_t = (r_1, \ldots, r_t)'$ and the past matrix of market risk factors $\boldsymbol{F}_t = (\boldsymbol{f}_t^{(1)}, \ldots, \boldsymbol{f}_t^{(d)})$, we want to explain r_t and to predict the next value r_{t+1} of a hedge fund index returns. The jth column $\boldsymbol{f}_t^{(j)} = (f_1^{(j)}, \ldots, f_t^{(j)})'$ represents the past values of jth risk factor, $j = 1, 2, \ldots, d$.

Following a classical regression setting, the factor-based replication approach at time t is carried out in two steps. At step one, risk factor exposures are estimated on past data running the regression:

$$r_i = \beta_0 + \sum_{j=1}^{d} \beta^{(j)} f_i^{(j)} + \varepsilon_i, \quad i = 1, 2, \ldots, t-1.$$

Given the estimated loading factors $(\widehat{\beta}_0, \widehat{\beta}^{(j)})$ from step one, the replicated return is identified in step two as:

$$\widehat{r}_t = \widehat{\beta}_0 + \sum_{j=1}^{d} \widehat{\beta}^{(j)} f_t^{(j)}.$$

One powerful critique to the simple approach outline above is its static nature. It is well known that hedge fund managers enjoy extraordinary flexibility and employ a large range of dynamic and complex investment strategies to achieve their return target. Hence, using a static methodology which estimates the exposure to the market risk factors on the whole past data, as the one described above, might fail to capture the dynamic features that most likely characterize the hedge fund return series. A number of responses to this critique have been proposed in the recent hedge fund return replication literature: rolling windows regression, regime-switching, Kalman filter and particle filter approaches (see for instance (among others), Roncalli and Teïletche (2008), Amenc, Martellini, Meyfredi, and Ziemann (2009)). Often, those variations do not provide important gains in the out-of-sample replication performance. Additionally, the more complicated ones lose the relative simplicity and intuitiveness of the purely linear factor-based approach.

In an attempt to account for the dynamic nature of the exposure of hedge fund index returns to market risk factors, we advocate a locally linear, *performance-adaptive replication* framework. The approach is conceptually simple and has the advantage of preserving the intuitive appeal of the factor-based model. The working hypothesis motivating the sequential algorithm is rather intuitive: the exposure of hedge fund investments to the set of style-relevant market risk factors should be similar during periods that are characterized by similar market conditions. This hypothesis engenders two important changes in the linear methodology.

The first change affects the aforementioned step one. One introduces a data selection process which individuates the periods characterized by market conditions 'close' to the current ones. Only the hedge fund index returns and the market risk factors of the periods identified by this data selection process are deemed suitable to contribute to the estimation of the loading factors for the current replication portfolio.

Formally, the similarity between the market conditions characterizing two time periods is measured by the Euclidean distance between the values of a set of pre-defined variables[5]. In other words, the values these variables take at a given moment t are a proxy for the market conditions at time t. For the simplicity of presentation we suppose that this vector is one-dimensional, i.e. the value of one given variable (risk factor or return)

[5]These variables could be market risk factors or returns. The issue of the choice of the set that functions as a proxy for the market conditions is discussed in the sequel.

provides a good proxy for the market conditions[6]. Let us denote by s_t the value of this variable at time t and the sequence of past k values of the market condition variable by $\boldsymbol{s}_{k,t} = (s_{t-k+1}, \dots, s_t)$. For a given d small, define[7]

$$S_t^{(k)} = \{k \le i \le t : \ \| \boldsymbol{s}_{k,i} - \boldsymbol{s}_{k,t} \| \le d\} \tag{2.2}$$

the set of all past time instances when the market conditions as measured by the k past values of the variable s were similar to the current market conditions at time t.

If $\boldsymbol{r}_{S_t^{(k)}}$ is the vector of hedge fund index returns that correspond to the instances $i \in S_t^{(k)}$, $\boldsymbol{F}_{S_t^{(k)}}$ the matrix of ones in the first column and d column vectors of market risk factors which belong to the same instance as the hedge fund return series, and $f_t^{(j)}$ the value of the j-th risk factor at replicating time period t, we define the vector of loading factors by running an OLS regression as:

for explaining hedge fund index returns:

$$\widehat{\boldsymbol{\beta}}^{ex}(S_t) = \underset{\boldsymbol{\beta}^{ex}}{\operatorname{argmin}} \| \boldsymbol{r}_{S_t^{(k)}} - \boldsymbol{F}_{S_t^{(k)}} \boldsymbol{\beta}^{ex} \| .$$

Hence, the explained return $r_t^{ex,k}$ is defined by:

$$r_t^{ex,k} = \widehat{\beta}^{(ex,0)}(S_t) + \sum_{j=1}^{d} \widehat{\boldsymbol{\beta}}^{(ex,j)}(S_t) f_t^{(j)}. \tag{2.3}$$

We explain the hedge fund index return r_t at time t as a linear combination of the market risk factors $f_t^{(j)}$.

for predicting hedge fund index returns:

$$\widehat{\boldsymbol{\beta}}^{pred}(S_t) = \underset{\boldsymbol{\beta}^{pred}}{\operatorname{argmin}} \| \boldsymbol{r}_{S_t^{(k)}} - \boldsymbol{F}_{S_t^{(k)}-1} \boldsymbol{\beta}^{pred} \|$$

$$\text{, where } S_t^{(k)} - 1 = \{i - 1 : i \ \in \ S_t\}.$$

Finally, for a given set of variables describing the market conditions S_t, the predicted return $r_{t+1}^{pred,k}$ is defined as:

[6]Later, in the data analysis discussed in Sections 2.4 and 2.5, market conditions proxies will be a vector valued time series whose components at time t are selected from the hedge fund return r and the risk factors f. Various set of combinations between risk factors and index returns are used depending of the hedge fund index under study. The process yields a range (7 for Equity Hedge and Merger Arbitrage, 11 for Barclay CTAs and for HFR Global Macro index) of variables used to characterize the state of the market. For example, in explaining/predicting Equity Hedge index, we defined a combinations of market condition variables based on 2 risk factors and index returns as follows: 3 vectors (hedge fund index r, risk factors $f^{(1)}$ and $f^{(2)}$); 4 matrices (built on factors ($f^{(1)}$, $f^{(2)}$); (r, $f^{(1)}$); (r, $f^{(2)}$); and (r, $f^{(1)}$, $f^{(2)}$)).

[7]$\| \cdot \|$ stands for the Euclidean norm.

$$r_{t+1}^{pred,k} = \widehat{\beta}^{(pred,0)}(S_t) + \sum_{j=1}^{d} \widehat{\beta}^{(pred,j)}(S_t) f_t^{(j)}. \tag{2.4}$$

We predict the hedge fund index return r_{t+1} at time t as a linear combination of the market risk factors $f_t^{(j)}$.

The second difference with the linear methodology consists in a aggregating the replicating portfolios yielded by the

- different sets of pre-defined variables that measure the similarity between the market conditions characterizing two periods as well as
- different distances d_l for measuring the similarity of the market conditions.

This aggregation is based on the past performance of each elementary replication strategy to produce a final replicating portfolio. To include this generalization in our notations, returns in Equations (2.3) and (2.4) will be denoted in the sequel by 3 indexes $*, k, l$, where $*$ stands for explained (ex) or predicted (pred) returns; k for the lookback period and l representing the choice of market condition variables.

This addition is of a particular importance as it addresses the issue of the choice of the market conditions proxy, i.e. the set of variables to be used to identify the time instances that are characterized by similar market conditions. Since we do not have *a priori* knowledge to help us determine the composition of such a set, the sequential method lets the data decide which is the most successful market conditions proxy, i.e. the configuration that most successful identifies the similarities in market conditions. The performance of each set is measures by the accuracy of the explanation/prediction of past returns. We say that one set of variables is better at measuring the similarities in market conditions if the replicating portfolio associated with it better explains/predicts the past returns. This performance measure is dynamic and, as a result, at different instances in time, different sets of variables can be best picking up the similarities in the market conditions. The replicating portfolio at time t is constructed to resemble those produced by the most successful sets. In what follows we provide a more formal description of this additional step.

Following Györfi and Lugosi (2001), the performance of each of the different implementations is measured by the *cumulative square replication error*

$$L_t(r^{*,k,l}) = \frac{1}{t} \sum_{i=1}^{t} \left(r_i^{*,k,l} - r_i \right)^2, \tag{2.5}$$

where $*$ stands for predicting or explaining hedge fund returns.

Finally, the structure of the final replication portfolio is obtained by combining the array of explained/predicted values $r_t^{ex,k}/r_{t+1}^{pred,k}$ defined in equations (2.3) and (2.4) according to their performance (equation (2.5)):

$$r_t^{ex} = \sum_{k,l} v^*_{k,l,t}\, r_t^{ex,k,l}, \text{ and } r_{t+1}^{pred} = \sum_{k,l} v^*_{k,l,t}\, r_{t+1}^{pred,k,l},$$

where the weighing function $v^*_{k,l,t}$ must satisfy the usual constraints for combining a portfolio of concurrent estimates, namely $0 \leq v^*_{k,l,t} \leq 1$ and $\sum_{k,l,t} v^*_{k,l,t} = 1$.

Given $\{q_{k,l}\}$, a probability distribution on the set of all pairs (k,l) of positive integers such that for all k and l, $q_{k,l} > 0$ and fixing the learning parameter[8] $\eta_t > 0$, the weights are defined by:

$$w^*_{k,l,t} = q_{k,l} e^{-\eta_t (t-1) L_{t-1}(r^{*,k,l})}$$

and their normalized values are then expressed as

$$v^*_{k,l,t} = \frac{w^*_{k,l,t}}{\sum_{i,j=1}^{\infty} w^*_{k,l,t}}. \tag{2.6}$$

The key idea of combining several concurrent replicators is simple: to improve the final replication decision. Practically, the worse an estimate performed in the recent past, the less it will contribute to the final replication value. The final estimate is constantly updated according to the recent performances of competitor estimates.

2.2.1 Alternative Replication Strategies

We compare the sequential investment methodology implemented as described above with a linear replication approach as well as with a simple extension that is intended to account for possible changes in market risk factors exposures, the so called rolling-windows regression. For the former technique, at each prediction/explaining time t, the procedure uses all the past information available $\boldsymbol{X}_t = (\boldsymbol{r}_t, \boldsymbol{F}_t)$ to estimate the loading factors. This procedure is equivalent to the performance adaptive factor-based rule, with $d = \infty$, the linear replicator that uses the full history to estimate the model and guesses the future value of hedge fund index return. Consequently, for $t > k+1$,

- explaining hedge fund index returns using the classical linear method:

$$\widehat{\boldsymbol{\beta}}^{ex}(\boldsymbol{X}_t) = \underset{\boldsymbol{\beta}^{ex}}{\operatorname{argmin}} \, || \, \boldsymbol{r}_t - \boldsymbol{F}_t \boldsymbol{\beta}^{ex} \, ||^2$$

[8] Throughout this paper, we fixed the leaning parameter η_t as $1/\sqrt{t}$. For a deep discussion on the best practical choice of η_t, see Biau, Bleakley, Györfi, and Ottucsak (2009).

The linearly explained returned is given by:

$$r_t^{ex} = \widehat{\beta}^{(ex,0)} + \sum_{j=1}^{d} \widehat{\boldsymbol{\beta}}^{(ex,j)} f_t^{(j)}.$$

- predicting hedge fund index returns using the classical linear method:

$$\widehat{\boldsymbol{\beta}}^{pred}(\boldsymbol{X}_t) = \underset{\boldsymbol{\beta}^{pred}}{\operatorname{argmin}} \mid\mid \boldsymbol{r}_t - \boldsymbol{F}_{t-1}\boldsymbol{\beta}^{pred} \mid\mid^2$$

  The linear predicted value is given by:

$$r_{t+1}^{pred} = \widehat{\beta}^{(pred,0)} + \sum_{j=1}^{d} \widehat{\boldsymbol{\beta}}^{(pred,j)} f_t^{(j)}.$$

In the rolling-windows regression context, the factor loadings at time t are determined by regressing the hedge fund return series from $t - \tau + 1$ to t on the set of market risk factors associated. The determination of the window length, τ, is arbitrary and implies a trade-off between the dynamic characteristic of the coefficients estimated and statistical accuracy.

Therefore, for a chosen τ and $t > \tau$

- explaining hedge fund index returns using rolling-windows regression:

$$\widehat{\boldsymbol{\beta}}^{ex}(\boldsymbol{X}_{\tau,t}) = \underset{\boldsymbol{\beta}^{ex}}{\operatorname{argmin}} \mid\mid \boldsymbol{r}_{\tau,t} - \boldsymbol{F}_{\tau,t}\boldsymbol{\beta}^{ex} \mid\mid^2$$

  The rolling-windows explained value is given by:

$$r_t^{ex} = \widehat{\beta}^{(ex,0)} + \sum_{j=1}^{d} \widehat{\boldsymbol{\beta}}^{(ex,j)} f_t^{(j)}.$$

- predicting hedge fund index returns using rolling-windows regression:

$$\widehat{\boldsymbol{\beta}}^{pred}(\boldsymbol{X}_{\tau,t}) = \underset{\boldsymbol{\beta}^{pred}}{\operatorname{argmin}} \mid\mid \boldsymbol{r}_{\tau,t} - \boldsymbol{F}_{\tau,t-1}\boldsymbol{\beta}^{pred} \mid\mid^2$$

  The predicted value is derived as:

$$r_{t+1}^{pred} = \widehat{\beta}^{pred} + \sum_{j=1}^{d} \widehat{\boldsymbol{\beta}}^{(pred,j)} f_t^{(j)}.$$

Another plausible alternative choice of model could has been a model where lagged hedge fund historical returns are added to the set of risk factors. However, our objective is not to find the best possible model but to compare our dynamic procedure with the static one well discussed in the literature.

To assess the performance quality of each prediction/explaining mechanism, we use four criteria:

1. *Mean Square Replication Error* (MSE): To quantify the difference between the realized returns and the values derived by the replication technique.

2. *Cumulative Square Replication Error* (CMSE): To illustrate the time-varying evolution of mean square error. The figure highlights the historical performance of each estimation.

3. *Percentage direction of change* (DOC): Computes the directions of changes (positive or negative) which are correctly estimated.

4. *Correlation between the estimated series and the realized series* (CORR): Assess the statistical relationships between the estimated and real series.

2.3 Data

This section presents the data used in our empirical investigation. We begin by discussing the hedge fund indexes we try to replicate. Then, we comment the choice of relevant style-specific market risk factors.

2.3.1 Indexes

The analysis is performed on 5 hedge funds indexes: 3 Chicago-based Hedge Fund Research (HFR) indexes and the Barclay CTAs (Commodity Trading Advisors) index.

The Barclay CTAs database is widely recognized throughout the world as the largest, most comprehensive, available Commodity Trading Advisors database. Generally, CTAs are funds primarily trading listed commodity and financial futures contracts. CTAs, also denominated *managed futures* are by no means homogeneous investment vehicles. CTA managers employ a large range of strategies and of asset classes. CTAs represent an important investment style within the universe of alternative-investment. While historically CTAs managers have been presented separate from hedge fund vehicles, the difference between the two has been constantly eroding. The Barclay CTAs index is the most followed and tracked benchmark by managed futures investors. In 2010, the index tracked 533 programs[9] worldwide.

HFR is one of the most widely used commercial database to which hedge fund managers voluntarily report their performances. HFR indexes

[9] This number was 491 in 2008 and currently, 602 CTAs are included in the index.

represent the global standard for performance measurement across all aspects of the alternative-investment industry. Our empirical analysis is implemented on three monthly HFR hedge fund indexes spanning different time interval: HFR Global Macro from January 1994 to December 2009, Equity Hedge from January 1994 to April 2010 , Merger Arbitrage from January 1994 to March 2010

Indexes are constructed using robust methods and qualitative selection process[10]. While many researchers have been aware of potential biases such as survivorship bias, liquidation bias, back-fill bias, and selection bias in hedge fund databases resulting from the nature of the data collection process, we believe the indexes possess sufficient data quality to perform our analysis. Additionally, since we focus on replicating hedge fund indexes, any relevant biases should affect the replicated and the original series identically.

2.3.2 The choice of risk factors

As mentioned in the introduction, we use sets of strategy-specific market risk factors to replicate returns of different-style hedge fund indexes. Our specific choice of the strategy-specific sets is informed by the extensive empirical literature on the topic which offers strong evidence of style-specific factors sensitivities. Returns of different investment styles are exposed to different sets of relevant market risk factors. A discussion of the sets of strategy-specific market risk factors follows.

Trend-following strategies. Our approach generalizes a restricted version of Fung and Hsieh (2001) model specification. The authors extend the pioneer work on mutual funds investment vehicles Sharpe (1992) to model the returns of managed futures funds. They shown that trend-following strategies have returns similar to a dynamically managed option-based strategy called a look-back straddle. As a consequence, the authors used four look-back factors on stocks, bonds, currencies and commodities to track CTA fund's returns.

Long/short equity strategies. (Fung and Hsieh, 2004b) highlight the strong positive exposure of this type of funds to the equity market as well as long/short investment dynamics in small-cap/large-cap stocks.

Directional strategies. To describe the risk structure of such funds, such as Global Macro, (Fung and Hsieh, 2004a) advocated a seven factors model: two equity factors (the excess return of the S&P 500 and the spread between the Small Cap index and the Large Cap index), two bond factors (the spread between the ten-year US Treasury bond and the three-month US Treasury bill and the return of the Baa corporate bond above the ten-year US Treasury bond) and the three look-back factors on currencies,

[10]Visit: http://www.hedgefundresearch.com, http://www.barclayhedge.com for detail information on index construction methodology.

bonds and commodities. Our approach generalizes a restricted version of the linear approach described in this article. We implemented the more parsimonious version, using 4 factors on the HFR Global Macro index.

Merger arbitrage strategies. Mitchell and Pulvino (2001) empirically show that the returns stream of such hedge funds exhibit similar statistical properties as the monthly excess return of S&P 500 while Fung and Hsieh (2007) pointed the existence of a US mutual fund, called the Merger Fund with an investment strategy similar to this class of risk arbitrageurs.

To summarize, in line with the academic consensus, we use the appropriate subset of market risk factors among the following observed market risk factors[11] as regressors to explain and predict hedge fund index returns:

1. Barclay CTAs index: monthly returns from January 1994 to September 2008.
   - 4 Fung-Hsieh look-back straddle factors: bond, commodity, stock, and short term interest rate.
2. HFR Global Macro Hedge fund index: monthly returns from January 1994 to March 2010.
   - 4 factors: S&P 500 index return, Fung-Hsieh look-back straddle commodity factor, TED spread and the Major currency dollar index.
3. HFR Equity Hedge fund Index: monthly returns from January 1994 to March 2010.
   - 2 spread factors: S&P 500 index return - one-month US Treasury bill rate, Russell 2000 index return - S&P 500 index return.
4. HFR Merger Arbitrage Index: monthly returns from June 1994 to March 2010.
   - 2 factors: S&P 500 index return - one-month US Treasury bill rate and the returns of the Merger Fund.

Needless to say, the factor model specifications discussed above are strong assumptions in our empirical investigation. By setting in advance and by keeping unchanged the set of market risk factors, we implicitly assume that a given hedge fund strategy is exposed to the the same risk factors. The regressor selection issue is not new in the empirical financial

[11]Data are downloaded from various sources: look-back straddle factors from Fung-Hsieh data library; S&P 500 and the three month US Treasury bill from Fama-French data library; Russell 2000 equity and the Merger Fund indexes from yahoo finance; US Treasury 10y, Moody's Baa, TED spread and the Major currency dollar index from the FED data library; USD Barclay Global High Yield bond index from DataStream.

literature and there is no consensus on the way to deal with the matter. In this article, since the objective is to investigate the performance-adaptive, factor-based rules in replicating hedge fund index returns compared to a linear factor based approach, we make abstraction of the factor selection issue. That is, we use well defined, established set of observable market risk factors. To guarantee consistent out-of-sample performance we are as parsimonious as possible in term of the number of factors. Ultimately we make sure that the factors exhibit a low cross-correlation profile and are economically interpretable in the style-specific framework. Therefore, for each hedge fund strategy, a factor will be discarded from the predictor list, if it does not provide sufficient prediction ability. A last caveat: if the interest of the researcher is to replicate hedge fund returns for an investment purpose, factors such as the Fung-Hsieh look-back straddles would be difficult to implement. Replication rules which use those factors should consider more liquid and simple to implement alternatives.

2.3.3 Data summary statistics

As a preliminary data analysis, we document in Table 2.1 several univariate summary statistics on the four hedge fund index under study. First, we concentrate on some. Panel A provides descriptive statistics of the returns series which include location (mean, median), dispersion (standard deviation), shape (skewness, kurtosis) and extremes (minimum, maximum) parameters of each index distribution. The mean and the median return are positive for all indexes. Monthly dispersion is larger for CTAs, HFR Global Composite and Equity Hedge indexes than for the HFR Merger Arbitrage index. For HFR Global Composite, and Merger Arbitrage, the minimum is larger in absolute value than the maximum. This suggests that for those indexes, the extreme down movements are more pronounced than the most extreme up movements. Both the most positive and as well as the most negative extreme values, for all four indexes are large as compared with the standard deviation of the series. These findings hint at the facts that the distribution of returns is asymmetric and has fat tails. This is confirmed by the value of the skewness and excess kurtosis: HFR Global Composite, Equity Hedge, Merger Arbitrage indexes are negatively skewed and all indexes have the kurtosis larger than 3, the value expected from the a normal distribution.

Finally, Panel B of Table 1 provides the Jarque-Bera test statistic for the hypothesis of normally distributed returns. For all indexes, except Barclay CTAs, the Jarque-Bera test indicates the null hypothesis of normally distributed is rejected at 5% significance level.

2.4 Implementation issues

This section presents the concrete details of the implementation, i.e choice of parameters and the algorithm description. As explained in the methodology section, replicating portfolios are constructed for different proxy of market conditions, different number of past values of the proxy used in measuring the similarity of market conditions (parameter k) as well as for different values of the maximal distance between the values of these proxy (parameter $d_{k,l}$). The choice of the last two parameters is as follows: $k = 1, \ldots, K$ and $l = 1, \ldots, L$, fixing $K = 3$ and $L = 10$. Additionally we used the uniform probability distribution $\{q_{k,l}\} = 1/(K \times L)$ for the weighting scheme of replication portfolios indexed over k and l. The maximal distance $d_{k,l}$ is defined by:

$$d_{k,l}^2 = 0.0001 \cdot 1 \cdot l \cdot k$$

Hence, for each k, we have $l = 10$ different values of the radius $d_{k,l}$ and a set of 10 different replicating portfolios.

For all the indexes the replication exercise starts in May 2001, i.e. roughly the first seven years of the samples are used to provide the needed initial estimates. For every time series of hedge fund index returns $(r_1, \ldots, r_T)$ and for every $88 < t \leq T - 1$ (here, $T = 171$ or 195) we use the past part of the sample $(r_1, \ldots, r_t)$ of hedge fund returns and market risk factors to identify the past time instants i displaying similar market conditions to t the current time period, and explain/predict the value $r_t^{ex,k}/r_{t+1}^{pred,k}$. The similarity set is defined as follows,

$$S_t^{(k,l)} = \{k \leq i \leq t : \ || \boldsymbol{s}_{k,i} - \boldsymbol{s}_{k,t} || \leq d_{k,l}\}$$

where by s_t we denoted the value of the market conditions proxy variable at time t and by $\boldsymbol{s}_{k,t} = (s_{t-k+1}, \ldots, s_t)$, the sequence of past k values of the market condition variable

For a fixed k and t, as l increases, $d_{k,l}$ increases and the similarity set $S_t^{(k,l)}$ gets bigger and bigger. Consequently, the elementary model converges to the linear replication using the full history.

A practical restriction is imposed on the similarity set $S_t^{(k,l)}$, a minimum size requirement. We restricted each elementary strategy to have a minimum of 24 data points (2 years) to estimate the regression parameters. By controlling the size of local information used to construct the estimates, and allowing the time-varying replicators to become more informative as the sample size increases, we can ensure that our results are consistent under standard regularity conditions.

The choice of market risk factors does not depend of parameters, k, l. Indeed, they are set of variables well defined in the hedge fund literature (refers to Sections 2.1 and 2.3). Depending of the hedge fund index under study, market conditions proxies are identified on various set of combinations between risk factors and index returns. At each prediction time, the strategy identifies the set of vectors (index return and market risk factors) in the past that are 'close' to the actual vector according to Equation(2.7). The final market conditions proxies is then a mix of past hedge fund returns and risk factors. The process yields a range (7 for Equity Hedge and Merger Arbitrage, 11 for Barclay CTAs and HFR Global Macro index) of variables used to characterize the state of the market. Every market proxy provide a portfolio of $K \times L$ elementary replication strategies.

Additionally, each elementary replication strategy is evaluated according to their recent past performances. Then, a performance-adaptive combining procedure is implemented.

Used for comparison purpose, our benchmarks the classic and rolling-window linear predict/explain schemes are implemented as described in Section 2.2. For the former, at each time t, all the past information available $\boldsymbol{X}_t$ is collected to estimate the time-varying factors sensitivities and for the latter, we fix, as it is common in the empirical literature, the window length τ to 48 months. This procedure is still dynamic, but far less flexible as the performance-adaptive factor-based rule.

To conclude the section, we give a short description of the algorithm implementing the *performance-adaptive replication* strategies.

Algorithm 1: Performance-Adaptive Replication

At each replication time t, given $\boldsymbol{r}_t$, the vector of past returns and $\boldsymbol{F}_t$, the matrix of past risk factors:

1. *For each market conditions proxy* $\mathbf{s}$ *and for each pair* (k,l), $k = 1, \ldots, K$ *and* $l = 1, \ldots, L$, *compute the history list* $S_t^{(k,l)}$. *Run the appropriate regressions to obtain* $\widehat{\boldsymbol{\beta}}^*(S_t)$. *Calculate* $r_t^{ex,k,l}$ $(r_{t+1}^{pred,k,l})$, *the explained (predicted) return and* $L_t(r^{*,k,l})$, *the cumulative mean square error (CMSE) .*

2. *Combine individual explanations/predictions: exponentially pool the estimates according to their past performances (CMSE), a probability distribution* $q(k,l)$ *over the set of all proxy and all pairs* (k,l) *and the learning parameter* η_t*:*

$$r_t^{ex} = \sum_{\mathbf{s},\, k,l=1}^{K,L} v_{k,l,t}^{ex}\, r_t^{ex,k,l}, \text{ and } r_{t+1}^{pred} = \sum_{\mathbf{s},\, k,l=1}^{K,L} v_{k,l,t}^{pred}\, r_{t+1}^{pred,k,l}.$$

3. *t=t+1.*

2.5 Empirical results

This section presents the results of our empirical analysis. We first document the explaining/prediction performances of our replication strategy (performance-adaptive) against the two alternative benchmarks: the purely linear 'static' and rolling-windows factor-based strategies. The power of each strategy is assessed according to the criteria outlined in Subsection 2.2.1. Next, we report the market risk factor exposures in (2.1) estimated for each of the four indexes. In particular, we examine to what extend the selected factors contribute to the replication. We then describe the impact of various proxy of market conditions to the final explained/predicted value. Each elementary sequential strategies are weighted according to Equation (2.6). Finally, we test the results of our proposal under a more restrictive experimental setting in which the data selection window is fixed in the past.

2.5.1 Explaining/Prediction performances

To measure the quality and compare the three methods, we used a range of four criteria. First, we compute the mean square replication error *MSE*. Secondly, as described in Equation (2.5), we calculate the cumulative square replication error. Thirdly, we report the percentage of explained/predicted values that correctly guess the direction of hedge fund index return movements. We will assume that with a percentage direction of change (*DOC*) larger or equal to 50%, the strategy has the ability to estimate the direction of the index movements. Finally, to assess the relationship between the explained/predicted series and the realized one, we calculate the Spearman's correlation coefficient (*CORR*).

Table 2.2 in Exhibit 1 shows that, for all hedge fund categories under study our proposed replication strategy performs significantly better than the two alternative benchmarks in explaining/predicting monthly hedge fund index returns. The performance of the three specifications is summarized by the relative mean square replication error. The ratio[12] is computed as MSE^a/MSE^o. A ratio bigger than one indicates that the performance of the alternative benchmarks in replicating hedge fund index returns is poorer than that based on the performance-adaptive proposal. The results are more striking for Barclay CTAs and HFR Equity Hedge indexes. Our replication rule is approximately 10% to 400% better in explaining, and around 50% to 430% superior in predicting than the two alternative strategies according to the relative *MSE*. The Ratio column in Table 2.2, shows that this gap is even more pronounced for the month-to-month prediction strategy. This suggests that both 'static' and rolling-windows model specifications fail to predict the future value of hedge fund returns. This failure is directly associated to the risk factor characterizations. We shall return to this important issue in the sequel.

Panels A.2 to D.2 of Table 2.2, display the Spearman's correlation coefficient. The figures highlight a pronounced statistical dependence between the performance-adaptive replicated series and the realized hedge fund indexes returns. The *CORR* values for the predicted series ranged from 95% for Barclay CTAs, 92% for Equity Hedge, 80% for Global Macro, and 66% for Merger Arbitrage indexes. For the 'static' setting, correlation

[12] Where 'a' stands for alternative benchmarks and 'o' for our replication strategy.

coefficients are low and range from around −15% for the Barclay CTAs index, to 11% with the Equity Hedge and Global Macro indexes. The results for the explained series are of comparable magnitude, though in this case both alternative benchmark replication strategies show a better performance. The direction of change values are systematically bigger than 60% for the performance-adaptive strategy (for example, 95% for the predicted Barclay CTAs index and 85% for Equity hedge fund index) and again, generally larger than in the classical linear method, specially for the predicted cases.

Compared to the rolling-windows[13] factor-based rule, the proposed methodology performed even better. The magnitude of the dominance is larger for all four quality measures adopted in the current article. Across all hedge fund indexes under study, the rolling-windows replication routine performs worse than the classical linear.

Table 2.2 shows that although the performance of our empirical investigation varies considerably across hedge fund categories, the *performance-adaptive replication* strategy exhibits superior quality under the four comparison criteria assumed in the present study. The implemented experiment spans 4 distinct hedge fund categories from directional funds (Equity Hedge index, Barclay CTAs, and Global Macro) to arbitrage fund (Merger Arbitrage) and our findings are pertinent around all styles. The performance can also be appreciated through inspection of third-row entries in many figures. When inquiring those plots, we see how our explained/predicted series closely track the realized ones. Our empirical findings seems to corroborate the fact that a sizeable percentage of hedge fund index returns could be replicated ex-ante or explained. In the next section we investigate the contributions of various market risk factors in the replication process.

2.5.2 Factor exposures

Having established the superiority of the *performance-adaptive replication* strategy in explaining/predicting the different categories of hedge fund index returns in our sample, we document the drivers behind this performance. The main interest is in the estimated coefficients related to each replication strategy. For that purpose, we report the value for the t-statistic associated to each beta estimated.

At every replication time, one performs a regression many times according to the number of market condition proxies. The value of the t-statistic for each loading is obtained across the 4 hedge fund indexes under scrutiny. Figures 2.1, 2.2, 2.4, 2.5, 2.7, 2.8, 2.10, and 2.11 depict the distribution of the t-statistics for the estimated factor exposures. The

[13] In the rolling-windows replication strategy, we have tried various window sizes (τ) of 24 (as in Hasanhodzic and Lo (2007)), 36, 48, and 60, without noticing any improvement. We report the results of window size equal to 48.

dotted vertical lines represents the left and right quantiles at 5% level of significance. The t-statistic less than 1.96 in absolute value - between the the vertical dotted lines - means that the loading are not statistically significant.

Starting with predicting hedge fund returns, we observe along the histograms of t-statistics that very few values fall outside the vertical dot lines. The risk factor specifications do not provide sensible contribution to the replicated series. The picture is approximately identical for all the indexes. A close inspection of the second-row entries of Figures 2.2, 2.5, 2.8, and 2.11 reveals that the prediction quality of the *performance-adaptive replication* strategy is due to the estimated $\hat{\beta}_0$, also called *alpha* in the financial literature. The factors that we use, although used by the literature, are not significant. The quality of the replication exercise relies entirely in the capacity of each sequential strategy to forecast the index's *alphas*.

If we turn to the explaining part, all three model specifications exhibit relatively good performances in three indexes with the exception of the HFR Global index. This can be seen through row 3 of Figures 2.1, 2.4, 2.7, and 2.11. In general, the estimated loadings are more often statistically significant at 5% than in the prediction case. From the histograms, we observe a large proportion of t-statistic values on the left and right side of the dot vertical lines. These results are in lined with several contributions in which it is demonstrated that, the multiple factor model analysis explain a portion of hedge fund returns (See among others, Fung and Hsieh (2006), and Hasanhodzic and Lo (2007)). The superior performance of our replication strategy resides on its ability to explain the time varying evolution of index's *alphas*. A good example to illustrate this feature is given by the second and third-row entries of Figures 2.4 and 2.6 respectively. The figures display the histograms of estimated *alphas*'s t-statistic and the dynamics of the explained HFR Equity Hedge index's *alphas* respectively. While the three replication strategies show significant t-statistics from the second-row entries of Figure 2.4, we see from the lower part of Figure 2.6 that the classical linear and rolling-windows replications miss to uncover the time varying index *alphas*, getting only decreasing *alphas* estimated.

Regarding the HFR Global Macro index, our findings are mixed. In this case, we observe that the replication process is slightly more difficult. Though our strategy performs better than the 'static' linear and the rolling-windows, the factor model does not seen to work in this case. While the *performance-adaptive replication* strategy correctly identifies the index returns's direction of change, it fails to reproduce its variability. Unreported results using the Fung-Hsieh seven factors model specification do not provided further improvements. Figures 2.10 and 2.11 show that whereas the Fung-Hsieh primitive trend followers commodity and Major Currency US Dollar index factors exhibit statistically significant explanatory power, none factors have the structure to forecast Global Macro index returns.

The global picture is that the multiple factor model characterization exhibit more explanatory power than predicting capability. In both cases, the major sources of hedge fund index returns are not the estimated betas times the risk premia, but the style specific *alphas* associated.

2.5.3 Portfolio of elementary sequential strategies

To gain further insight into the performance-adaptive replication strategy, we present in the first-row entries of Figures 2.3, 2.6, 2.9, and 2.12 the dynamic of weights ($v_{k,l,t}$) associated to each estimate according to various proxy of market conditions. Every estimate comes out from the combination of $K \times L$ elementary sequential strategies. The contributions of each elementary strategy is determined by how good they performed in the past. The final replicated return is obtained from an exponential weighting technique as explained in Section 2.2. The flexibility of the replication process allows each elementary strategy to propose its own replication portfolio, characterized by various set of input variables. Visual inspection of the weight dynamics indicates a strong heterogeneity among sequential strategies, across hedge fund indexes. Although for HFR Equity Hedge, Merger Arbitrage, and Global Macro indexes practically all defined proxies contribute to the final replication decision, for Barclay CTAs just two out of eleven contribute. However, for all indexes under study the first elementary sequential strategy tends to be the most successful replicator, having larger weights according to its good performances. Strategy number one is the elementary strategy which individuates the market conditions that are 'similar', from the hedge fund index historical return series. This finding means that the most reliable information of future market conditions for explaining/predicting hedge fund returns arose from the index itself, and not from market risk factors.

2.5.4 Robustness analysis

To further evaluate the robustness of our findings, we have implemented a more restrictive empirical analysis, the principal results of which are presented in this section.

We divided the data in two non overlapping periods with 12 months gap between the two data sets: a fixed training set (from January 1994 to December 2003), where we apply the data selection scheme and get the regression coefficients; and a test set where we explain/predict (from January 2005 to the end of the sample, $T = 171$ or 195). Therefore given the time series vector of hedge fund index returns $(r_1, \ldots, r_T)$, for each $131 \leq t \leq T - 1$ we used the fix training set of data $(r_1, \ldots, r_{119})$ to extract the time instances that are 'close' to the current one. Consequently the elementary predicted (explained) values $\hat{r}_{t+1}$ ($\hat{r}_t$) of monthly hedge fund indexes are obtained by regressing the string of data collected in the fixed past. In other words, instead of moving the data selection interval as the replicating time t increases, we fixed it to be a given time interval. The final replication values is derived by weighting individual strategies according to their past performances as explained in Section 2.2. It follows that, as t grows, we will look for 'similar' market conditions in a more distant past.

The results of the *performance-adaptive replication* strategy under this restrictive setting are reported in Table 2.3 and are consistent with the general findings outlined in Table 2.2. The performance-adaptive rules still perform better than the classical linear and rolling-windows methods under 'normal' replication framework (see Subsection 2.2.1). Unreported results concerning the coefficient t-statistics are in line with what expressed in Subsection 2.5.2.

The empirical robustness analysis confirms the conclusion of the moving-window performance-adaptive implementation. That is, the technique is superior to the standard 'static' and rolling-windows linear methods, in terms of the four criteria listed in the current article. Additionally, the quality of the performance-adaptive replicator is not undermined by how distant in the past we look for data which are 'similar' to those of the replication time.

2.6 Concluding remarks

The *performance-adaptive replication* strategy proposed in this paper is a methodology that extends the widely used purely linear and rolling-windows to the purpose of hedge fund index replication. The proposal enables to relax the strongly stationarity assumption in the classical linear replication approach. This extension allows us to identify the main sources behind the time-varying evolution of hedge fund returns.

We apply the local linear, performance-adaptive method to the replication of 4 monthly hedge fund index returns sampled over 17 years. We find that the proposed performance-adaptive setting is statistically superior to the standard linear ('static' and rolling-windows) framework in terms of mean square replication error, Spearman's correlation coefficient, and direction of change. We demonstrate the usefulness of the methodology as a tool to individuate predictive patterns amongst hedge fund indexes.

We also find that the factors used in the literature for hedge fund replication to be of little help in predicting future returns. However, their explanatory abilities are mixed. We show that *performance-adaptive replication* approach can, occasionally, better forecast the dynamic changes of the aggregate fund style-specific *alphas*.

Acknowledgements

The authors are grateful to the participants of the international conference on mathematical finance and economics, Istanbul, and conference in computational financial econometrics, London, 2011 for their insightful comments and suggestions.

Exhibit 1

Table 2.1 – *The table presents some descriptive statistics and the Jarque Bera normality test at 5% level for six hedge fund indexes under study. In the upper part of the table, we present, for the various indexes, the monthly basic statistics and the lower part of the table reports the normality test. The analysis is performed on the entire data sample including Barclay CTAs index, HFR Global Composite, Equity Hedge, and Merger Arbitrage indexes. Bold P-value results highlight significance at 5% level.*

Panel A.1 General descriptive statistics				
	Global Macro Index	Barclay CTAs	Equity Hedge	Merger Arbitrage
Observations	192	177	196	195
Minimum	-0.06	-0.0477	-0.0946	-0.0569
Median	0.0064	0.0036	0.0112	0.0086
Arithmetic Mean	0.0077	0.0052	0.0095	0.0072
Maximum	0.0682	0.0645	0.1088	0.0312
SE Mean	0.0014	0.0017	0.0019	0.0008
St. deviation	0.0196	0.0228	0.0270	0.0108
Skewness	0.1311	0.2789	-0.2014	-1.7218
Kurtosis	0.9920	0.0853	2.2177	6.2781

Panel B.1 Jarque-Bera normality test				
	Global Macro Index	Barclay CTAs	Equity Hedge	Merger Arbitrage
Jarque-Bera Test	8.42	2.35	41.49	416.59
P-value	**0.01**	0.31	**0.00**	**0.00**

Table 2.2 – *Performance comparison between the* performance-adaptive replication *(our strategy), 'static' and rolling-windows regression setting on explaining/predicting monthly hedge fund index returns. The experiment is performed on* 4 *distinct hedge fund categories. The table presents the result derived combining* $K \times L$ *elementary replicators according to their recent past performances expressed by the* 24*-Months cumulative square replication error. The quality of each empirical scheme is assessed with* 3 *criteria: MSE (Mean Square Replication Error), DOC (the percentage of the direction of change correctly forecasted) and CORR (Correlation between the explained/predicted series and the realized series).*

Panel A.2 Barclay CTAs Index				
Models	Goals	MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.92	0.89
	Predicted		0.95	0.95
'Static' Setting	Explained	3.62	0.40	0.73
	Predicted	3.60	-0.15	0.53
Rolling-Windows	Explained	4.34	0.27	0.54
	Predicted	4.08	0.07	0.40

Panel B.2 HFR Equity Hedge Index				
Models		MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.89	0.80
	Predicted		0.92	0.85
'Static' Setting	Explained	1.38	0.84	0.85
	Predicted	2.17	0.11	0.44
Rolling-Windows	Explained	1.49	0.82	0.85
	Predicted	2.30	-0.08	0.32

Panel C.2 HFR Merger Arbitrage Index				
Models		MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.71	0.77
	Predicted		0.66	0.75
'Static' Setting	Explained	1.10	0.68	0.76
	Predicted	1.48	0.21	0.43
Rolling-Windows	Explained	1.24	0.64	0.74
	Predicted	1.59	-0.26	0.45

Panel D.2 HFR Global Macro Index				
Models		MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.70	0.83
	Predicted		0.80	0.87
'Static' Setting	Explained	1.43	0.03	0.35
	Predicted	1.57	0.11	0.11
Rolling-Windows	Explained	1.52	0.17	0.45
	Predicted	1.7	0.10	0.29

Table 2.3 – *Performance comparison between the* performance-adaptive replication, *'static' and rolling-Windows regression setting in explaining/predicting monthly hedge fund index returns. The experiment is performed on* 4 *distinct hedge fund categories in a more restricted framework, where the data selection and estimation is performed in a fix time interval in the past from January 1994 to December 2003. The replication is performed from January 2005. The table presents the result derived combining $K \times L$ concurrent estimates according to their recent past performances expressed by the* 24-*Months cumulative square replication error. The quality of each empirical scheme is assessed with* 3 *criteria: MSE (Mean Square Replication Error), DOC (the Percentage of the direction of change (positive or negative) correctly forecasted) and CORR (Correlation between the explained/predicted series and the realized series).*

Panel A.2 Barclay CTAs Index				
Models	Goals	MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.73	0.74
	Predicted		0.70	0.74
'Static' Setting	Explained	1.43	0.38	0.70
	Predicted	1.40	-0.34	0.36
Rolling-Windows	Explained	1.87	0.19	0.55
	Predicted	1.52	-0.16	0.20

Panel B.2 HFR Equity Hedge Index				
Models		MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.88	0.84
	Predicted		0.88	0.85
'Static' Setting	Explained	1.29	0.83	0.84
	Predicted	1.79	0.29	0.43
Rolling-Windows	Explained	1.38	0.81	0.84
	Predicted	1.92	0.14	0.32

Panel C.2 HFR Merger Arbitrage Index				
Models		MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.66	0.80
	Predicted		0.79	0.84
'Static' Setting	Explained	1.56	0.68	0.76
	Predicted	1.35	0.21	0.44
Rolling-Windows	Explained	1.75	0.63	0.73
	Predicted	1.47	-0.20	0.41

Panel D.2 HFR Global Macro Index				
Models		MSE^a/MSE^o	*CORR*	*DOC*
Our Strategy	Explained		0.75	0.72
	Predicted		0.75	0.83
'Static' Setting	Explained	1.46	0.23	0.54
	Predicted	1.84	0.10	0.17
Rolling-Windows	Explained	1.74	0.25	0.6
	Predicted	2.05	0.14	0.37

Figure Captions

Throughout Exhibit 2, the empty space on figures indicates those periods where the data selection process does not identify enough similar values in the past.

Figures 2.1, 2.2, 2.3: These figures display the results of the explained/predicted Barclay CTAs index returns. Figures 2.1 and 2.2 express respectively the findings: t-test statistics and the evolution of the final replication series. The two vertical dotted lines reflect the left and right quantiles at 5% level of significance. The t-statistic less than 1.96 in absolute value - between the the vertical dotted lines - means that the loading are not statistically significant. From left to right, 'static', rolling-windows, and the performance-adaptive replication strategies respectively. The factor model specification is defined by the Fung-Hsieh look-back straddle factors: ptsbd (bond factor), ptsir (short term interest rate factor), ptscom (commodity factor), and ptsstk (stock factor). Alpha stands for the abnormal returns. In Figure 2.3, the upper plots depict the evolution of the weights associated to various approximated market conditions, while the lower plots highlight the estimated index-specific alphas in explaining/predicting Barclay CTAs index.

Figures 2.4, 2.5, 2.6: These figures display the results of the explained/predicted HFR Equity Hedge. Figures 2.4 and 2.6 express respectively the findings: t-test statistics and dynamic of the final replication series. The two vertical dotted lines reflect the left and right quantiles at 5% level of significance. The t-statistic less than 1.96 in absolute value - between the the vertical dotted lines - means that the loading are not statistically significant. From left to right, 'static', rolling-windows, and the performance-adaptive replication strategies respectively. The factor model specification is defined by two factors: Rm-rf (excess return on the market, is S&P 500 index return minus the one-month Treasury bill rate) and russ-Rm (spread between the Russell 2000 index return and the S&P 500 index return). Alpha stands for the abnormal returns. In Figure 2.6, the upper plots depict the evolution of the weights associated to various approximated market conditions, while the lower plots highlight the estimated index-specific alphas in explaining/predicting HFR Equity Hedge index.

Figures 2.7, 2.8, 2.9: These figures display the results of the explained/predicted HFR Merger Arbitrage. Figures 2.7 and 2.8 express respectively the findings: t-test statistics and dynamic of the final replication series. The two vertical dotted lines reflect the left and right quantiles at 5% level of significance. The t-statistic less than 1.96 in absolute value

- between the the vertical dotted lines - means that the loading are not statistically significant. From left to right, 'static', rolling-windows, and the performance-adaptive replication strategies respectively. The factor model specification is defined by two factors: Rm-rf (excess return on the market, is S&P 500 index return minus the one-month Treasury bill rate) and MERF (return of the Merger Fund). Alpha stands for the abnormal returns. In Figure 2.9, the upper plots depict the evolution of the weights associated to various approximated market conditions, while the lower plots highlight the estimated index-specific alphas in explaining/predicting HFR Merger Arbitrage index.

Figures 2.10, 2.11, 2.12: These figures display the results of the explained/predicted HFR Global Macro. Figures 2.10 and 2.11 express respectively the findings: t-test statistics and dynamic of the final replication series. The two vertical dotted lines reflect the left and right quantiles at 5% level of significance. The t-statistic less than 1.96 in absolute value - between the the vertical dotted lines - means that the loading are not statistically significant. From left to right, 'static', rolling-windows, and the performance-adaptive replication strategies respectively. The factor model specification is defined by the four factors: S&P 500 index return (SP500), Fung-Hsieh look-back straddle commodity factor (ptscom), TED spread (TED.Spread) and the Major currency dollar index (MDINX). Alpha stands for the abnormal returns. In Figure 2.12, the upper plots depict the evolution of the weights associated to various approximated market conditions, while the lower plots highlight the estimated index-specific alphas in explaining/predicting HFR Global Macro index (form left to right).

Figure 2.13: These sub-figures display the evolution of the explained/predicted hedge fund index returns under the more restrictive fixed windows experiment. The first-column entries are explained series and the second-column entries are predicted ones. From up to down we have Barclay CTAs, Equity Hedge, Merger Arbitrage and Global Macro indexes, respectively.

Exhibit 2

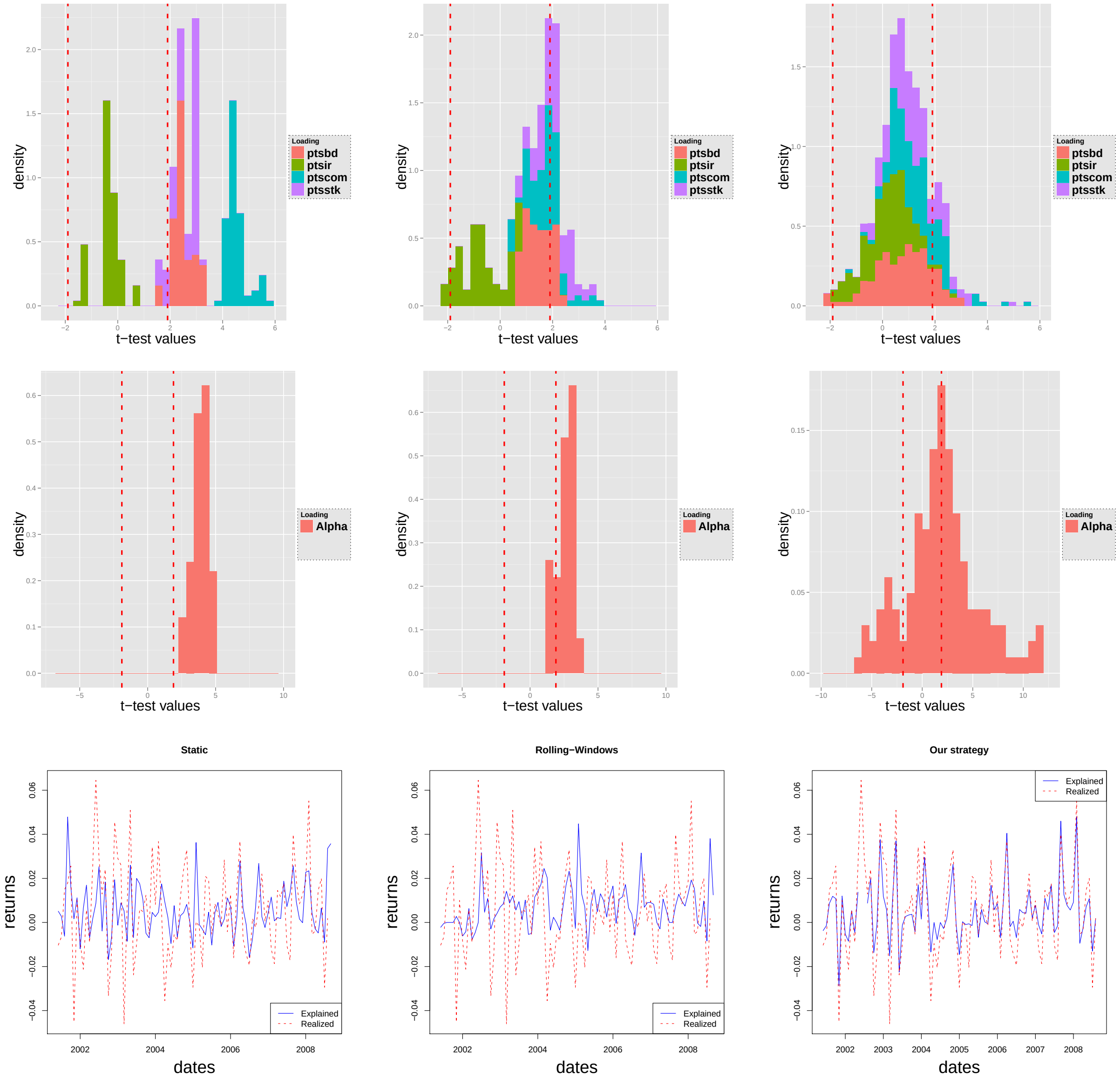

Figure 2.1 – *Explained Barclay CTAs Index: From left to right, 'static', rolling-windows, and* performance-adaptive *replication strategies. The factor model specification is defined by the Fung-Hsieh look-back straddle factors: ptsbd (bond factor), ptsir (short term interest rate factor), ptscom (commodity factor), and ptsstk (stock factor).* Alpha *stands for the estimated values of* β_0. *Refer to Figure Captions section for additional details.*

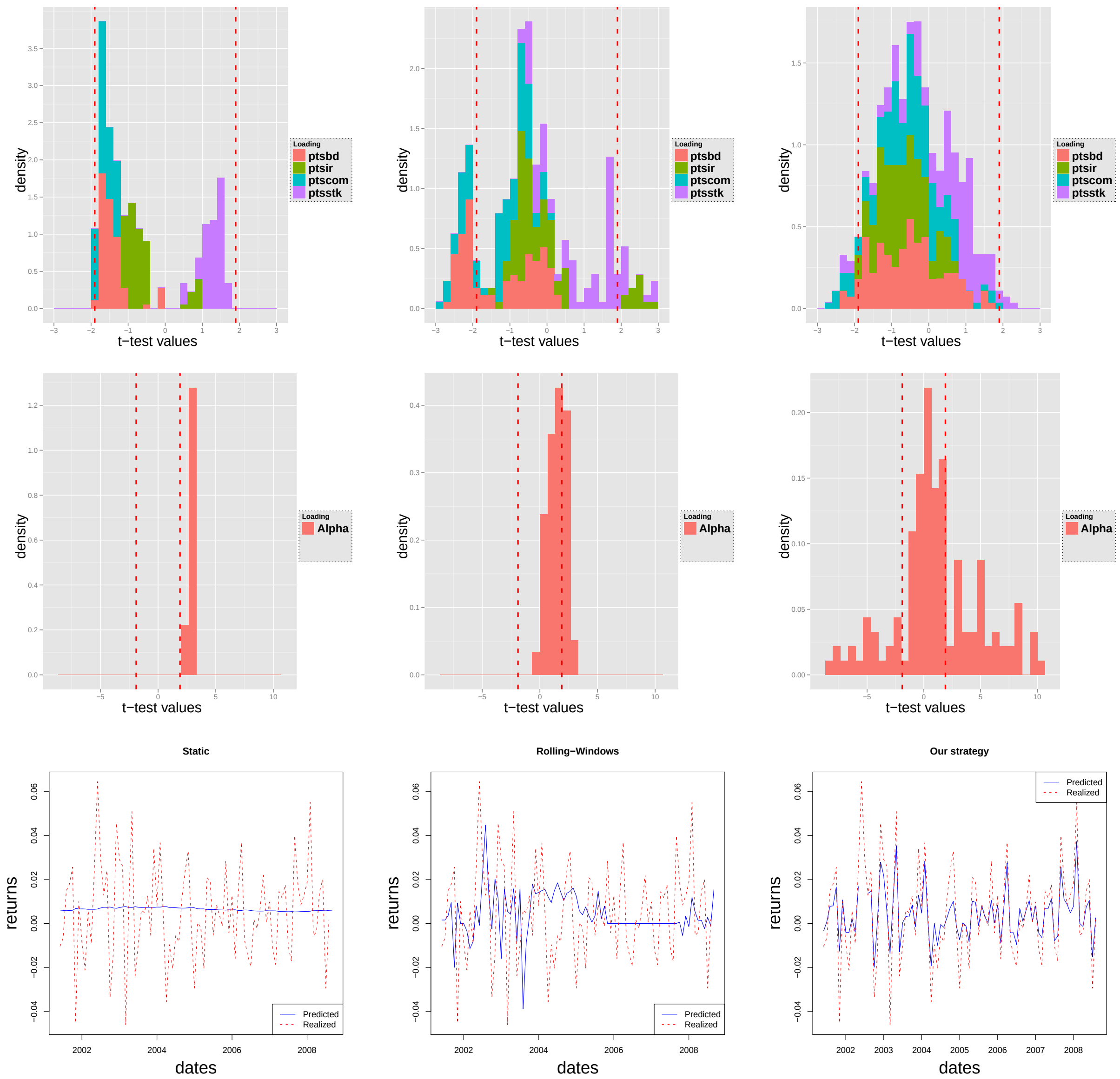

Figure 2.2 – *Predicted Barclay CTAs Index: From left to right, 'static', rolling-windows, our strategy. The factor model specification is defined by the Fung-Hsieh look-back straddle factors: ptsbd (bond factor), ptsir (short term interest rate factor), ptscom (commodity factor), and ptsstk (stock factor).* Alpha *stands for the estimated values of* β_0. *Refer to Figure Captions section for additional details.*

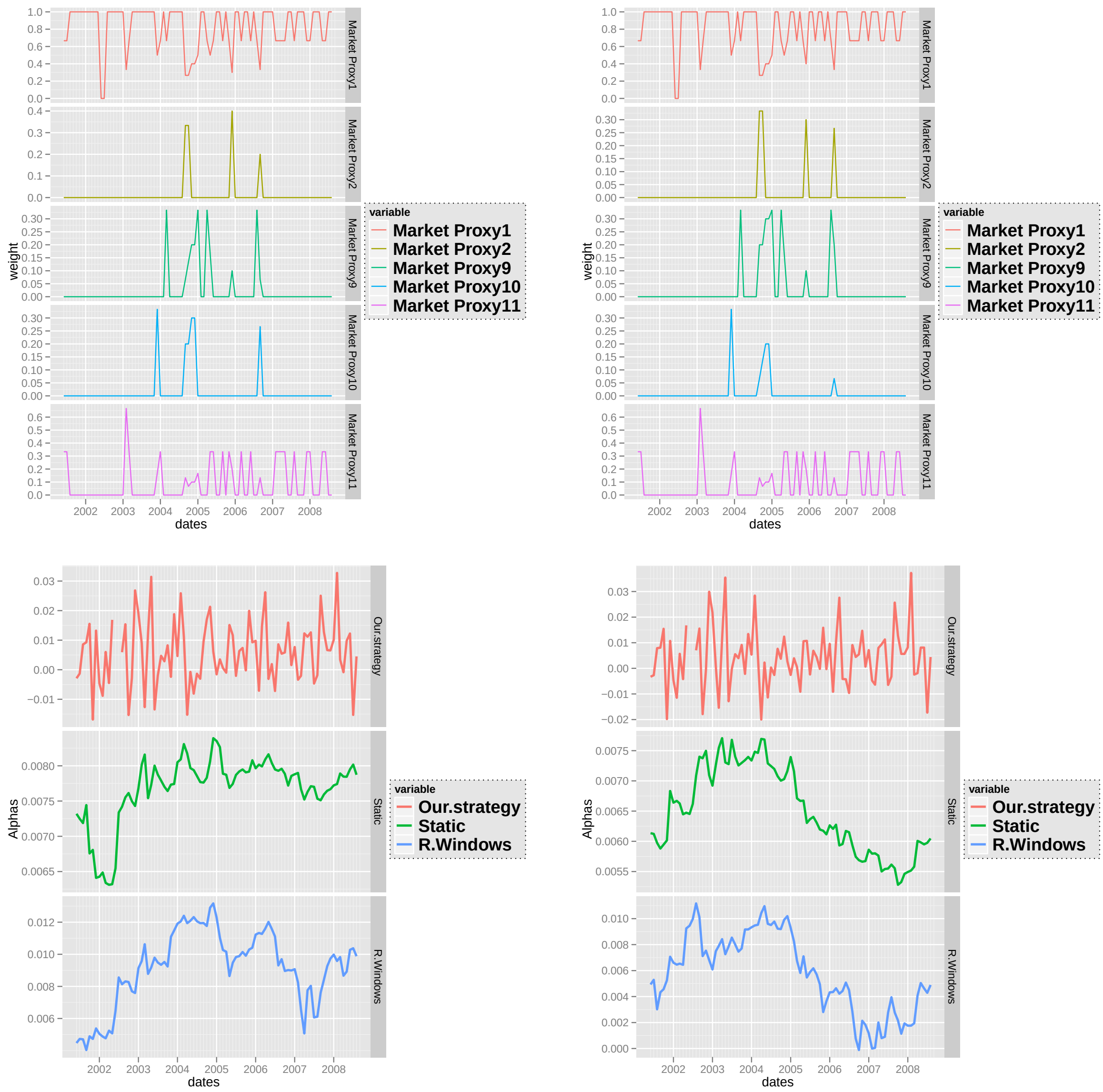

Figure 2.3 – *Barclay CTAs Index: Elementary strategy weights (first-row entries) and* Alphas, *the estimated values of* β_0 *(second-row entries). The order from left to right corresponds to explaining , and predicting strategies respectively. Refer to Figure Captions section for additional details.*

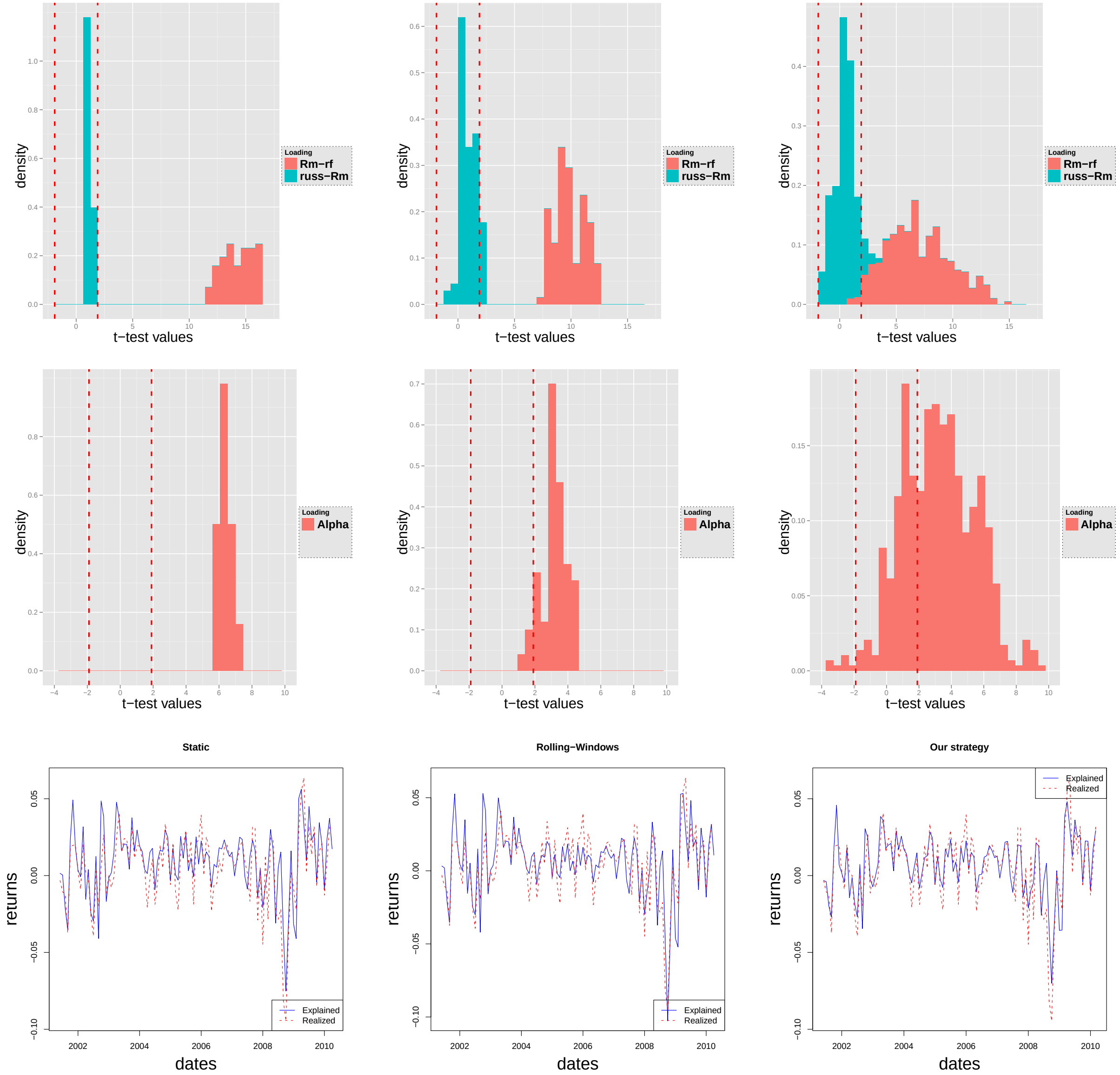

Figure 2.4 – *Explained Equity Hedge Index: From left to right, 'static', rolling-windows, our strategy. The factor model specification is defined by two factors: Rm-rf (excess return on the market, is S&P 500 index return minus the one-month Treasury bill rate) and russ-Rm (spread between the Russell 2000 index return and the S&P 500 index return).* Alpha *stands for the estimated values of* β_0*. Refer to Figure Captions section for additional details.*

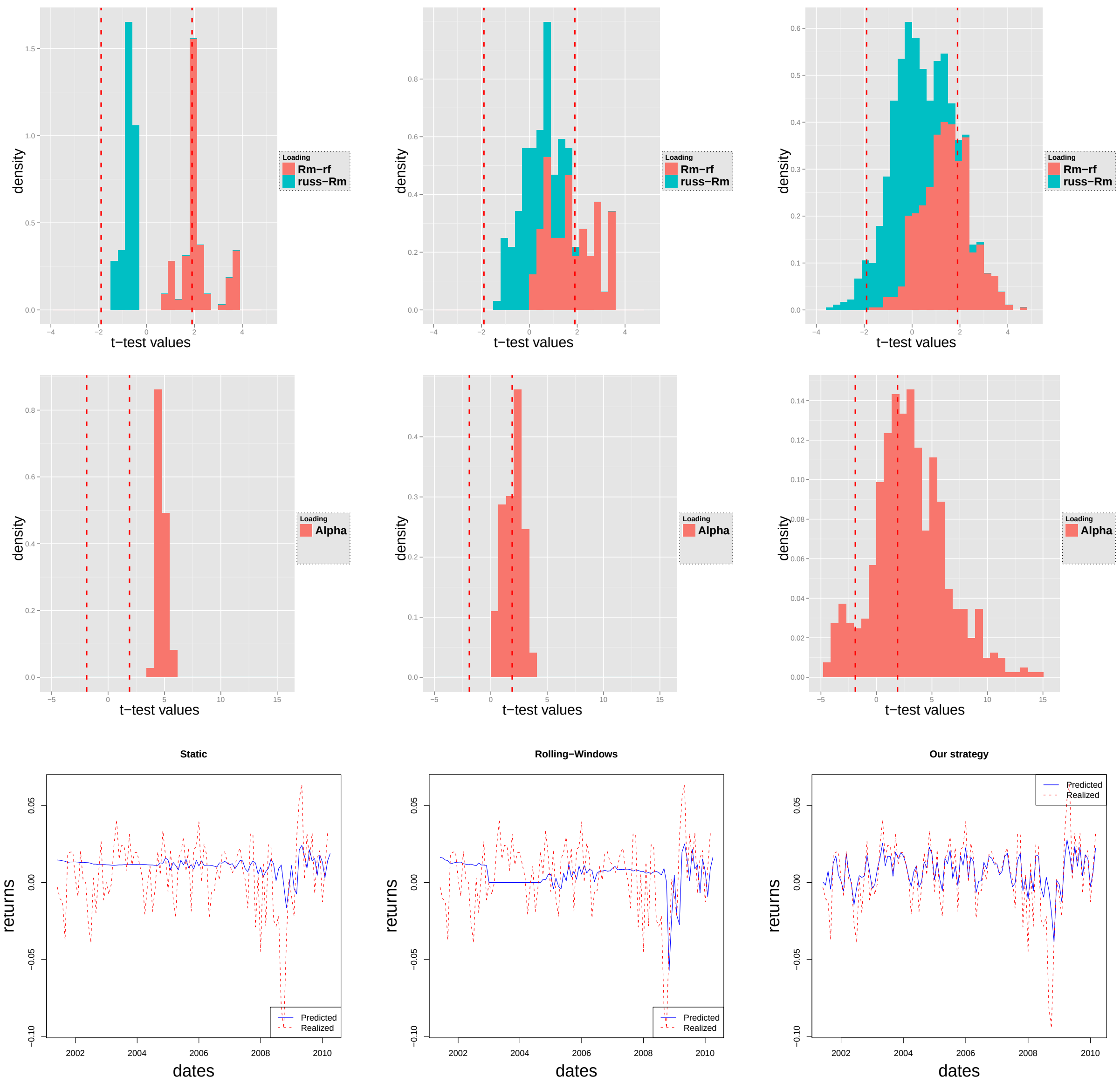

Figure 2.5 – *Predicted Equity Hedge Index: From left to right, 'static', rolling-windows, our strategy. The factor model specification is defined by two factors: Rm-rf (excess return on the market, is S&P 500 index return minus the one-month Treasury bill rate) and russ-Rm (spread between the Russell 2000 index return and the S&P 500 index return).* Alpha *stands for the estimated values of* β_0. *Refer to Figure Captions section for additional details.*

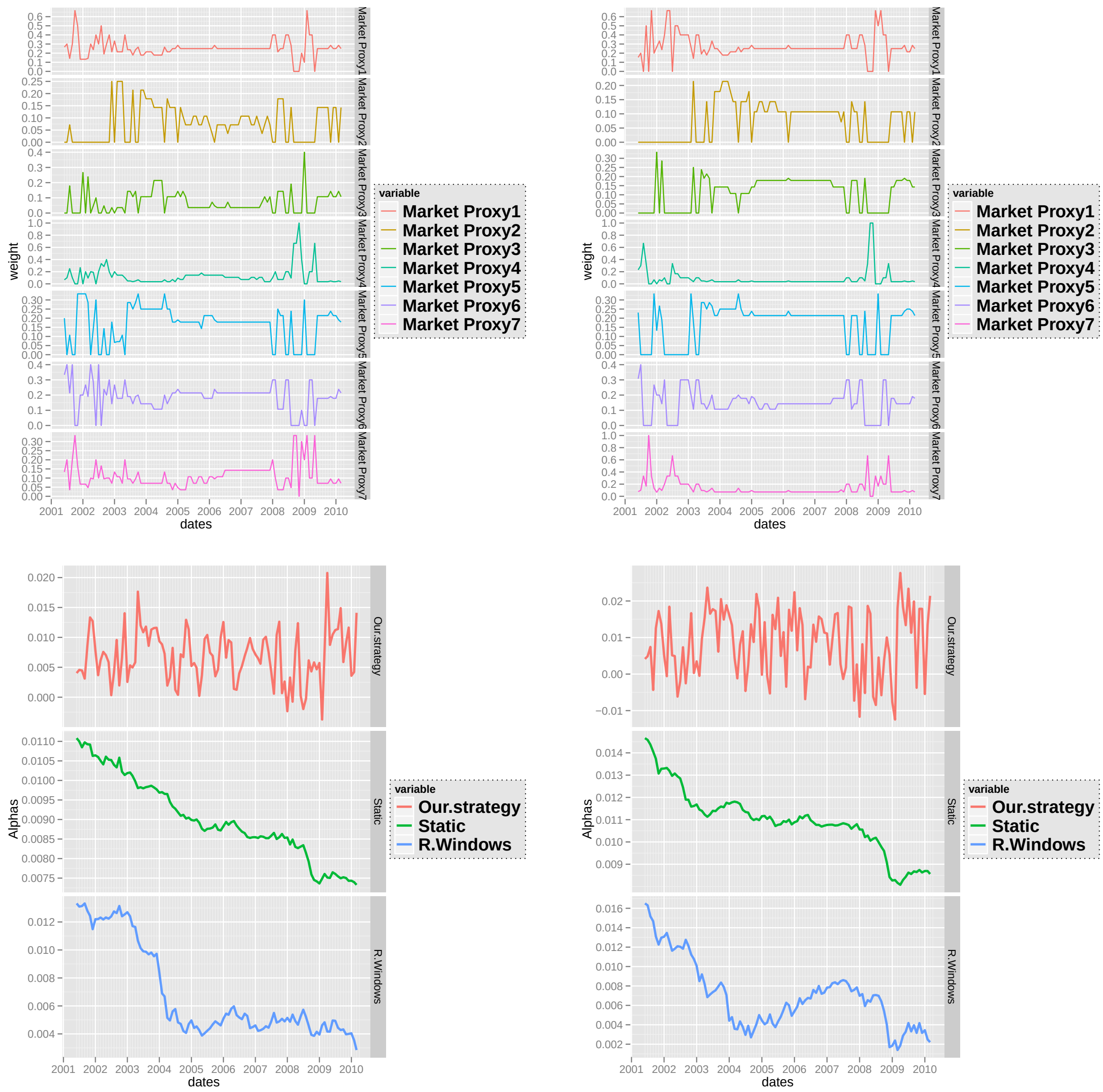

Figure 2.6 – *Equity Hedge Index: Elementary strategy weights (first-row entries) and* Alphas*, the estimated values of* β_0 *(second-row entries). The order from left to right corresponds to explaining , and predicting strategies respectively. Refer to Figure Captions section for additional details.*

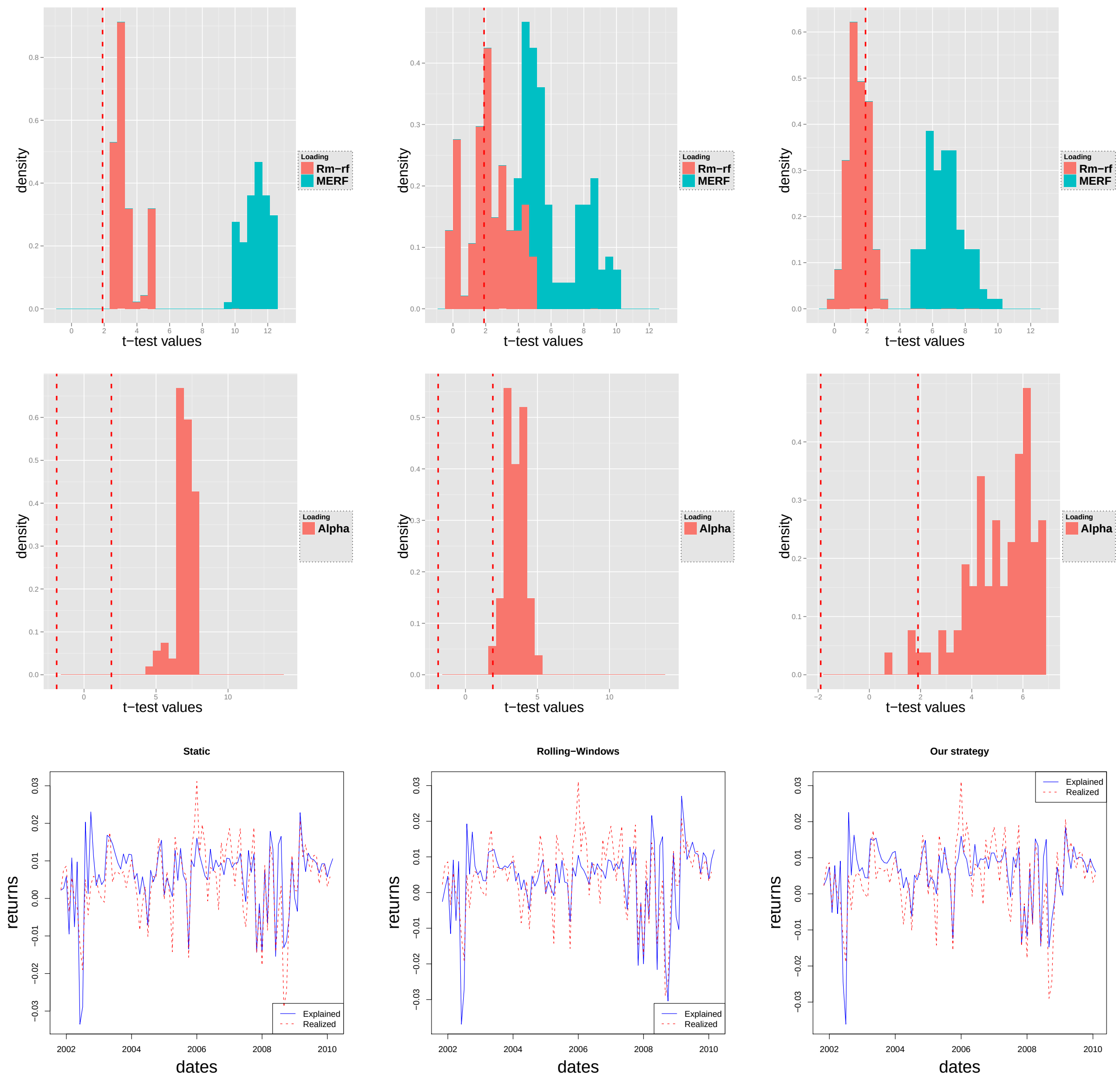

Figure 2.7 – *Explained Merger Arbitrage Index: From left to right, 'static', rolling-windows, our strategy. The factor model specification is defined by two factors: Rm-rf (excess return on the market, is S&P 500 index return minus the one-month Treasury bill rate) and MERF (return of the Merger Fund).* Alpha *stands for the estimated values of β_0. Refer to Figure Captions section for additional details.*

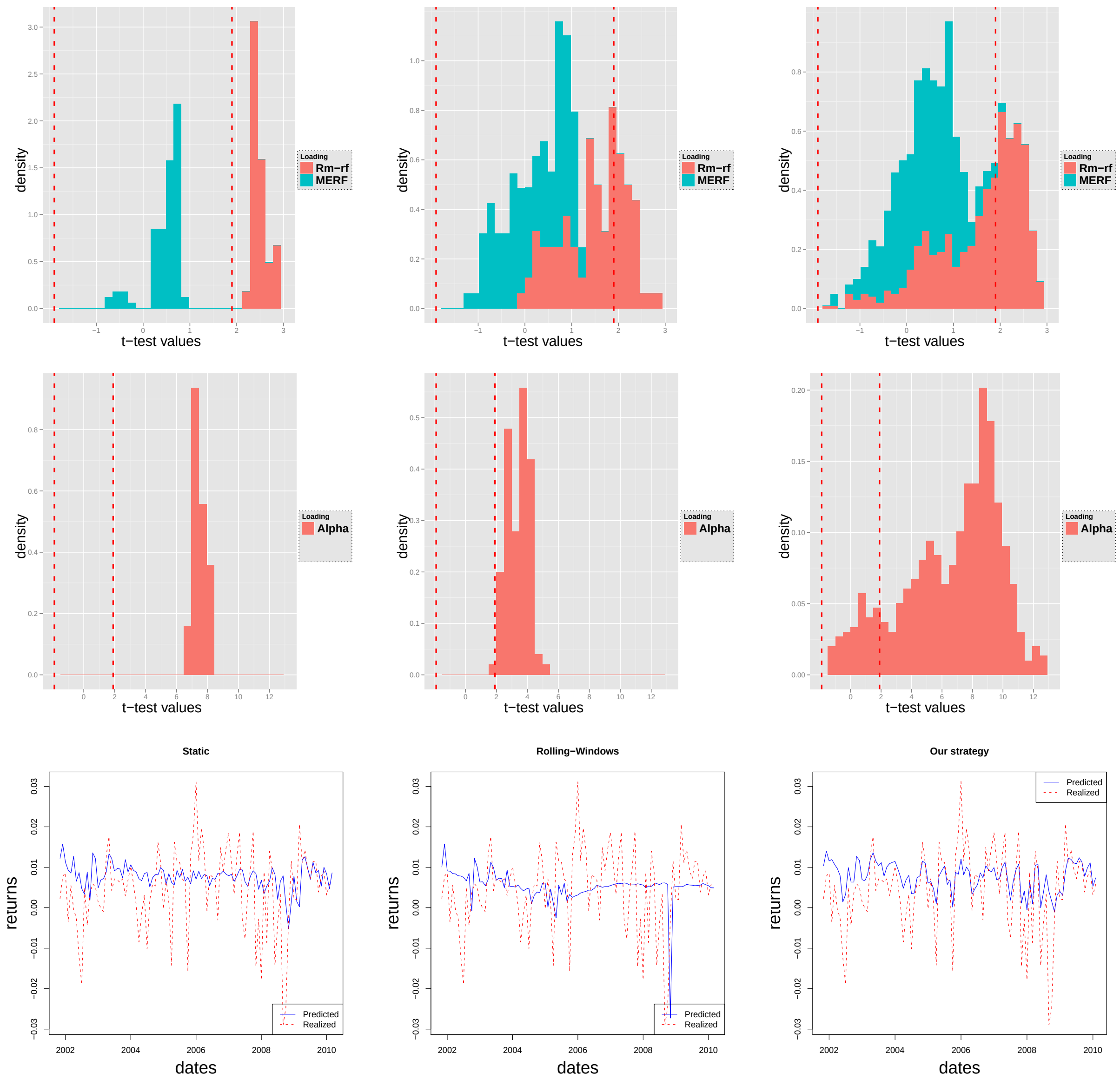

Figure 2.8 – *Predicted Merger Arbitrage Index: From left to right, 'static', rolling-windows, our strategy. The factor model specification is defined by two factors: Rm-rf (excess return on the market, is S&P 500 index return minus the one-month Treasury bill rate) and MERF (return of the Merger Fund).* Alpha *stands for the estimated values of* β_0*. Refer to Figure Captions section for additional details.*

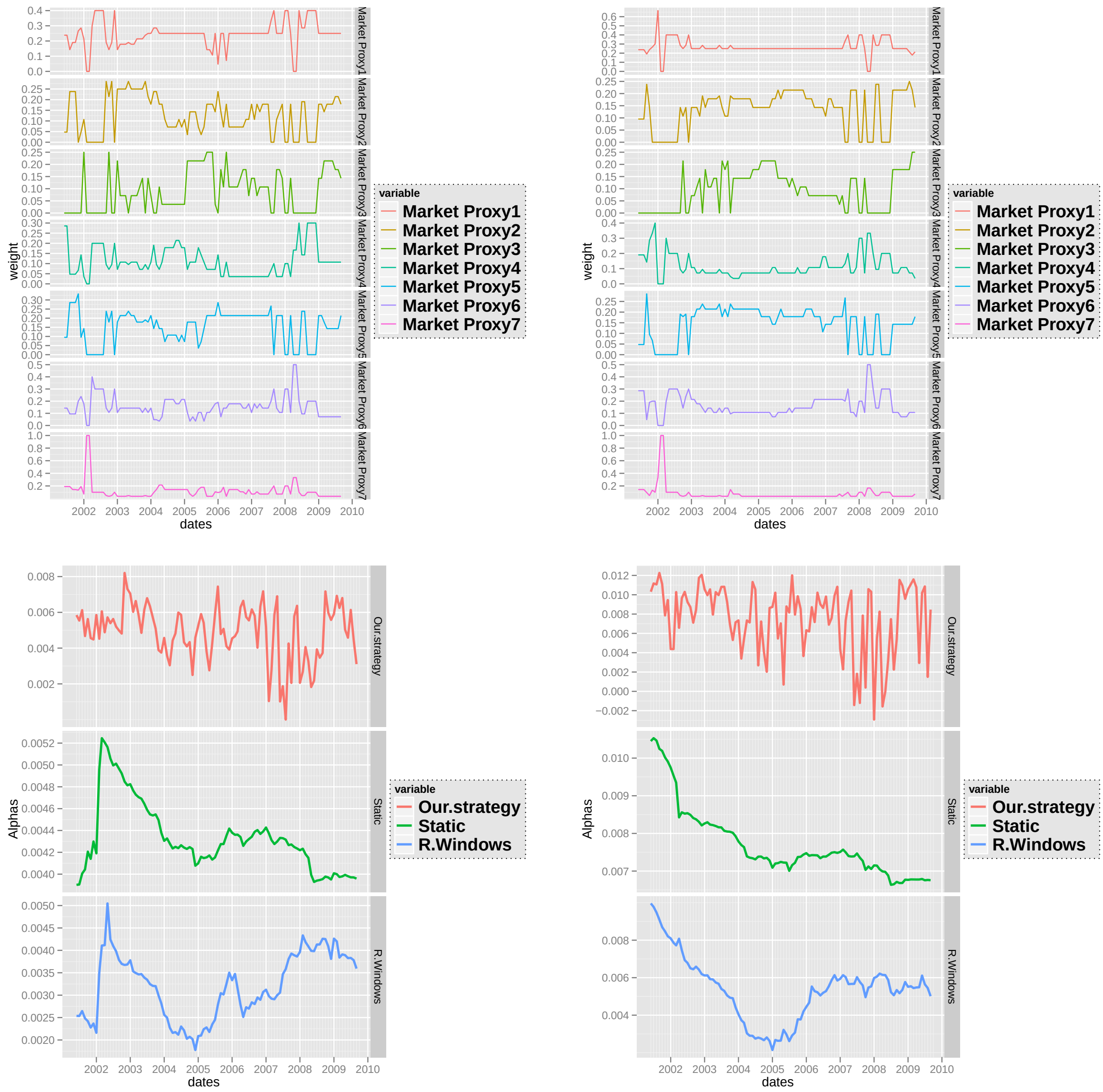

Figure 2.9 – *Merger Arbitrage Index: Elementary strategy weights (first-row entries) and* Alphas, *the estimated values of* β_0 *(second-row entries). The order from left to right corresponds to explaining , and predicting strategies respectively. Refer to Figure Captions section for additional details.*

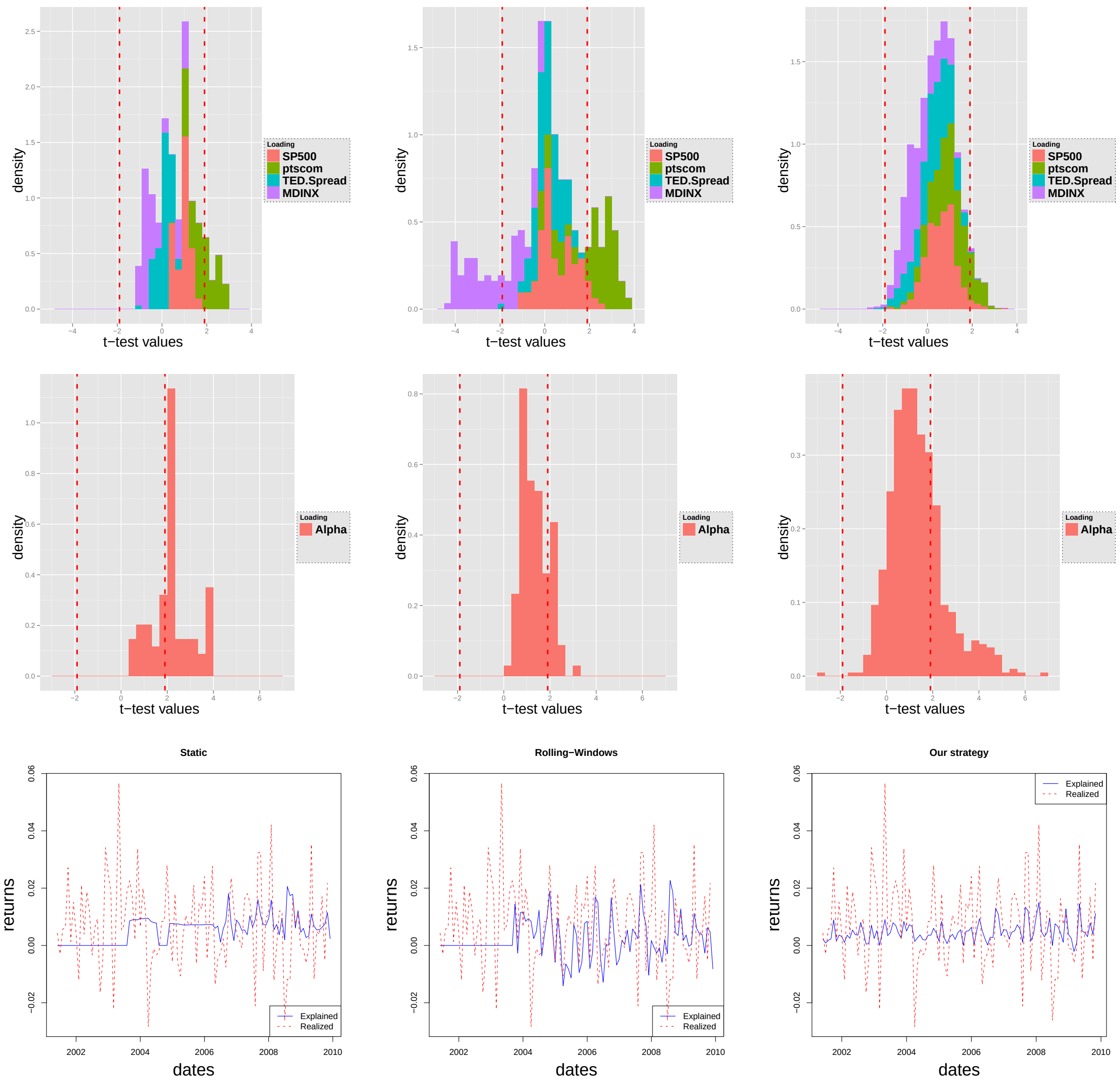

Figure 2.10 – *Explained Global Macro Index: From left to right, 'static', rolling-windows, our strategy. The factor model specification is defined by the four factors: S&P 500 index return (SP500), Fung-Hsieh look-back straddle commodity factor (ptscom), TED spread (TED.Spread) and the Major currency dollar index (MDINX).* Alpha *stands for the estimated values of* β_0*. Refer to Figure Captions section for additional details.*

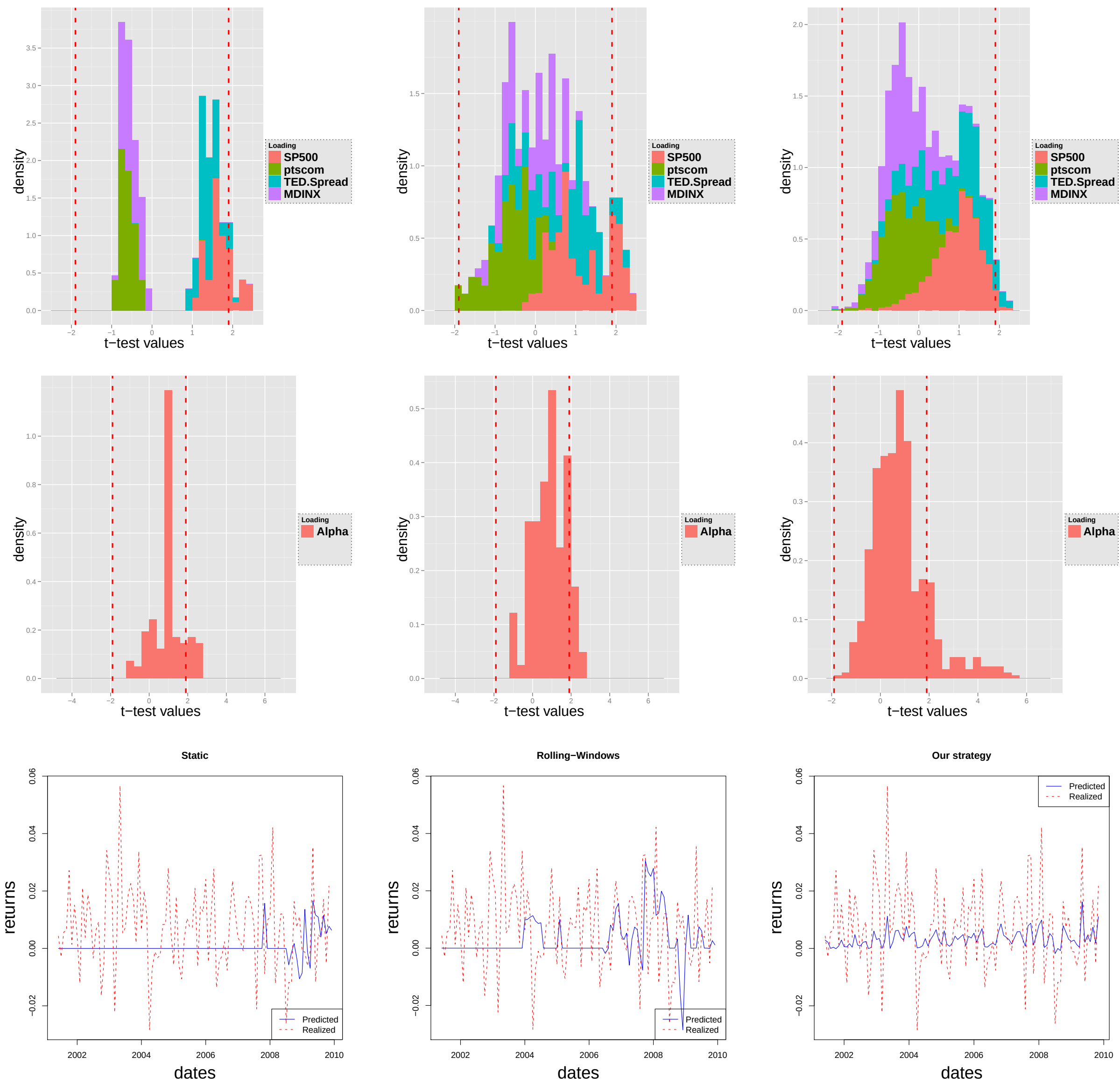

Figure 2.11 – *Predicted Global Macro Index: From left to right, 'static', rolling-windows, our strategy. The factor model specification is defined by the four factors: S&P 500 index return (SP500), Fung-Hsieh look-back straddle commodity factor (ptscom), TED spread (TED.Spread) and the Major currency dollar index (MDINX).* Alpha *stands for the estimated values of β_0. Refer to Figure Captions section for additional details.*

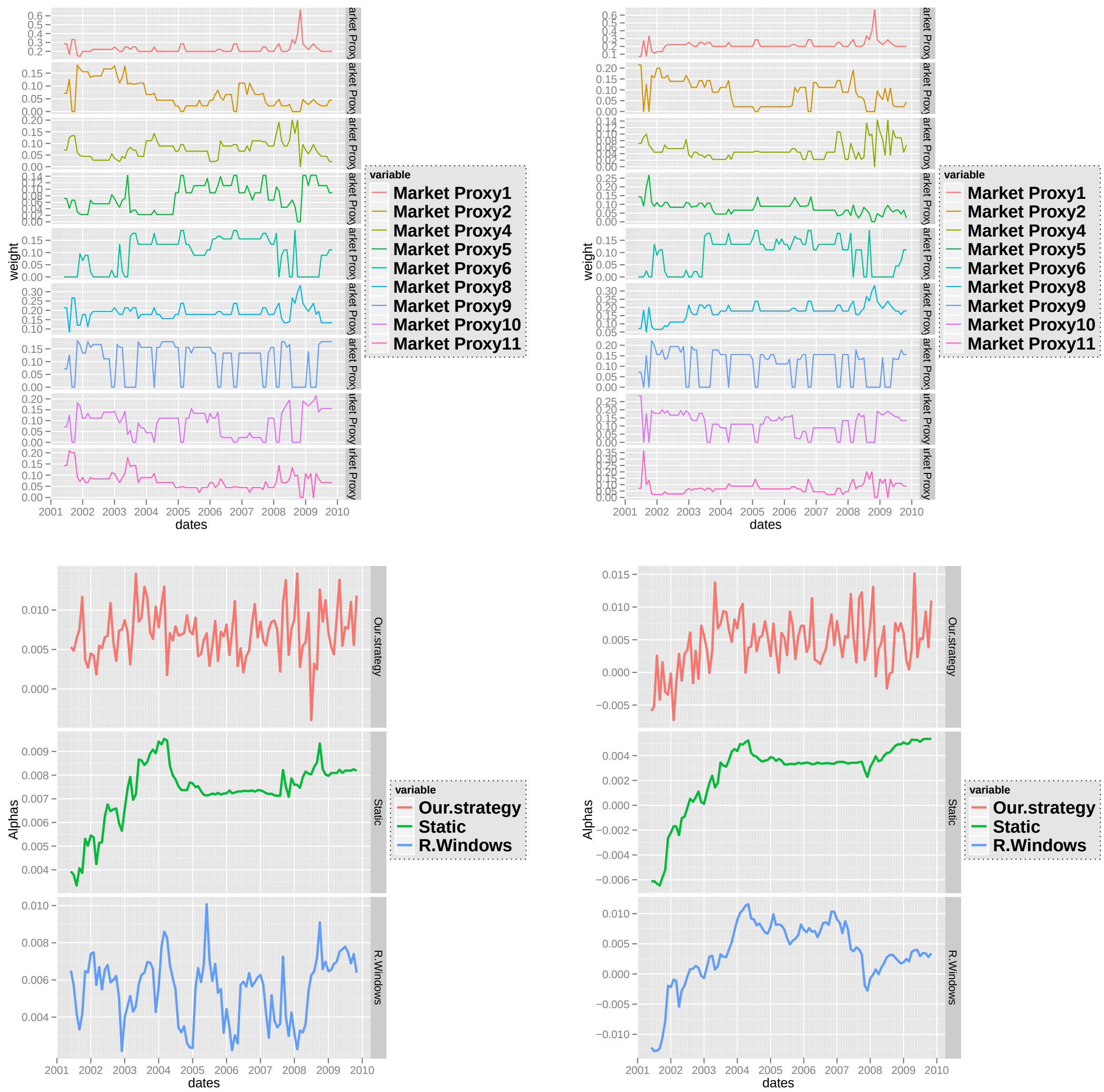

Figure 2.12 – *Global Macro Index: Elementary strategy weights (first-row entries) and* Alphas*, the estimated values of β_0 (second-row entries). The order from left to right corresponds to explaining , and predicting strategies respectively. Refer to Figure Captions section for additional details.*

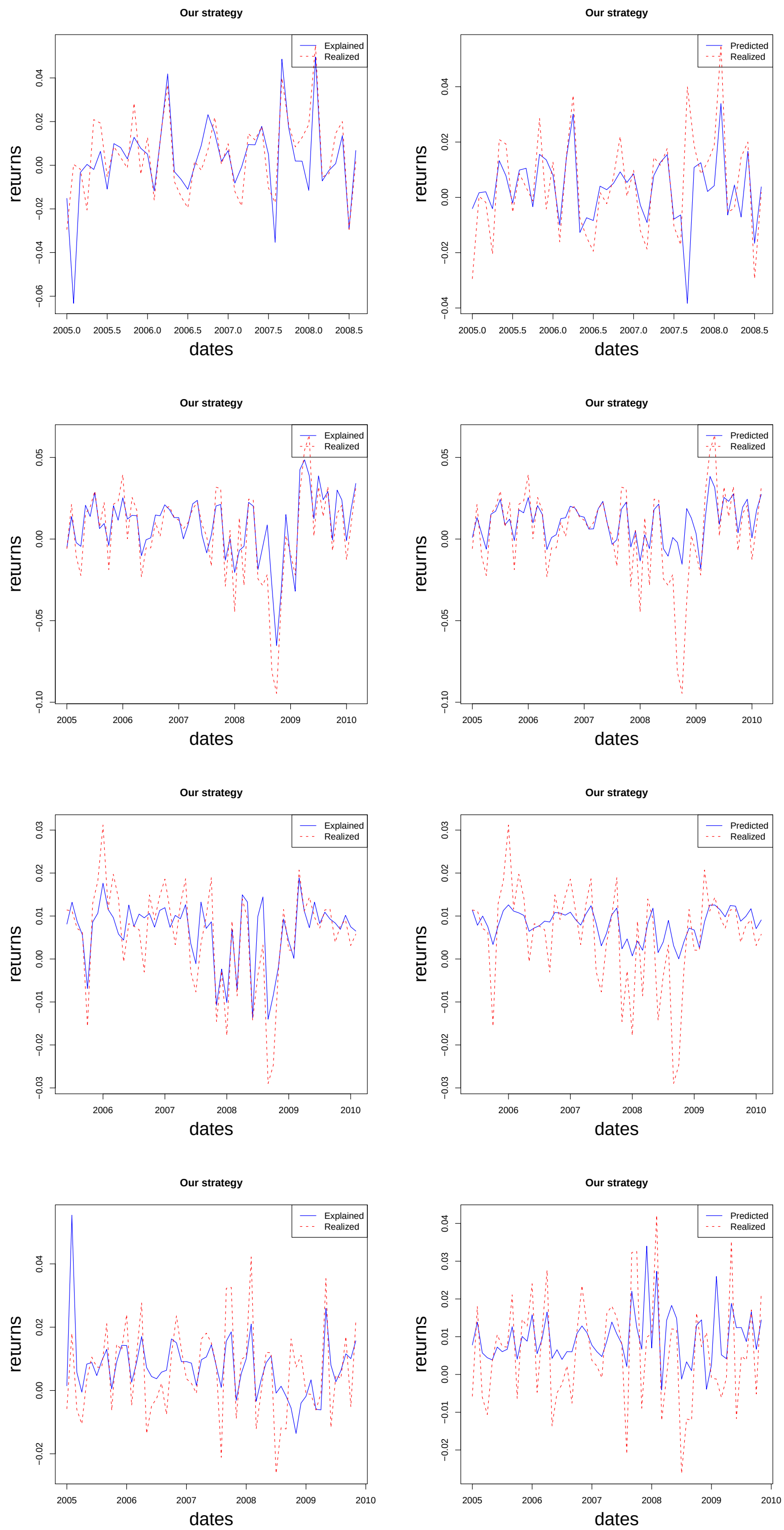

Figure 2.13 – *Fixed windows experiment: The order from left to right corresponds to explaining, and predicting strategies respectively. From top to bottom, we have Barclay CTAs, Equity Hedge, Merger Arbitrage and Global Macro indexes. Refer to Figure Captions section for additional details.*

3 Performance Based Diversification or How To Create a Multistrategy hedge fund?

Abstract

This article investigated the implications of distribution free investment strategies on constructing portfolio of individual hedge funds. The author proposes a dynamic, performance-adaptive asset allocation model that allows to optimally diversify across multiple hedge funds styles. The approach to be followed is related to the adaptive allocation of resources between elementary concurrent from the perspective of the theory of sequential investment strategies. The methodological frame gives up the common global stationary hypothesis and approximates locally, a nonstationary paradigm. The approach is then evaluated in a multivariate basis, by examining the performances of several time evolving portfolio strategies in a sample of 16 funds. Empirical experiments are conducted across 5 different hedge fund categories as classified by the Hedge Funds Research and Barclay CTAs databases. We find that the dynamic performance-adaptive allocation strategies amongst hedge funds yield superior annualized average performances, compared to various alternative benchmarks. These findings are robust to different hedge fund restriction provisions such as, lockup and redemption periods.[1]

Keywords: Hedge fund, Portfolio allocation, Kernel estimation, Nonparametric, Nonstationary

[1]This chapter is a reprint of: Tafin Djoko (2013a) Performance Based Diversification or How To Create a Multistrategy hedge fund? Accepted for publication in *Journal of Applied Finance and Banking*.

3.1 Introduction

This article reports a methodology for constructing portfolios of hedge funds that drops the common assumption of global stationarity. Nonstationary modeling has a long history in empirical financial analysis. The reader may refer to Hsu, Miller, and Wichern (1974), Starica and Granger (2005), and Perron and Qu (2010) for a more recent investigation. Working under nonstationarity has always been motivated by the belief that financial market microstructure is continuously changing, with direct direct consequences in asset price dynamics.

To date, portfolio managers tend to use mean-variance (MV) optimization techniques to construct optimal risk-return balanced asset allocation. However, to perform a MV allocation of resources, the manager has to provide the next-period variance-covariance matrix. One easiest way is to use the historical covariance matrix. A common extension of MV allocation is the construction of volatility timing strategy using the dynamic conditional correlations (DCC) model of Engle (2002). In this regard, Switzer and Omelchak (2011) focus on the case where an investor or funds of funds manager is concerned with the volatility of a portfolio of hedge fund indexes. Their work successfully extends the static MV asset allocation framework to allow for time varying volatility of returns. However, like most of the multivariate methodological frame for financial returns, the former model assumes that the volatility is homogeneous and stationary. Additionally, the standard MV portfolio construction involving hedge funds has been subject to criticisms in the empirical financial literature. The most common criticism is that MV analysis is suitable, exclusively for normally distributed returns, which is far from being a hedge fund characteristic. Empirical analyzes of hedge fund return series tend to exhibit asymmetry and excess kurtosis, which imply theoretical issues to implement a proper MV portfolio construction.

Assuming a MV criterion directly implies that the data generating process of returns is *Gaussian*, or at least that moments over the first and the second are trivial for asset allocation. Early empirical study reveals that, the MV criterion fails to approximate the expected utility in non-normality cases. In this vein, Harvey and Siddique (2000) express the view that systematic skewness (co-skewness with the market return) yields economic value to investors. In a recent manuscript, Jondeau and Rockinger (2011) examine the additional value of taking the complexity of time varying higher moments into the underlying distribution of returns. Jondeau and Rockinger (2011) estimate the model in a Bayesian configuration and show that higher moment investment strategies (up to order four) significantly outperform the MV strategy.

MV portfolio modeling and extensions of it are dominated by stationary long-memory, conditional framework. In this paper, I advocate an approximate nonstationary paradigm to construct a dynamic portfolio allocation of capital across individual hedge funds. Due to the unregulated status of their business model, hedge fund managers enjoy enormous flexibility in pursuing investment returns. The industry investment philosophy is time varying and evolves according to the structure of the market. Therefore, investing in hedge funds through a stationary-constructed portfolio of individual hedge fund strategies can bring to seriously underestimating the risk associated to the portfolio, with potential drawback in terms of market timing ability and risk management. The changing nature of the data generating process requires a consequent adjustment of the approximating stationary approach. The goal of the article is to investigate the economic benefits of locally stationary versus global stationary (parametric) models.

The building block of our locally stationary model is to identifying periods of time where market conditions are 'similar' to the current trading period. This specification will provide intervals of homogeneity in which an optimal allocation of resources can be decided. As a proxy of market conditions, the author intends to capitalize on the vast literature on factor-based hedge fund replication. I use an approximate version of Hasanhodzic and Lo (2007)'s six factors model, to identify the variables that characterize the market at each trading period. The first step of the method consists of identifying the instances where the state of the market, i.e. the risk factor specification, was 'similar' to the current conditions. Then, only the hedge fund return information from these instances are the input to construct the current optimal allocation strategy.

The definition of the 'distance' between past and current market conditions is, of course, a crucial aspect in applying the methodology. Since we do not have *a priori* knowledge of the variables (market factors) that are instrumental in defining the state of the market, a refinement of the methodology puts in competition a number of elementary strategies. Each one of them is characterized by a proxy for the market conditions, i.e. a given set of relevant variables (and a depth of the history of these variables) whose values are used to define 'similar' or 'close' market conditions. For each one of the market condition proxy, a portfolio is constructed, based on the time instances identified by the proxy as 'similar' to those of the allocation moment. The final investment decision follows then an adaptive, time varying allocation of resources between different proxy portfolios based on their recent past performance. See Section 3.2 for further details.

Once homogeneous market proxies have been identified, I implemented two types of locally stationary models: A fully nonparametric and an asymptotic MV. The later is an extension of Györfi, Lugosi, and Udina (2006) kernel-based, sequential investment strategy in which the investor maximizes his wealth in the long run without any assumption of the underlying probabilistic structure of the data. The portfolio allocation is time evolving and distribution free. The former is a Markowitz-type portfolio in which a risk-adverse investor carries each time a mean-variance optimization, however without knowing the statistical characterizations of the underlying process. See György and Vajda (2007) additional information.

Research on investment strategies with minimal assumptions on the underlying distribution driving the available assets, has been addressed by various authors such as Algoet and Cover (1988), Algoet (1992). Györfi and Shäfer (2003) has shown the existence of universal portfolios, which can achieved asymptotically, the maximum rate of growth of capital without any knowledge of the data generating process. Given the complex structure of hedge fund investment vehicles, I work with minimal assumptions on funds return's distribution. In fact, the only assumption is that monthly fund's gain factor ((1+$r_{j,t}$), where $r_{j,t}$ is the return of fund j at month t) forms a locally stationary and ergodic process. In theory, this assumption allows to the allocation strategy to reach in the long, the maximum rate of growth of capital that can be obtained knowing completely the distribution of the underlying process.

As the paper addresses the issue of dynamic portfolio allocation in a nonstationary paradigm, a comparison with equally weighted constantly rebalanced and dynamic mean-variance portfolios is implemented. The parametric, stationary dynamic MV optimization requires a statistical model to predict subsequent volatility and the conditional expected returns. The dynamic conditional volatility forecast is carried out using the Weide (2002) generalized orthogonal GARCH(1,1) model.

The main contribution of this article is directly related to the nonparametric structure of the analysis. Indeed, I move away from the parametric and conditional perspective, to a nonstationary and distribution free paradigm. This work is the first, at our knowledge, that clearly account for time variability and nonparametric design in constructing portfolios of individual hedge funds. Empirical studies are implemented in various hedge fund data sets, including the Chicago-based HFR and Barclay CTAs databases. Our findings suggest that unconditional models are able to uncover and exploit hidden past dependence structures between funds's gain factors. The model allows to incorporate time variation and is constantly out-of-sample. A clear advantage of this model is that it is completely data driven and remains computationally tractable even when several funds are incorporated.

The remainder of the paper is structured as follows. The methodology behinds the portfolio construction strategy is documented in Section 3.2. Section 3.3 describes the data used in the empirical exercise and formulate our approach for selecting a restricted basket of funds to construct portfolios of hedge funds. In Section 3.4 the author discusses some implementation issues and a detailed characterization of the proposed investment process. Empirical results are outlined in Section 3.5. The main characteristics of the portfolios constructed under several alternative optimizers are outlined. Finally, I provide some robustness assessment of core findings. A tentative conclusion is outlined in Section 3.6. Additional information are formulated in Appendices 3.7.1 and 3.7.2.

3.2 Methodology

This section describes the nonparametric portfolio allocation methodology to construct dynamic portfolios of hedge funds. The investment process consists of a performance-adaptive allocation of resources between a finite set of elementary strategies, from the perspective of the theory related to sequential investment strategies. The analysis is distribution free and the only mathematical assumption used is that the monthly fund's relatives Net Asset Value (NAV) form a locally stationary and ergodic process. Under this assumption, the asymptotic rate of growth of the constructed portfolio has a well-defined maximum, which can be achieved without knowing the data generating process. As in Györfi, Lugosi, and Udina (2006) and György and Vajda (2007), I approximate a kernel-based allocation mechanism in order to obtain an optimal and robust growth rate of the capital, in a finite investment horizon.

3.2.1 Background

The goal here is to construct a multistrategy portfolio of hedge funds, implementing a dynamic allocation algorithm that optimally diversifies and maximizes in the long run the investor wealth over multiple hedge funds styles. The portfolio predictions relied only on past market information.

To crystallize the context of the analysis, consider a hedge fund universe of m managers, $i = 1, \ldots, m$, spanning an heterogeneous range of investment strategies. The vector $\mathbf{X} = (x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}_+^m$ is the vector of m nonnegative numbers representing fund's relatives NAV at a given trading period, where the jth component $x^{(j)} \geq 0$ of $\mathbf{X}$ represents the ratio between the current and previous NAV of fund's manager j $\left(x^{(j)} = NAV_t^{(j)} / NAV_{t-1}^{(j)} = 1 + r_{j,t}\right)$. Basically, $x^{(j)}$ is the coefficient by which capital allocated in fund j grows during the trading period. We consider an investment setting where an investor or fund of hedge funds (FOF) manager is allowed to allocate his capital at the beginning of each trading period[2] according to a portfolio vector $\boldsymbol{\alpha} = (\alpha^{(1)}, \ldots, \alpha^{(m)})$. The j$th$ component $\alpha^{(j)}$ of $\boldsymbol{\alpha}$ denotes the weight of the investor's capital allocated in fund j. Throughout the article, I assume the investment strategy is self-financing (no borrowing) and the proceed is fully reinvested. This

[2]We will further relax this assumption to account for hedge fund redemption restrictions such as lockup period and redemption period.

specification implies a long only portfolio selection strategy with $\alpha^{(j)} \geq 0$ and $\sum_{j=1}^{m} \alpha^{(j)} = 1$.

Different hypothesis may be formulated on the path dependent process behind a time varying portfolio selection. For a more general algorithm, the $\boldsymbol{\alpha}$-portfolio vector depends on market history, characterized the sequences of funds's realized gain factors and factor model specification. The state of the market is approximate by the Hasanhodzic and Lo (2007)'s six factors model representation. Let suppose that there are m risky, open for investment funds and their market dynamic is well described by a sequence of market vector $\mathbf{X}_1, \mathbf{X}_2, \ldots \in \mathbb{R}^m_+$, where the jth component $x_t^{(j)}$ of $\mathbf{X}_t$ indicates the amount obtained, net of fee after investing an unit of capital in the jth fund at the t_{th} trading period. The factor model is described by a sequence of vector $\mathbf{Y}_1, \mathbf{Y}_2, \ldots \in \mathbb{R}^6$. To ease the notation, for $i \leq t$, $\mathbf{X}_{i,t}$ denotes the array of market vectors $(\mathbf{X}_i, \ldots, \mathbf{X}_t)$.

Therefore, at each trading period $t = 1, 2, \ldots$, we want to perform an optimal allocation of resources based on past market information $\mathcal{I}_{1,t-1} = (\mathbf{X}_{1,t-1}, \mathbf{Y}_{1,t-1})$. Then, $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}(\mathcal{I}_{1,t-1})$ denotes the portfolio vector chosen by the investor in the t_{th} trading period, after observing the past behavior of the market.

Because we are primarily interested in maximizing the long tern rate of growth of investor's capital, I describe the wealth accumulation dynamic. Starting with an initial capital C_0, the end-of-period wealth of investment strategy $\mathbf{I}$, after t trading periods is

$$C_t = C_0 \prod_{i=1}^{t} \langle \boldsymbol{\alpha}_i, \mathbf{X}_i \rangle = C_0 \exp^{\sum_{i=1}^{t} \log \langle \boldsymbol{\alpha}_i, \mathbf{X}_i \rangle} = C_0 \exp^{tW_t(\mathbf{I})},$$

where $W_t(\mathbf{I})$ indicates the average growth rate

$$W_t(\mathbf{I}) = \frac{1}{t} \sum_{i=1}^{t} \log \langle \boldsymbol{\alpha}_i, \mathbf{X}_i \rangle$$

Since it is rationally to assume that the objective of any investor is to maximize his wealth, it is straightforward that the maximization of $C_t = C_t(\mathbf{I})$ is equivalent to the maximization of $W_t(\mathbf{I})$.

The present work moves away from the MV portfolio framework and concentrates on a hypothetical investor concerned to maximize his end of period capital in a dynamic optimization setting. Following Györfi, Lugosi, and Udina (2006), I adopt a fully nonparametric setting. We do not assume any parametric structure on the distribution of the sequence of hedge fund's relatives NAV. Under the mild hypothesis, Algoet and Cover (1988) show that the best possible choice of the sequential investment problem is the so-called *log-optimum portfolio* $\mathbf{I}^\star = \{\alpha_t^\star(\cdot)\}$. That is, in trading period t the optimal allocation $\alpha_t^\star(\cdot)$ is such that,

$$\mathbb{E}\{\log\langle \boldsymbol{\alpha}_t^\star, \mathbf{X}_t\rangle \mid \mathcal{I}_{1,t-1}\} = \max_{\boldsymbol{\alpha}(\cdot)} \mathbb{E}\{\log\langle \boldsymbol{\alpha}_t^\star, \mathbf{X}_t\rangle \mid \mathcal{I}_{1,t-1}\}$$

No investment rule can have a faster rate of growth than the *log-optimal portfolio*, where a full knowledge of the distribution of the process is required. However, Algoet (1992) shows that there exist strategies reaching a identical rate of growth without knowing the distribution of the underlying stochastic process, the so-called *universal portfolio*. Following Algoet's scheme, Györfi, Lugosi, and Udina (2006) provided a more practical and general version of the *universal portfolio* concept. For the scope of our study, I implement the so-called kernel-based portfolio optimization.

3.2.2 Time varying kernel-based portfolio selection

Given our interest in constructing portfolios of hedge funds relying uniquely on predictions based on past market conditions, I build a model that enables us to collect the historical sequences of data that are informative to current market states. This section introduces the kernel-based sequential investment strategy, which exploits market condition proxies to estimate the optimal portfolio weights of funds. We describe and implement the uniform kernel version. The approach is closely related to Györfi, Lugosi, and Udina (2006), who form a nonparametric *log-optimum portfolio* strategies.

Basically, the idea behind the *kernel-based* ($\mathbf{I}^K$) is quite simple and intuitive. Under a well defined set of parameters, the strategy develops a flexible algorithm to individuate different sequences of data to estimate the optimal portfolio weights by maximizing the investor wealth. The approach is completely data driven and consists firstly, to identify those months in the past where the state of the market as described by a set of factors, was 'similar' to the current conditions. Only the the hedge fund return information from the time instances that follow are then used to determine the optimal allocation of capital amongst funds.

Since we do not have *a priori* knowledge of the variables (risk factors) that are instrumental in defining the state of the market, the methodology puts in competition a number of elementary investment strategies. Each of these elementary investment strategies is characterized by a given set of relevant variables whose values are used to define 'similar' market conditions. Each elementary strategy hence proposes its own allocation, based on the time instances that it identified as 'close' to the current investment period. The final investment decision follows then an adaptive, time varying allocation of resources between these 'experts'. The allocation is based on their recent past performance. What follows provides a detailed description of the methodology that I implement in the sequel.

The working hypotheses motivating the *kernel-based* sequential algorithm are rather intuitive: First I allow the state of the market (characterized by a set of risk factors) to have an impact to the risk-return structure of hedge fund and second, the dynamics of these impacts should be similar during periods that are characterized by similar market conditions.

The $\mathbf{I}^K$ investment strategy is constructed as follows. First I introduce a data selection process which individuates periods characterized by similar market conditions. The market is identified by an approximate version of Hasanhodzic and Lo (2007)'s six factors model. The similarity between market conditions characterizing two time periods is measured by the Mahalanobis distance between the values of all past approximate market vectors. Let $\mathbf{Y}_{t-1}$ be the vector of market proxy at time $t-1$ and $\mathbf{Y}_{k,t-1} = (Y_{t-k}, \ldots, Y_{t-1})$ the sequence of past k values of the market condition variable. For a given d small, define[3]

$$S_t^{(k)} = \{k < i < t : \; || \mathbf{Y}_{t-k,i-1} - \mathbf{Y}_{t-k,t-1} ||_M \leq d\} \tag{3.1}$$

the set of all past time instances, following those when the market conditions as measured by the k past values of the variable $\mathbf{Y}$ were similar to the last seen vector $\mathbf{Y}_{t-k,i-1}$.

If $\mathbf{X}_i$ is the array of hedge fund gain factors that correspond to the instances $i \in S_t^{(k)}$, then for each $t > k+1$, elementary portfolios are estimated as

$$\begin{aligned} \boldsymbol{\alpha}(S_t) &= \underset{\boldsymbol{\alpha} \in \Delta_m}{\operatorname{argmax}} \prod_{\{i \in S_t^{(k)}\}} \langle \boldsymbol{\alpha} \, , \mathbf{X}_i \rangle \\ &= \underset{\boldsymbol{\alpha} \in \Delta_m}{\operatorname{argmax}} \frac{\sum_{\{i \in S_{t,(k,l)}\}} \log \langle \boldsymbol{\alpha} \, , \mathbf{X}_i \rangle}{| \, S_t^{(k)} \, |} \end{aligned}$$

if $S_t^{(k)} \neq \emptyset$ and

$$\boldsymbol{\alpha}(\mathbf{X}_{1,t-1}) = \underset{\boldsymbol{\alpha} \in \Delta_m}{\operatorname{argmax}} \prod_{\{\tau\}} \langle \boldsymbol{\alpha} \, , \mathbf{X}_{\tau,t-1} \rangle$$

otherwise. That is, I use the past τ length of data to estimate the portfolio vector. In this work, τ is fixed at 24 months and Δ_m is the simplex of all m-dimensional vectors with nonnegative components summing up to one.

The second step consists in aggregating elementary sequential portfolios yielded by the

[3] $|| \cdot ||_M$ in Equation (3.1) stands for the Mahalanobis distance. Also refers as the statistical distance, it takes into account the correlation the correlation between variables when computing distances. Lets $\vec{x}$ and $\vec{y}$ two random vectors of identical distribution with covariance matrix Σ, the Mahalanobis distance may be defined as $d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})' \Sigma^{-1} (\vec{x} - \vec{y})}$.

- different choices of the parameter k that measure the similarity between various market conditions as well as

- different distances d_l for measuring the similarity of the market conditions.

The aggregation is based on the past performance, in terms of cumulative wealth of each elementary strategy to produce a final allocation portfolio. This further refinement is of particular importance as it addresses the issue of the choice amongst various market conditions. The sequential method puts in competition concurrent proxies and follows a performance adaptive allocation of capital based on their recent wealth. In what follows I provide a more formal description of this additional step.

Given $\{q_{k,l}\}$, the probability distribution on the set of all pairs (k,l) of positive integers such that for all k and l, $q_{k,l} > 0$ and fixing the learning parameter[4] $\eta_t > 0$, the weights are defined by:

$$w_{k,l,t} = q_{k,l} e^{\eta \log C_{t-1}^{(k,l)}}.$$

where $C_{t-1}^{(k,l)}$ is the wealth accumulated by the pair (k,l) elementary strategy up to time $t-1$ with initial investment C_0. Their normalized values are then expressed as:

$$v_{k,l,t} = \frac{w_{k,l,t}}{\sum\limits_{k,l}^{K,L} w_{k,l,t}}. \tag{3.2}$$

Obviously, the weighing function $v_{k,l,t}$ must satisfy the usual constraints for combining elementary sequential strategies, namely $0 \leq v_{k,l,t} \leq 1$ and $\sum_{k,l,t} v_{k,l,t} = 1$ and each elementary strategies will receive a large weight if its past performance was relatively good.

Finally, the combined portfolio $\boldsymbol{\alpha}^{hf}$ at time t is then defined by weighting the set of $K \times L$ predicted portfolios according to their past performances and the positive distribution probability $\{q_{k,l}\}$ as follows:

$$\boldsymbol{\alpha}^{hf} = \sum_{k,l}^{K,L} v_{k,l,t} \boldsymbol{\alpha}(\mathbf{X}_{1,t-1}).$$

Therefore, under the dynamic *kernel-based* strategy $\mathbf{I}^K$, the investor's capital accumulated at time t can be expressed after some simplification as

$$C_t = \sum_{k,l}^{K,L} q_{k,l} C_t^{(k,l)}. \tag{3.3}$$

[4]Thorough this paper, I fixed the leaning parameter η_t as $1/\sqrt{t}$. For a deep discussion on the best practical choice of η_t, see Biau, Bleakley, Györfi, and Ottucsak (2009).

where $C_t^{(k,l)}$ is the final capital arose from the strategy associated to the pair (k, l).

The key idea of combining several concurrent portfolios is simple: to improve the final investment decision. Basically, the worse a portfolio performed in the recent past, the less it will contribute to the final allocation. The final prediction is constantly updated according to the recent performances of competitors.

3.2.3 Markovitz-type kernel-based portfolio selection

As an alternative unconditional allocation model, I perform the György and Vajda (2007) asymptotic MV kernel-based portfolio ($\mathbf{MV}^K$). The author extends the Markowitz (1952) MV characterization in a multi-period setting. While the approach is as in the dynamic *kernel-based* strategy $\mathbf{I}^K$, it differs from the fact the investor is concerned with a Markovitz-type utility function.

Just like before, for each pair (k, l) of positive integers, elementary portfolios $\tilde{\boldsymbol{\alpha}}$ are estimated as follows[5]:

$$\tilde{\boldsymbol{\alpha}}(\mathbf{X}_{1,t-1}) = \underset{\boldsymbol{\alpha}\in\Delta_m}{\text{argmax}} \left((1-2\lambda) \sum_{\{i\in S_t^{(k)}\}} (\langle \boldsymbol{\alpha} \, , \mathbf{X}_i \rangle - 1) \right.$$
$$\left. -\lambda \sum_{\{i\in S_t^{(k)}\}} (\langle \boldsymbol{\alpha} \, , \mathbf{X}_i \rangle - 1)^2 + \frac{\lambda}{\mid S_{t,(k,l)} \mid} \left(\sum_{\{i\in S_t^{(k)}\}} \langle \boldsymbol{\alpha} \, , \mathbf{X}_i \rangle \right)^2 \right) \tag{3.4}$$

if $S_t^{(k)} \neq \emptyset$, and used otherwise the past 24 months of data, as it is the case in Equation (3.2), to estimate the optimal portfolio weights. $S_t^{(k)}$ is the aforementioned similarity set outlined in Equation (3.1) and λ measures the investor's coefficient of relative risk aversion.

Furthermore, $\mathbf{MV}^K$ elementary portfolios are then aggregated as in the ($\mathbf{I}^K$) strategy according to:

$$\tilde{\boldsymbol{\alpha}}^{hf} = \sum_{k,l} \tilde{v}_{k,l,t} \tilde{\boldsymbol{\alpha}}(\mathbf{X}_{1,t-1}).$$

where $\tilde{v}_{k,l,t}$ are derived as in Equation (3.2). Using the same representation as in Equation (3.3), the wealth achieved by the $\mathbf{MV}^K$ strategy is given by

$$\tilde{C}_t = \sum_{k,l}^{K,L} q_{k,l} \tilde{C}_t^{(k,l)}.$$

[5]In appendix B, I provide additional mathematical details on the kernel-based MV portfolio selections.

3.2.4 Alternative Optimizers

The scope of this section is to illustrate alternative portfolio allocation strategies for comparison purpose. We compare the dynamic *kernel-based* strategy $\mathbf{I}^K$ to three different optimizers.

Our initial benchmark strategy is the equally weighed constantly rebalanced portfolio (**EWCRP**). We consider an investor who equally allocated her initial capital according to $\boldsymbol{\alpha}_0 = (1/m, \dots, 1/m)$ and at each trading period, rebalanced her portfolio to respect this uniform allocation constraint. This naive diversification strategy does not involve any optimization and estimation errors related to the plug-in of sample means and variance-covariance matrix.

Secondly, I implement the dynamic mean-variance portfolio selection technique. Under the standard mean-variance optimization, the investor selects $\boldsymbol{\alpha}_t$ to maximize the quadratic objective function

$$\mathcal{Q}(\boldsymbol{\alpha}_t) = \boldsymbol{\alpha}_t^{'}\mu_t - \frac{\lambda}{2}\boldsymbol{\alpha}_t^{'}\Sigma_t\boldsymbol{\alpha}_t \tag{3.5}$$

where μ_t is the conditional mean vector of the fund gain factors, Σ_t is the conditional variance-covariance matrix, λ expresses the investor's coefficient of relative risk aversion, and $\boldsymbol{\alpha}_t$ are nonnegative and sum up to one. To solve the problem in Equation (3.5), I consider an investor who replaces the mean and the covariance matrix by their sample counterparts $\hat{\mu}$ and $\hat{\Sigma}$, respectively. The investor believes that the conditional expected return and covariance are time-varying and will adjust his portfolio weights accordingly. Therefore, I use an autoregressive of order one model ($AR(1)$) to estimate the fund expected returns and the Weide (2002) generalized orthogonal GARCH(1,1) model to forecast the time varying conditional covariance matrix under the assumption of a joint normal distribution. In what follows, I refer to this investment strategy[6] as **GO-GARCH(1,1)**. The **GO-GARCH(1,1)** estimation method is based on common eigenvectors of the observed data. It does not suffer of numerical convergence problem[7].

The performance of the various strategies are compared using several measures. The first measure is the wealth achieved by each investment strategy and the annualized average yield (AAY). Then I compute the so-called Sharpe ratio, in an annualized basis. I estimate this ratio using the sample mean and variance of the excess returns for each strategy considered over the monthly US treasury bill. I also report a relative performance measure based on a modified version of the Sharpe ratio (mSR)

[6] In appendix B, complementary details are outlined on the **GO-GARCH(1,1)** dynamic MV portfolio selections.

[7] A sample size around $1000 - 1500$ is generally assumed to be the minimum for a precise estimation of GARCH(1,1) model. In our multivariate case the limit is even higher. Given the short history and low frequency of hedge fund data, a correct estimation of GARCH(1,1) model is not guaranteed. The choice of generalized orthogonal GARCH(1,1) is motivated by the limited sample size.

introduced by Graham and Harvey (1997), which provides the measure of out-performance of a given investment strategy over an alternative with different level of risk

$$mSR = \frac{\sigma_a}{\sigma_K}(\mu_K - \mu_a), \tag{3.6}$$

where μ_K and σ_K, μ_a and σ_a are the annualized average yield and the annualized volatility of the $\mathbf{I}^K$ investment strategy and those of the various alternative strategies, respectively.

Considering the dynamic structure of our analysis, it is important to assess the possible effect of transaction costs. Indeed, the gain of a dynamic investment strategy may be, partially or totally offset by transaction costs related to portfolio turnover. However, there is no general consensus among academicians about the right range of values in the financial industry. The issue is even more complicated for non traditional asset classes like hedge funds. For these reasons, to account for the transaction costs, I adopt the approach proposed by Han (2006). Since in practice it is difficult to derive the realized transaction costs, the author computes tc^{be}, the break-even transaction cost. The break-even transaction cost measures the level of transaction costs that a given investor will be indifferent to pay when selecting between the $\mathbf{I}^K$ investment strategy and the alternative allocation decisions. The break-even transaction cost between the two strategies is then defined as

$$tc^{be} = \frac{\bar{r}_p^K - \bar{r}_p^a}{tv^K - tv^a},$$

where $tv = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\left|\boldsymbol{\alpha}_{j,t} - \frac{\boldsymbol{\alpha}_{j,t-1}\mathbf{X}_{j,t}}{1+r_{p,t}}\right|$ is the monthly average of the value traded in all individual hedge fund in the portfolio, $\bar{r}_p^K$ is the average return of the portfolio constructed from the strategy $\mathbf{I}^K$, and $\bar{r}_p^a$ is the average return form one the alternative optimizers. If transaction costs are a fixed proportion ε of the value traded in the portfolio, the average transaction cost in each strategy is then $\varepsilon \times tv$. Therefore, if an investor has transaction costs smaller than tc^{be}, she will prefer the $\mathbf{I}^K$ investment strategy; otherwise, the investor will be better off with one of the three alternatives. In other words, high break-even costs is synonym to low portfolio turnover rates.

Finally, I calculate out-performance rate $\mathcal{S}$, that is, the percentage of allocation periods for which the $\mathbf{I}^K$ investment strategy performs better than one of the various alternative strategies. This measure is an expression of the *kernel-based* strategy ability to capture time-varying investment opportunities by discovering significant patterns in the local structure of the data in the past.

3.3 Data

This section presents the data used in our empirical investigation. I begin by discussing the individual hedge funds and commodity trading advisors (CTAs) in the Hedge Fund Research (HFR) and Barclay CTAs databases. Then, the section follows presenting the market risk factor specification.

3.3.1 Funds and sample restrictions

A proper study of hedge fund returns requires accurately measured data. Such a quality is a primary concern in the field since hedge fund managers *voluntarily* provide information to databases and the industry lacks an uniform reporting standard. Despite these difficulties a number of commercial databases are currently providing hedge fund information both at individual fund level as well as at aggregated level (indexes).

To analyze the relevance of the proposed methodology in constructing portfolio of hedge funds, I focus on live funds and use two main sources of hedge fund data: the Hedge Fund Research (HFR), and Barclay Commodity Trading Advisors (CTAs) databases. The HFR consists of returns and basics information for individual funds and fund of hedge funds from January 1981 to September 2008, and the Barclay CTAs database contains CTAs returns and fund specific comprehensive information from January 1980 to September 2008.

The HFR database is composed of returns reported on different frequencies (mainly monthly) and additional qualitative/quantitative information such as main strategy, sub-strategy, asset under management, etc. on 5230 individual funds and 2720 Funds-of-Funds that are still active on September 30, 2008. The data provider groups individual fund data in four main categories[8] depending on their sub-strategies: Macro, Relative Value, Equity Hedge, and Event-Driven. Out of the 5230 numbers of individual funds, 940 are classified as Macro, 2796 as Equity Hedge, 961 are Relative Value, and 533 are grouped as Event-Driven funds.

The Barclay CTAs data system is widely recognized by both practitioners and academicians as the largest, most comprehensive, available Commodity Trading Advisors sample. Generally, CTAs are funds primarily trading listed commodity and financial futures contracts. The database consists of 981 reporting funds as of September 2008. CTAs, also denominated Managed futures are by no means homogeneous investment vehicles. CTAs managers employ a large range of strategies and asset classes. Combined, the two databases provide a rather completed and detailed picture of the hedge funds universe.

For the scope of our empirical analysis, I impose a set of filters on both

[8] The reader may refer to **Appendix A** for a description on HFR hedge fund classification

databases. First I selected funds that reported in U.S. dollar net of fee on a monthly basis. Then I required that each fund has at least 15 years of non-missing reported returns. Additionally, for each strategy, I extract the group of funds with the longest consecutive stretch of non-missing returns and keep those funds having at least 10 years of reported assets under management (AUM). I end up with a final sample of 429 funds, which consists of 356 funds in HFR and 73 in Barclay CTAs. All funds in the final sample are opened for new investment. Since the author is interested in constructing multistrategy portfolios of hedge funds, I finally constituted for each hedge fund category, an heterogeneous basket of 16 funds based on their average first year reported AUM. According to Lhabitant (2004), in general Funds-of-Funds managers hold a basket of underlying funds ranging between 15 to 50. However, Lhabitant and Learned (2003) extending the previous work of Amin and Kat (2003), shown that approximately 10 funds are sufficient to construct a diversified portfolio of hedge funds. In line with this approach and the fund of hedge funds industry practice, I choose to fix the number of funds to sixteen and implement an asset-related selection process. We ranked the funds per average AUM and for each category, selected the five lowest and the top five, three funds directly below the median, the median ranking fund, and two funds above the median. A detailed representation of the data set is described in Panel A. of Table 3.1. Empirical experiments are performed on 5 styles (Global Macro, Equity Hedge, Relative Value, Event-Driven, and CTAs), excluding the Funds-of-Funds category. Naturally, this data selection process creates an additional survivorship bias in the residual time series of funds. I apply our empirical investigation on different structure of the data set to provide some robustness checks of our main results. In addition, since the purpose of our empirical study is to construct portfolios of hedge funds that maximize in the long run the investor wealth, and to study the portfolios relative performances under different allocation strategies, the fund selection routine may not be a big issue.

Furthermore, for consistency, I examine the impact of hedge fund redemption restrictions on the performances of the dynamic portfolio allocation. In general, the hedge fund industry uses what is commonly called share restriction provisions[9] such as redemption, lockup, and advance notice periods. Various authors have pointed the importance of redemption restrictions on hedge fund business model. Conclusive results on the subject are mixed. Some studies provide evidence that redemption restrictions are associated to excessive risk taking with fund's managers and potential costs on investors (Ang and Bollen, 2010; Klaus and Rzepkowski, 2009). Other studies, however, suggests that share restrictions

[9]Hedge funds restrictions also contain provisions such side pockets, gates, redemption suspension, $\cdots$. Some of those restrictions are not reported in the databases considered in the present study. We refer the reader to **Appendix A** for a short variables definitions.

are linked to higher hedge fund performance (Liang, 1999; Aragon, 2007; Agarwal, Daniel, and Naik, 2009). In general, these studies conclude that redemption restrictions affect various hedge fund characteristics such as fund flows, returns, and risk. It is then important to analyze the evolution of our dynamic allocation strategies under the complexity of hedge fund restriction provisions. Panel B of Table 3.1 reports average value of several characteristics for the sample of funds under investigation. along this panel, three variables capture restriction provisions: redemption period and notice period in days, and lockup period in months. Inspection of Panel B reveals a strong heterogeneity amongst hedge fund categories. Lockup periods are on average null for Barclay CTAs and Global Macro funds, with a maximum value of 8.25 months for Event-Driven. Redemption periods range between 3.75 and 170.83 days. Notice period measures are between 0 to 14.83. Over our samples, Barclay CTAs funds are far the least restrict.

3.3.2 Factor model specification

To define the variables that characterize the state of the market at every trading month, we approximate the Hasanhodzic and Lo (2007)'s six factors model. The authors show that, this factor model specification has a significant explanatory power amongst hedge fund strategies. They advocate the use of factors[10] that correspond to the main drivers affecting the hedge fund's risk-return tradeoff:

The equity market: the S&P 500 total return.

Currencies: the US Dollar major currencies index.

The bond market: the month end-to-month end return on the Moody's corporate Aa bond index.

Credit: the monthly change in the Moody's corporate Baa bond index less the three months treasury bill (month end-to-month end).

Commodity: the Goldman Sachs commodity index total return.

Volatility: the first difference of the end-of-month value of the CBOE volatility index.

[10] Data are downloaded from various sources: S&P 500 total return from DataStream; Goldman Sachs commodity (GSCI) and the Volatility (VIX) indexes are from yahoo finance; US three months treasury bill and Moody's corporate Aa bond, Moody's corporate Baa bond and US Dollar major currencies indexes from the FED bank of St. Louis data library.

3.3.3 Data summary statistics

We begin our analysis reporting some preliminary statistical characteristics on the various fund categories under our study. The data are monthly, ranging from July 1993 to December 2007 for Global Macro, December 1993 to August 2007 for Equity Hedge, December 1993 to January 2008 for Event-Driven, September 1993 to December 2007 for Relative Value, and from October 1993 to June 2008 for the Barclay CTAs category. From Panels A to E of Table 3.2, except in few cases, average returns are in general positive and significant. Through all categories, volatilities range between 18% and 0.1% per month. The less volatile hedge fund strategy is the Relative Value group. The highly volatile strategies are Global Macro and Barclay CTAs funds. The measure of asymmetry (skewness) is heterogeneous across individual funds and main strategies. Barclay CTAs and Macro funds are in general positively skewed, indicating that booms occur more often than crashes in those categories. In opposite, Relative Value and Event-Driven strategies are more often negatively skewed. In all investment styles, kurtosis measures are larger than 3, which is not consistent with the normality assumption. The hypothesis of normality in fund's return series is generally rejected by both the Jarque-Bera and Lilliefors tests.

Regarding time dependency in individual hedge fund performances, the natural way of testing it is the Ljung-Box statistic (ρ). We estimate the Ljung-Box statistic for returns and squared returns to assess the presence of serial correlation in the first and second moment, respectively. Taking apart Relative Value and Event Driven funds, there is no strong evidence that the series are serially correlated. However, squared returns are more often serially correlated at 95% confidence level, which indicates temporal correlation in volatilities.

3.4 Implementation details

In this section, we discuss some practical implementation issues. First, I focus on the *kernel-based* portfolio selections, namely, the $\mathbf{I}^K$ and the $\mathbf{MV}^K$ optimizers. As described in Section 3.2, elementary strategies are constructed for different proxy of market conditions, different number of past values of the proxy used in measuring the similarity of market conditions (parameter k) as well as for different values of the maximal distance between the values of these proxies (parameter $\mathrm{d}_{k,l}$). The choice of the last two parameters is as follows: I initially define the two positive integers as $l = 1, \cdots, L$ and $k = 1, \cdots, K$, fixing $L = 10$ and varying $K = 7$. Additionally I use the uniform probability distribution $\{q_{k,l}\} = 1/(K \times L)$ for the weighting scheme of elementary strategies indexed over k and l. The maximal distance $d_{k,l}$ is defined by:

$$d_{k,l}^2 = \exp((0.002 \cdot k \cdot m) + (0.02 \cdot m \cdot l)) + (k/2)$$

Then, for each k, we have $l = 10$ different values of the radius $d_{k,l}$ and an array of 10 different elementary portfolios.

For all hedge fund categories, I fix the starting point of the dynamic portfolio prediction algorithm after 12 months. The first year of the data is dedicated to the initial parametrization. Therefore, the first data selection process defined in Equation (3.1) uses this initial window to identify the similarity set. However, this window increases in a monthly basis with the trading period.

The final maximized wealth at time t is obtained by exponentially weighting the $K \times L$ elementary sequential strategies according to their past performances as described in Equations (3.3, 3.5).

The initial investment capital C_0 is set equal to 1. The estimation of the portfolio weights for MV dynamic strategies is implemented for different choices of the investor's relative risk aversion coefficient. That is, we set λ equal to 5, 10, and 15.

To make the *kernel-based* portfolio construction implementable, some usual hypothesis are assumed. In addition to the assumption of local stationarity and ergodicity of the market, I assume that any fund is infinitively divisible and that all funds are opened for new investment at each trading period. The dynamic allocation problem is expressed here from the view of a long only agent investing in US dollars, so the annualized average yields are expressed in US$. Last, the $\mathbf{I}^K$ and the MV (kernel-based and GO-GARCH(1,1)) investment processes are solved using the R version of the Spellucci (1997) DONLP2 optimization routines.

3.5 Empirical analysis

We now focus on the empirical analysis of the performance of the various investment strategies illustrated in the present article. This section displays some results of the nonparametric *kernel-based* portfolio selection, compared to the three alternative investment strategies. We will provide evidence of the benefit that results of switching from a dynamic mean-variance allocation of resources to a kernel-based distribution free strategy. Table 3.4 documents the statistics of the wealth achieved by the various trading strategies under a month-to-month portfolio construction, while Table 3.5 reported the results of the strategies under different set of market restrictions. The evolutions of the accumulated capital for each strategy are displayed through Figures 3.1 to 3.10 (for a relative risk aversion of $\lambda = 5$). We tested the investment strategies on 5 distinct categories of hedge funds: HFR Global Macro, Equity Hedge, Event-Driven, Relative Value and the Barclay CTAs. For each category we constructed a heterogeneous panel of 16 individual hedge funds as described in Section 3.3. This means that $m = 16$.

3.5.1 Performance analysis of the dynamic investment strategies

The estimation of the optimal portfolio allocation is carried out each month by maximizing the average rate of growth of capital. Table 3.4 reports the sample performance of the $\mathbf{I}^K$, $\mathbf{MV}^K$, **GO-GARCH(1,1)** and **EWCRP** strategies for the 5 categories of hedge fund. The results for the two mean-variance specifications ($\mathbf{I}^K$, **GO-GARCH(1,1)**) are outlined for $\lambda = 5, 10, 15$. As the panels reveal, the $\mathbf{I}^K$ investment strategy show superior annualized average yield (AAY) across all styles. In the absence of transaction costs, the percentage values of AAY range from 40.80% for Barclay CTAs portfolio to 23.5% in Event-Driven funds. These numbers are always significantly bigger than those yield by the three alternative strategies. As expected, the ex-post annualized standard deviations (ASD) are also bigger, which translate into a smaller Sharpe ratios (SR), specially compared to the $\mathbf{MV}^K$ portfolios (for all values of λ). Except for Barclay CTAs funds, the SR of the dynamic $\mathbf{I}^K$ strategy are bigger than one across all hedge fund styles and are equal magnitude to **GO-GARCH(1,1)** and **EWCRP** strategy values.

On a relative basis, the success rate, $\mathcal{S}$, that is, the proportion of months for which the $\mathbf{I}^K$ investment strategy has superior return than the alternative optimizers, further illustrate the attractiveness of the former strategy. Over the sample of hedge fund styles under scrutiny, the success rate is between 50% and 75.82%, a range that is consistent with an upgrade quality to capture time varying investment opportunities. This evidence suggests that the out-performance of the dynamic $\mathbf{I}^K$ strategy is not due to some specific extreme events. The modified Sharpe ratio measure, mSR, is the return that the $\mathbf{I}^K$ strategy would have earned if it had the same risk as the alternative optimizer. Across Panels A to E, mSR values are positive and bigger for the the **GO-GARCH(1,1)** and **EWCRP** strategies than for the $\mathbf{MV}^K$ strategy.

When the investor considers the dynamic changes in the variance-covariance matrix under the *kernel-based* framework ($\mathbf{MV}^K$ strategy), the SR significantly increase for all values of λ and across the 5 hedge fund categories. Indeed, the $\mathbf{MV}^K$ strategy show superior SR values. For the same coefficient of relative risk aversion, the $\mathbf{MV}^K$ strategy tends to exhibit higher average annual yield and lower annualized volatility than the **GO-GARCH(1,1)** strategy. Accordingly, their relative performance (mSR and $\mathcal{S}$) to the $\mathbf{I}^K$ strategy are in favor of $\mathbf{MV}^K$. As a consequence, in contrast to the dynamic **GO-GARCH(1,1)** portfolios, the $\mathbf{MV}^K$ portfolios benefit from volatility timing. However, one has to take with cautious the value added by MV strategies since they are subjected to estimation risk (uncertainty regarding the estimation of $\hat{\mu}$ and $\hat{\Sigma}$).

The results documented above have been derived without transaction costs. It is clear that if we take into account transactio costs, the above presented performance will deteriorate. Table 3.4 also reveals the break-even transaction costs, tc^{be}, of the alternative three strategies for all levels of risk aversion. The numbers are listed in the last column entry from Panel A to E and are different across hedge fund categories and investment strategies. In general, highest tc^{be} suggests lower portfolio turnover rates for the strategy. For Global Macro, Equity Hedge, and Event-Driven funds, **GO-GARCH(1,1)** strategies show lower break-even costs than $\mathbf{MV}^K$ and **EWCRP** strategies. The results are opposite for Barclay CTAs and Relative Value funds. For the later, the tc^{be} are positive and higher with **EWCRP** strategies.

Last, but not least, Figures 3.1 to 3.10 depict the evolution of wealth accumulated for each model specification over the allocation period. One can clearly see that the $\mathbf{I}^K$ strategy significantly dominates the others. Overall, these results suggest that the time evolving *kernel-based* investment strategies provide significant performances for a risk-seeking investor who is willing to bear the risk - in terms of higher volatility. Amongst *kernel-based* strategies, the distribution free $\mathbf{I}^K$ strategy outperforms the $\mathbf{MV}^K$ counterpart. The results show the superiority of $\mathbf{I}^K$ and $\mathbf{MV}^K$ models, supporting the hypothesis that modeling hedge fund returns in a nonstationary, unconditional paradigm yields significant economic benefits.

3.5.2 Robustness assessment

To further evaluate the robustness of the analysis in constructing a multi-strategy portfolio of hedge funds, we have implemented an additional set of more restrictive empirical experiment, the principal results of which are presented in this section. These investigations are performed on the same data, however with different specification.

In a first analysis, we assess the effect of hedge fund redemption provisions on the portfolio performance characteristics. We adjusted the dynamic allocation of resources to take into account average restriction provisions. We focus on two variables. The lockup period in months, and the redemption period which is the number of days between two consecutive redemption dates. They vary across funds, and categories. On average, in the portfolio of 16 funds under study, the lockup period is between 0 (HFR Global Macro and Barclay CTAs) and 8.25 months (Event-Driven). Redemption period ranges between 45.08 (HFR Macro) to 170.83 days (Event-Driven). For the portfolio of HFR Global Macro and Barclay CTAs funds, with on average no lockup period, we used the full sample average (per style) restriction provisions to impose the time sequence of the investment routine. This yields 3 (2) months redemption periods and 12 (6) months lockup periods for HFR Global Macro (Barclay CTAs) category. For the three other hedge fund styles, we restrict ourself to what is displayed in Panel B. of Table 3.1. Therefore, the experiment is conducted in the following way: After our initial allocation of resources to the portfolio, we block this investment for the time indicates by the average lockup period. Next, the investment is dynamically re-allocated at each pre-specified redemption dates as reported in Panel B. of Table 1. Table 3.5 reports the absolute and relative performance for all strategies and level of risk aversion. As is evident, the performances of the various strategies are just slightly different from those obtained in the month-to-month portfolio construction. By respecting the hedge fund industry restriction provisions, the Sharpe ratios of $\mathbf{I}^K$ strategies improve for HFR Global Macro and Equity Hedge styles and are relatively unchanged for the rest. Once again, the success rate enhanced the $\mathbf{I}^K$ investment strategy, with percentage values between 48% to 69%. The risk-adjusted excess return (mSR) is always positif, with bigger values relative to **GO-GARCH(1,1)** and **EWCRP** strategies.

Secondly, we performed a data driven model evaluation method to question our fund selection process. The analysis intends to test whether different competing choices of sample constituents would lead to equivalent performance characteristics of the investment dynamics. The implemented test is an approximation of the procedure advised by Racine and Parameter (2009). The experiment is flexible and allows to defeat the inconveniences associated on relying on only one choice of data. In this article, we intend to construct the distribution of the investment strategy's true annualized average yield, by randomly testing the strategies under various structures of the data. Therefore, instead of using a fix sample of 16 funds selected according to their ranking AUM as explained in Section 3.3, we draw randomly a basket of funds of size 16, then perform our allocation strategies accordingly. This process is repeated 150 times, each time reassessing the performances of our dynamic portfolios across the four investment methodologies. The repeated investigation will produce vectors of length 150 of annualized performance statistics for all models under scrutiny. To discriminate between different allocation strategies, we use a paired t-test of differences in sample mean and a the Mann-Whitney-Wilcoxon test for differences in location. Some useful Boxplots are also represented to highlight dominance relationships between alternative optimizers. For what follows, we consider an application on HFR Equity hedge funds. The data set consists of 97 funds, obtained after imposing several filters as explained in Section 3.3. The results are reported for the coefficient of relative risk aversion, $\lambda = 5$. Observing 3.6 reveals that the $\mathbf{I}^K$ allocation specification is preferred to all three alternative investment strategies on the AAY basis. Both the t-test and the Mann-Whitney-

Wilcoxon p-values are close to zero. From Figure 3.11, the median value for the distribution free approach is equal to 26.2%, which is of equal magnitude to what achieved by the sample of 16 funds ranked according to their AUM. However, on the relative performance basis expressed by the Sharpe ratio distributions in Figure 3.12, the **GO-GARCH(1,1)** and **EWCRP** strategies performs better than the $\mathbf{I}^K$ and $\mathbf{MV}^K$ strategies. To summarize, the $\mathbf{I}^K$ allocation higher annualized yields, at the cost of an increase in the annualized volatility.

Overall, the results of the present section are in line with Subsection 3.5.1. The empirical robustness analysis confirms the dominance of the *kernel-based* investment strategies. That is, this investment technique over-performs the two alternative optimizers in terms of AAY and outperformance rate ($\mathcal{S}$), while these results are mixed on a risk-adjusted basis. Additionally, this dominance relation is in general unaffected by different market and statistical restrictions. These results suggest that, across our hedge fund category and investment horizon, allocating capital on the basis of an unconditional, nonparametric modeling is likely to provide sizeable benefits in the allocation process.

3.6 Conclusion

In this manuscript, the performance of various allocation strategies across individual hedge funds is investigated. We particularly illustrated the time varying *kernel-based* approach to solve the optimal allocation of resources within various hedge fund styles. An empirical study is performed on HFR and Barclay CTAs databases. For the first, in our knowledge, this article analyzes the risk-return characteristics of hedge fund portfolios when relaxing distribution hypothesis. Indeed, under the *kernel-based* framework, the hypothetical investor is concerned by the long term maximization of his average capital rate of growth in a time evolving optimization setting, without knowing the underlying distribution generating the hedge fund's NAV. I investigated the benefits of approximating nonstationary data locally on a multivariate portfolio of hedge funds.

I found that the distribution free investment strategy provides reliably results. Empirical findings suggest that there are distinct benefits in performance improvement for portfolios constructed under the *kernel-based* investment strategy. On a relative basis, the proportion of months for which the strategy outperforms alternative investment approaches is high. The annualized risk-return profile, expressed by the Sharpe ratio, is sometimes undermined by the high volatility associated to dynamically investing in hedge funds. Furthermore, this strategy ignores the volatility problem and it is particularly constructed for an investor concerned by any parametric hypothesis on the distribution. I further perform various experiments designed to assess the robustness of the results. Evidences suggest that our findings are solid, and the magnitude of the performances compared to two conditional alternative optimizers is invariant.

Several extensions to this paper may be considered. Future researches might address the issue of how to incorporate a risk measure in the optimization routine without distorting the completely nonparametric structure of the methodology. Transaction costs are also ignored in this article. Additional works are necessary to fully incorporated the cost related to the time-varying allocation of capital across hedge funds. Furthermore, investors do allocate resources to hedge funds in combination with different asset classes, in order to achieve the desired performance profile. Therefore, it might be interesting to assess the return pattern of an aggregate portfolio of hedge fund and several traditional asset classes, under the *kernel-based* investment strategy.

Clearly, the core idea of this manuscript is that there exist a class of distribution free allocation strategies that are able to uncover and exploit hidden structure in the past of hedge funds data, to optimally allocate capital in a multistrategy portfolio.

Acknowledgements

The author is grateful to professor Catalin Starica for his insightful comments and suggestions.

3.7 Appendices

3.7.1 Appendix A: Hedge Fund Characteristics

Variable Definitions

NAV$_t$: The month t net asset value of a given hedge fund. As Lhabitant (2004), we assume that the **NAV**$_t$ has been adjusted to take into account all realized and non-realized capital gains, accrued dividends and interest income, capital distributions, splits and all the impacts of equalization and crystallization. The gain factor or simple gross return at time $t + 1$, of a fund j is defined as: $\frac{\mathbf{NAV}_{j,t+1}}{\mathbf{NAV}_{j,t}} = (1 + r_{j,t+1})$.
Lockup Period: The minimum number of months that an investor has to wait, after his initial investment, before he can withdraw money from the fund.
Redemption Period: Number of days between two consecutive specified dates. Hedge fund investors are allowed to take back their capital only at these pre-specified intervals.
Notice Period: The advance number of days that a hedge fund investor is required to inform the fund of his willingness to withdraw his capital.
Incentive Fee (%): Also called Performance fee, it is the main source of hedge fund's profits. It is shaped as an option to provide incentives for a hedge fund manager to generate profits. Calculated as a fraction of profits generated above the high-water mark.
Management Fee (%): Designed to provide to the fund's manager enough money to cover his operating costs. It is a fraction of the fund's asset under management.
Minimum Investment($M): The minimum amount of money required by the fund to be accepted as an investor.
High Water Mark: Defines the fund's manager compensation. The clause means that the manager receives performance fee only on increases in the NAV of the fund in excess of the previous highest NAV.
Leverage: Percentage of capital borrowed by the fund to boost the potential profit of the strategy.

Investment Strategies

The term 'hedge funds' is a generic acronym associated to a pool of professionally managed capital. However, hedge funds are highly heterogeneous, covering a large range of investment styles with various approaches, objectives and performance characteristics. There is no consensus in the financial literature in the way to classify the wide basket of strategies implement in the hedge fund universe. For simplicity, we adopt Lhabitant (2004) classification which is closed to Chicago-based Hedge Fund Research database. Hedge funds are classified into five main strategies: tactical trading, equity hedge, event-driven, relative value and funds of funds.

Tactical trading strategies: Also called directional funds, the term refers to funds that speculate in the direction of market prices of commodities, currencies, bonds and/or equities. They trade on a discretionary or directional basis. CTAs and Global macro funds belong to this category.

Equity Hedge: Indicates managers implementing long/short investment strategies in equities. The strategy is not automatically market neutral. The category can be further divided into different sub-strategies, based on manager invest target as: Sector or/and geographically focus, emerging market, dedicated short bias, and market timers.

Event-driven: As their name claimed, event-driven funds are particularly interested in situations where the investment opportunity is associated to specific corporate events such as: mergers and acquisitions, bankruptcy, recapitalization, stock buybacks. Fund's managers invest mainly on equity, debt or trade claim from those companies. Risk arbitrage and distressed securities funds dominate this group.

Relative value: They intend to profit on pricing differences between similar or related assets such as options, futures, equities and debt. The underlying assumption is that the price gap between the two similar/related securities will return to its fair value as the investment horizon growths. This category regroups strategies such as: fixed income arbitrage, convertible arbitrage, statistical arbitrage, index arbitrage and mortgage-backed securities arbitrage.

Funds of funds: This category represents funds investing in a pool of hedge funds. The investment principle relies on the assumption that combining individual funds will reduce the risk and provide a more stable return in the long run. The manager may allocate his resource within a strategy, or in multiple strategies.

3.7.2 Appendix B: Alternative optimizers

Asymptotic Mean Variance Portfolio Model: Following György and Vajda (2007) we briefly define the conditional expected value of the Markowitz-type utility function as:

$$\begin{aligned}\mathbb{E}\{U_M(\langle \boldsymbol{\alpha}(\boldsymbol{X}_{1,T-1}), \boldsymbol{X}_T\rangle, \lambda) \mid \boldsymbol{X}_{1,T-1}\} &\overset{def}{=} \mathbb{E}\{\langle \boldsymbol{\alpha}(\boldsymbol{X}_{1,T-1}), \boldsymbol{X}_T\rangle \mid \boldsymbol{X}_{1,T-1}\} \\ &\quad -\lambda \mathbf{Var}\{\langle \boldsymbol{\alpha}(\boldsymbol{X}_{1,T-1}), \boldsymbol{X}_T\rangle \mid \boldsymbol{X}_{1,T-1}\}\end{aligned} \tag{3.7}$$

where $\boldsymbol{X}_T$ is the T-*th* day market vector, $\alpha(\boldsymbol{X}_{1,T-1})$ is the investor portfolio weight and λ is a positive constant coefficient representing the investor absolute risk aversion.

Under mild conditions, equation 3.7 can be expressed in as follow:

$$\begin{aligned}\mathbb{E}\{U_M(\langle\boldsymbol{\alpha}(\boldsymbol{X}_{1,T-1}),\boldsymbol{X}_T\rangle,\lambda)\mid \boldsymbol{X}_{1,T-1}\} &= (1-2\lambda)\mathbb{E}\{\langle\boldsymbol{\alpha}(\boldsymbol{X}_{1,T-1}),\boldsymbol{X}_T\rangle - 1\mid \boldsymbol{X}_{1,T-1}\}\\ &\quad -\lambda\mathbb{E}\{(\langle\boldsymbol{\alpha}(\boldsymbol{X}_{1,T-1}),\boldsymbol{X}_T\rangle-1)^2\mid \boldsymbol{X}_{1,T-1}\}\\ &\quad +1-\lambda+\lambda\mathbb{E}^2\{\langle\boldsymbol{\alpha}(\boldsymbol{X}_{1,T-1}),\boldsymbol{X}_T\rangle\mid \boldsymbol{X}_{1,T-1}\}.\end{aligned}$$

finally, the kernel-based mean-variance portfolio is expressed as:

$$\tilde{\boldsymbol{\alpha}}(\boldsymbol{X}_{1,t-1}) = \underset{\boldsymbol{\alpha}\in\Delta_m}{\operatorname{argmax}}\left((1-2\lambda)\sum_{\{i\in S_{t,(k,l)}\}}(\langle\boldsymbol{\alpha}\ ,\ X_i\rangle-1) - \lambda\sum_{\{i\in S_{t,(k,l)}\}}(\langle\boldsymbol{\alpha}\ ,\ X_i\rangle-1)^2 + \frac{\lambda}{\mid S_{t,(k,l)}\mid}\left(\sum_{\{i\in S_{t,(k,l)}\}}\langle\boldsymbol{\alpha}\ ,\ X_i\rangle\right)^2\right)$$

where $S_{t,(k,l)}$ is the similarity set associated to each pair (k,l).

GO-GARCH Portfolio Selection: To relax the notation, let's denote by X_t the vector of m funds's gain factors at time t. We assume that the dynamics of the gain factors vector is

$$X_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t,$$

where ε_t is the vector of unexpected returns. We estimated an $AR(1)$ model to filter out autocorrelation. Finally, I adjust the Weide (2002) **GO-GARCH(1,1)** model for the residuals of the $AR(1)$ to estimate the time varying conditional covariance matrix.

Briefly, the fundamental assumption behind the **GO-GARCH** model is the following:

Given an observed economic process $\{x_t\}$, there exists an uncorrelated components $\{y_t\}$ such that

$$x_t = Zy_t$$

where Z is the linear map that links the observed economic process to the unobserved components is assumed to be constant and invertible. Associated to the **GO-GARCH(1,1)** process:

$$\begin{aligned} x_t &= Zy_t\ y_t \sim N(0,H_t) \\ \boldsymbol{H_t} &= diag(h_{1,t},\cdots,h_{m,t}) \qquad (3.8)\\ h_{i,t} &= (1-\alpha_i-\beta_i)+\alpha_i y_{i,t-1}^2+\beta_i h_{i,t-1} \quad i=1,\cdots,m \qquad (3.9)\end{aligned}$$

with $\boldsymbol{H_0} = \boldsymbol{I}$ represents the unconditional covariance matrix of the components. Therefore, the conditional covariance matrix of $\{x_t\}$ are obtained by:

$$\boldsymbol{V}_t = \boldsymbol{Z}\boldsymbol{H}_t\boldsymbol{Z}'$$

Table and Figure Captions

Table 3.1: This table presents the fund's filtering procedure across all hedge fund categories (Panel A) and summary statistics of individual hedge fund characteristics (Panel B) such as: Leverage, High Water Mark, Redemption period, Management Fee, Incentive Fee, Notice period, Lockup period, Minimum Investment.

Table 3.2: This table presents summary statistics on individual hedge funds gain factors, across all styles through Panel A to E: the Mean, the standard deviation [SD], the skewness, the kurtosis, the minimum [Min], the maximum [Max], the t-statistic of mean, skewness and kurtosis, the Jarque-Bera [JB Test] and Lilliefors normality test statistics, and the [JB Test] 95% P-value. The critical value of the Lilliefors test at 95% confidence level is $0.805/\sqrt{T}$, where T is the sample size.

Table 3.3: This table presents summary serial correlation statistics on individual hedge funds amongst all styles through Panel A to Panel E: the first-order serial correlation of gain factors [$\rho(r)$] and of squared gain factors [$\rho(r^2)$], with their respective 95% P-value.

Table 3.4: This table presents the annualized summary statistics on the optimal portfolio constructed across the 5 styles under scrutiny. Through Panel A to E, the results of various allocation strategies are reported. Several measures of performance are outlined: the annualized average yield [AAY], the annualized standard deviation [sdev], the Sharpe ratio [SR], the modified Sharpe ratio [mSR] as defined in Equation (3.6), the out-performance rate [$\mathcal{S}$], that is, the percentage of allocation periods for which the $\mathbf{I}^K$ investment strategy performs better than one of the various alternative strategies. The coefficient of relative risk aversion, λ are set equal to $5, 10$, and 15.

Table 3.5: This table presents the annualized summary statistics on the optimal portfolio constructed across the 5 styles under scrutiny, in the case of respecting hedge funds redemption restrictions such as Lockup period and Redemption period. Through Panel A to E, the results of various allocation strategies are reported. Several measures of performance are outlined: the annualized average yield [AAY], the annualized standard deviation [sdev], the Sharpe ratio [SR], the modified Sharpe ratio [mSR] as defined in Equation (3.6), the out-performance rate [$\mathcal{S}$], that is, the percentage of allocation periods for which the $\mathbf{I}^K$ investment strategy performs better than one of the various alternative strategies. The coefficient of relative risk aversion, λ are set equal to $5, 10$, and 15.

Table 3.6: This table presents the results of robustness assessment of the fund selection process. We test the effect of drawing various samples of funds, on the dominance relationships between our four investment strategies. The experiment is performed on the HFR Equity hedge fund category. Test statistics on paired t-test [t-statistics] of differences in sample mean and the Mann-Whitney-Wilcoxon test [M-W tests] for differences in location are reported, with their respective 95% P-values in brackets.

Figures 3.1, 3.3, 3.5, 3.7, 3.9: These figures display respectively the evolution of the wealth achieved, investing in the HFR Global Macro, CTAs Barclay, HFR Equity Hedge, HFR Relative Value, and HFR Event-Driven funds under our four investment strategies. The investor's coefficient of relative risk aversion, $\lambda = 5$.

Figures 3.2, 3.4, 3.6, 3.8, 3.10: These figures display respectively the evolution of the wealth achieved, investing in the HFR Global Macro, CTAs Barclay, HFR Equity Hedge, HFR Relative Value, and HFR Event-Driven funds under our four investment strategies, in the case of respecting hedge funds redemption restrictions such as Lockup period and Redemption period. The investor's coefficient of relative risk aversion, $\lambda = 5$.

Figures 3.11, 3.12: These figures are Boxplots of respectively, the annualized average yield [AAY] and the annualized Sharpe ratio for the 150 samplings of HFR Equity Hedge funds. Those Boxplots intends to highlight the the dominance relationships between our four investment strategies in different structure of the data set. The coefficient of relative risk aversion, $\lambda = 5$.

Table 3.1 – *HFR individual funds summary characteristics*

Panel A: Data Filtering Process			
HF Categories	Initial	After Filtering	Ranked per AUM
	# of funds	#-of-funds	#-of-funds
HFR Macro	940	58	16
HFR Equity Hedge	2796	97	16
HFR Relative Value	961	45	16
HFR Event-Driven	533	53	16
HFR Funds of Funds	2720	103	16
Barclay CTAs	981	73	16
Total	8931	429	96

Panel B: Summary of Fund Characteristics					
Main Strategy	Macro	Equity Hedge	Event-Driven	Relative Value	Barclay CTAs
Leverage (%)	92	50	50	80	94
High Water Mark (%)	83.3	66	81	73	6
Redemption period (days)	45.08	131.25	170.83	58	3.75
Management Fee(%)	2.66	1.02	1.31	1.2	2.1
Incentive Fee(%)	18.94	15.41	16.86	15.53	20.6
Notice period (days)	14.83	32.92	50	43.67	0
Lockup period (months)	0	7	8.25	1.6	0
Minimum Investment ($M)	5.24	3.33	0.82	1.89	5.84

Table 3.2 – *HFR individual funds general descriptive statistics*

Panel A: Global Macro funds

Funds	Mean	SD	Skewness	Kurtosis	Min	Max	Mean t-statistic	Skewness t-statistic	Kurtosis t-statistic	JB Test	P-value P-value	Lilliefors Test
Fund1	1.02	0.11	0.24	3.67	0.71	1.34	122.92	1.27	1.82	4.91	0.09	0.09
Fund2	1.01	0.05	-0.29	4.46	0.81	1.19	252.94	-1.55	3.94	17.89	0.00	0.06
Fund3	1.02	0.08	0.62	5.55	0.71	1.32	162.69	3.33	6.87	58.34	0.00	0.12
Fund4	1.01	0.06	-0.02	2.89	0.86	1.19	209.44	-0.08	-0.30	0.09	0.95	0.04
Fund5	1.01	0.08	0.53	4.56	0.81	1.28	174.31	2.85	4.20	25.73	0.00	0.08
Fund6	1.02	0.10	0.94	5.68	0.80	1.44	135.37	5.05	7.22	77.64	0.00	0.09
Fund7	1.01	0.03	0.81	5.57	0.92	1.16	401.53	4.38	6.91	67.03	0.00	0.08
Fund8	1.01	0.05	0.30	3.57	0.86	1.15	263.80	1.63	1.54	5.01	0.08	0.07
Fund9	1.01	0.07	0.63	3.83	0.83	1.28	183.94	3.39	2.23	16.49	0.00	0.09
Fund10	1.01	0.05	-0.24	4.65	0.80	1.14	286.40	-1.29	4.44	21.35	0.00	0.05
Fund11	1.01	0.03	0.55	3.45	0.94	1.12	427.48	2.97	1.20	10.28	0.01	0.07
Fund12	1.01	0.07	-0.66	9.01	0.64	1.23	202.47	-3.54	16.19	274.58	0.00	0.07
Fund13	1.01	0.04	0.04	2.78	0.89	1.12	308.09	0.22	-0.59	0.39	0.82	0.03
Fund14	1.01	0.05	0.54	4.18	0.90	1.20	282.50	2.90	3.18	18.52	0.00	0.07
Fund15	1.02	0.10	0.23	3.21	0.76	1.29	132.04	1.25	0.56	1.86	0.39	0.04
Fund16	1.01	0.03	0.41	4.39	0.93	1.12	456.57	2.21	3.74	18.84	0.00	0.07

Panel B: Barclay CTAs funds

Funds	Mean	SD	Skewness	Kurtosis	Min	Max	Mean t-statistic	Skewness t-statistic	Kurtosis t-statistic	JB Test	P-value P-value	Lilliefors Test
Fund1	1.00	0.11	0.20	6.94	0.47	1.37	117.28	1.07	10.71	115.78	0.00	0.18
Fund2	1.03	0.18	2.33	14.59	0.60	2.20	76.33	12.67	31.47	1151.25	0.00	0.16
Fund3	1.01	0.07	0.48	4.12	0.79	1.22	202.91	2.59	3.03	15.87	0.00	0.07
Fund4	1.01	0.03	0.67	4.40	0.93	1.13	417.48	3.64	3.80	27.72	0.00	0.07
Fund5	1.01	0.04	0.56	3.85	0.91	1.15	339.37	3.05	2.32	14.67	0.00	0.08
Fund6	1.01	0.03	0.69	4.67	0.94	1.13	502.34	3.76	4.54	34.71	0.00	0.06
Fund7	1.02	0.12	1.46	9.52	0.71	1.75	109.62	7.92	17.71	376.40	0.00	0.08
Fund8	1.01	0.02	3.89	21.59	0.98	1.16	544.71	21.11	50.49	2995.14	0.00	0.27
Fund9	1.00	0.02	0.81	4.96	0.96	1.06	863.82	4.38	5.33	47.59	0.00	0.08
Fund10	1.02	0.07	0.08	2.95	0.86	1.20	203.05	0.42	-0.14	0.20	0.91	0.04
Fund11	1.01	0.10	0.05	9.30	0.52	1.48	141.40	0.27	17.11	292.90	0.00	0.08
Fund12	1.02	0.09	0.51	6.22	0.75	1.44	158.23	2.79	8.74	84.21	0.00	0.06
Fund13	1.01	0.04	0.81	4.07	0.92	1.17	311.15	4.37	2.90	27.57	0.00	0.09
Fund14	1.01	0.04	0.03	2.87	0.89	1.12	315.51	0.18	-0.34	0.15	0.93	0.03
Fund15	1.01	0.05	0.05	3.85	0.84	1.17	275.80	0.29	2.30	5.35	0.07	0.04
Fund16	1.01	0.07	0.57	3.70	0.82	1.28	180.83	3.11	1.90	13.29	0.00	0.08

Panel C: Equity Hedge funds

Funds	Mean	SD	Skewness	Kurtosis	Min	Max	Mean t-statistic	Skewness t-statistic	Kurtosis t-statistic	JB Test	P-value P-value	Lilliefors Test
Fund1	1.02	0.06	0.80	8.00	0.78	1.31	213.71	4.22	13.11	189.70	0.00	0.11
Fund2	1.01	0.06	0.45	6.15	0.80	1.27	208.66	2.37	8.25	73.65	0.00	0.15
Fund3	1.01	0.04	-0.63	3.91	0.86	1.10	323.36	-3.28	2.38	16.43	0.00	0.05
Fund4	1.01	0.09	1.18	10.72	0.70	1.52	149.58	6.17	20.25	448.20	0.00	0.11
Fund5	1.01	0.05	-0.31	3.60	0.84	1.16	249.76	-1.64	1.58	5.21	0.07	0.06
Fund6	1.00	0.02	-0.36	3.77	0.95	1.05	806.07	-1.88	2.01	7.58	0.02	0.07
Fund7	1.01	0.06	-2.24	13.84	0.66	1.17	232.45	-11.73	28.42	945.51	0.00	0.15
Fund8	1.01	0.07	0.64	6.39	0.77	1.32	182.98	3.35	8.89	90.18	0.00	0.10
Fund9	1.01	0.05	1.26	16.62	0.81	1.35	253.70	6.60	35.72	1319.40	0.00	0.13
Fund10	1.01	0.06	-0.48	3.83	0.83	1.15	230.91	-2.51	2.19	11.10	0.00	0.06
Fund11	1.02	0.12	0.57	3.98	0.76	1.49	106.32	2.99	2.57	15.55	0.00	0.05
Fund12	1.01	0.03	0.69	12.28	0.87	1.18	429.18	3.62	24.32	604.71	0.00	0.10
Fund13	1.01	0.03	0.00	3.09	0.94	1.08	479.48	-0.02	0.23	0.05	0.97	0.05
Fund14	1.00	0.02	0.06	2.88	0.96	1.04	800.75	0.34	-0.32	0.21	0.90	0.03
Fund15	1.01	0.04	-0.65	4.13	0.85	1.10	332.23	-3.41	2.97	20.45	0.00	0.08
Fund16	1.01	0.04	-0.54	4.90	0.85	1.14	294.80	-2.85	4.99	33.04	0.00	0.07

Panel D: Event-Driven funds

Funds	Mean	SD	Skewness	Kurtosis	Min	Max	Mean t-statistic	Skewness t-statistic	Kurtosis t-statistic	JB Test	P-value P-value	Lilliefors Test
Fund1	1.01	0.05	0.18	6.50	0.80	1.25	242.76	0.97	9.31	87.55	0.00	0.10
Fund2	1.01	0.02	-2.10	11.61	0.92	1.04	838.09	-11.19	22.91	650.20	0.00	0.12
Fund3	1.01	0.01	0.52	6.52	0.98	1.05	1509.18	2.77	9.37	95.41	0.00	0.14
Fund4	1.01	0.01	-0.29	4.37	0.95	1.05	924.25	-1.54	3.65	15.73	0.00	0.07
Fund5	1.01	0.09	1.26	11.68	0.65	1.51	147.33	6.72	23.11	579.21	0.00	0.18
Fund6	1.01	0.02	0.60	5.16	0.93	1.10	553.59	3.17	5.75	43.16	0.00	0.08
Fund7	1.01	0.03	-1.46	8.70	0.86	1.06	515.28	-7.76	15.17	290.37	0.00	0.08
Fund8	1.01	0.03	-0.61	4.47	0.89	1.08	446.41	-3.23	3.91	25.76	0.00	0.04
Fund9	1.01	0.01	-0.43	4.56	0.95	1.05	919.68	-2.29	4.15	22.47	0.00	0.09
Fund10	1.01	0.03	-1.52	8.88	0.86	1.06	513.78	-8.07	15.65	309.90	0.00	0.09
Fund11	1.01	0.01	-0.18	3.43	0.97	1.05	1005.29	-0.95	1.14	2.20	0.33	0.04
Fund12	1.00	0.04	-0.48	5.58	0.86	1.12	340.37	-2.54	6.87	53.59	0.00	0.09
Fund13	1.01	0.04	0.95	8.47	0.86	1.24	296.82	5.07	14.55	237.26	0.00	0.10
Fund14	1.01	0.01	-1.03	6.29	0.97	1.03	1562.35	-5.51	8.77	107.21	0.00	0.08
Fund15	1.01	0.01	-1.17	7.46	0.93	1.04	899.26	-6.22	11.88	179.90	0.00	0.07
Fund16	1.01	0.04	-0.55	5.79	0.83	1.13	332.35	-2.91	7.43	63.60	0.00	0.08

Panel E: Relative Value funds												
Funds	Mean	SD	Skewness	Kurtosis	Min	Max	Mean t-statistic	Skewness t-statistic	Kurtosis t-statistic	JB Test	P-value P-value	Lilliefors Test
Fund1	1.02	0.08	-0.97	27.75	0.40	1.50	160.44	-5.20	66.25	4415.64	0.00	0.23
Fund2	1.01	0.01	-0.18	6.06	0.96	1.06	1066.13	-0.96	8.20	68.22	0.00	0.09
Fund3	1.01	0.00	3.04	15.92	1.00	1.03	3599.15	16.26	34.58	1460.35	0.00	0.24
Fund4	1.01	0.01	-0.74	4.48	0.98	1.02	1896.05	-3.99	3.96	31.57	0.00	0.09
Fund5	1.01	0.03	-5.02	42.48	0.75	1.10	439.97	-26.87	105.69	11892.48	0.00	0.22
Fund6	1.01	0.01	-0.40	5.74	0.95	1.05	1005.59	-2.12	7.34	58.32	0.00	0.12
Fund7	1.01	0.01	-0.90	5.94	0.96	1.03	1458.86	-4.84	7.86	85.28	0.00	0.08
Fund8	1.01	0.00	1.19	6.32	1.00	1.02	6153.22	6.35	8.88	119.17	0.00	0.09
Fund9	1.00	0.02	-0.58	5.92	0.89	1.07	561.73	-3.12	7.81	70.76	0.00	0.05
Fund10	1.01	0.01	-0.21	5.69	0.98	1.04	1701.42	-1.15	7.19	52.99	0.00	0.08
Fund11	1.01	0.02	-1.07	8.38	0.94	1.06	858.61	-5.70	14.41	240.27	0.00	0.11
Fund12	1.01	0.01	0.06	4.41	0.96	1.05	1025.82	0.34	3.79	14.45	0.00	0.07
Fund13	1.01	0.01	0.06	4.66	0.96	1.05	1021.13	0.34	4.44	19.87	0.00	0.08
Fund14	1.01	0.02	-3.01	20.21	0.90	1.04	878.87	-16.14	46.07	2382.68	0.00	0.19
Fund15	1.00	0.00	-0.53	5.92	0.99	1.02	3321.75	-2.82	7.83	69.21	0.00	0.07
Fund16	1.00	0.01	-0.48	3.78	0.98	1.02	2035.89	-2.58	2.08	10.98	0.00	0.07

Table 3.3 – *Ljung-Box statistics for individual HFR hedge fund returns and squared returns*

Panel A: Global Macro funds				
	$\rho(r)$	P-value	$\rho(r^2)$	P-value
Fund1	14.12	0.17	42.17	0.00
Fund2	9.50	0.49	22.58	0.01
Fund3	10.80	0.37	81.07	0.00
Fund4	23.66	0.01	13.25	0.21
Fund5	12.51	0.25	65.55	0.00
Fund6	16.98	0.07	15.55	0.11
Fund7	24.05	0.01	22.54	0.01
Fund8	7.46	0.68	8.07	0.62
Fund9	26.36	0.00	14.31	0.16
Fund10	12.98	0.23	3.82	0.96
Fund11	7.29	0.70	24.92	0.01
Fund12	13.50	0.20	3.08	0.98
Fund13	15.11	0.13	22.52	0.01
Fund14	10.25	0.42	2.94	0.98
Fund15	17.02	0.07	20.25	0.03
Fund16	17.89	0.06	45.59	0.00

Panel B: Barclay CTAs funds				
	$\rho(r)$	P-value	$\rho(r^2)$	P-value
Fund1	20.63	0.02	30.33	0.00
Fund2	29.21	0.00	38.47	0.00
Fund3	5.54	0.85	9.58	0.48
Fund4	9.59	0.48	10.84	0.37
Fund5	19.82	0.03	9.13	0.52
Fund6	20.26	0.03	7.68	0.66
Fund7	8.56	0.57	2.07	1.00
Fund8	62.78	0.00	42.40	0.00
Fund9	10.54	0.39	21.16	0.02
Fund10	21.65	0.02	27.30	0.00
Fund11	11.29	0.34	1.99	1.00
Fund12	14.74	0.14	43.16	0.00
Fund13	26.36	0.00	30.12	0.00
Fund14	14.49	0.15	20.65	0.02
Fund15	14.04	0.17	22.88	0.01
Fund16	25.30	0.00	13.89	0.18

Panel C: Equity Hedge funds				
	$\rho(r)$	P-value	$\rho(r^2)$	P-value
Fund1	16.40	0.09	9.52	0.48
Fund2	13.26	0.21	72.61	0.00
Fund3	5.53	0.85	14.45	0.15
Fund4	25.58	0.00	42.44	0.00
Fund5	3.91	0.95	18.34	0.05
Fund6	22.92	0.01	15.15	0.13
Fund7	22.74	0.01	39.76	0.00
Fund8	25.16	0.01	37.88	0.00
Fund9	14.70	0.14	1.47	1.00
Fund10	6.06	0.81	23.24	0.01
Fund11	13.33	0.21	54.29	0.00
Fund12	5.64	0.84	21.55	0.02
Fund13	13.86	0.18	22.08	0.01
Fund14	6.16	0.80	13.87	0.18
Fund15	5.78	0.83	16.62	0.08
Fund16	14.50	0.15	37.12	0.00

Panel D: Event-Driven funds				
	$\rho(r)$	P-value	$\rho(r^2)$	P-value
Fund1	22.38	0.01	15.55	0.11
Fund2	24.26	0.01	21.71	0.02
Fund3	53.03	0.00	53.47	0.00
Fund4	35.94	0.00	22.45	0.01
Fund5	16.03	0.10	82.44	0.00
Fund6	38.91	0.00	22.12	0.01
Fund7	30.86	0.00	12.24	0.27
Fund8	33.44	0.00	14.79	0.14
Fund9	42.42	0.00	11.08	0.35
Fund10	29.31	0.00	10.94	0.36
Fund11	24.86	0.01	9.74	0.46
Fund12	17.66	0.06	25.05	0.01
Fund13	19.01	0.04	31.39	0.00
Fund14	41.17	0.00	11.56	0.32
Fund15	17.67	0.06	28.91	0.00
Fund16	10.76	0.38	28.69	0.00

Panel E: Relative Value funds				
	$\rho(r)$	P-value	$\rho(r^2)$	P-value
Fund1	12.99	0.22	30.11	0.00
Fund2	36.91	0.00	32.87	0.00
Fund3	9.62	0.47	9.17	0.52
Fund4	44.99	0.00	10.51	0.40
Fund5	61.25	0.00	36.42	0.00
Fund6	21.20	0.02	111.10	0.00
Fund7	63.55	0.00	6.50	0.77
Fund8	506.62	0.00	207.20	0.00
Fund9	31.02	0.00	24.61	0.01
Fund10	15.21	0.12	29.15	0.00
Fund11	40.45	0.00	54.56	0.00
Fund12	45.96	0.00	97.36	0.00
Fund13	42.25	0.00	87.17	0.00
Fund14	35.97	0.00	26.22	0.00
Fund15	79.59	0.00	67.37	0.00
Fund16	11.79	0.30	12.35	0.26

Table 3.4 – *Annualized Portfolio Performance Statistics*

Panel A: Global Macro funds

$K = 7,\ \&\ L = 10$

Strategy	λ	AAY(%)	ASD(%)	ASR	mSR	$\mathcal{S}$(%)	tc^{be}
I^K		28.58	23.86	1.20			
	5	23.44	11.07	2.12	0.41	52.26	0.46
MV^K	10	20.45	9.01	2.27	0.55	50.32	0.43
	15	18.66	8.14	2.29	0.64	49.68	0.45
	5	14.01	10.56	1.32	1.10	56.13	-0.33
GO-GARCH(1,1)	10	13.87	10.50	1.32	1.11	56.13	-0.38
	15	13.83	10.48	1.31	1.11	55.48	-0.39
EWCRP		13.67	11.72	1.16	1.24	58.06	0.16

Panel B: Barclay CTAs funds

$K = 7,\ \&\ L = 10$

Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
I^K		40.80	45.84	0.89			
	5	25.42	12.71	2.00	0.47	52.53	-3.36
MV^K	10	21.69	10.04	2.16	0.54	51.90	-4.40
	15	19.27	8.90	2.16	0.59	51.27	-4.58
	5	13.52	11.94	1.13	0.81	59.49	2.26
GO-GARCH(1,1)	10	13.36	11.85	1.12	0.81	59.49	2.07
	15	13.31	11.82	1.12	0.81	59.49	2.04
EWCRP		13.16	11.05	1.19	0.80	59.49	0.40

Panel C: Equity Hedge funds

$k = 7,\ \&\ L = 10$

Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
I^K		29.33	19.31	1.52			
	5	25.52	12.21	2.09	0.54	53.42	0.72
MV^K	10	22.39	9.63	2.33	0.72	54.79	0.80
	15	20.46	8.32	2.46	0.81	56.85	0.84
	5	11.93	7.81	1.52	1.53	65.75	-4.23
GO-GARCH(1,1)	10	11.81	7.78	1.51	1.53	65.75	-12.73
	15	11.77	7.76	1.51	1.53	65.75	-24.22
EWCRP		13.45	10.72	1.25	1.86	65.75	0.17

Panel D: Event-Driven funds							
$K = 7, \ \& \ L = 10$							
Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
I^K		23.49	13.77	1.68			
	5	20.14	9.04	2.78	0.71	47.02	2.56
MV^K	10	18.85	7.72	3.17	0.77	50.99	0.77
	15	18.07	7.00	3.44	0.81	52.32	0.75
	5	13.22	10.43	1.26	3.09	63.58	-0.26
GO-GARCH(1,1)	10	13.15	10.42	1.26	3.10	63.58	-0.27
	15	13.13	10.41	1.26	3.11	63.58	-0.27
EWCRP		12.72	6.35	1.99	1.46	63.58	0.12

Panel E: Relative Value funds							
$K = 7, \ \& \ L = 10$							
Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
I^K		25.55	23.29	1.10			
	5	21.57	10.03	2.15	0.30	52.29	-0.50
MV^K	10	19.47	6.71	2.90	0.37	54.90	-1.04
	15	18.57	5.47	3.39	0.39	55.56	-1.20
	5	10.99	4.12	2.65	0.76	75.82	-0.24
GO-GARCH(1,1)	10	10.89	4.07	2.66	0.77	75.82	-0.24
	15	10.85	4.05	2.66	0.77	75.82	-0.24
EWCRP		11.03	2.76	3.98	0.71	71.90	0.22

Table 3.5 – *Hedge Fund Redemption Restrictions and Annualized Portfolio Performance Statistics*

Panel A: Global Macro funds

$K = 7,\ \&\ L = 10$

Strategy	λ	AAY(%)	ASD(%)	ASR	mSR	$\mathcal{S}$(%)	tc^{be}
$\mathbf{I}^K$		29.07	23.89	1.22			
	5	21.20	11.83	1.79	0.66	50.35	1.12
$\mathbf{MV}^K$	10	18.76	9.75	1.92	0.74	50.35	1.12
	15	17.22	8.90	1.93	0.80	51.05	1.17
	5	13.93	10.49	1.32	1.14	50.35	-1.57
GO-GARCH(1,1)	10	13.87	10.47	1.32	1.14	50.35	-1.61
	15	13.85	10.46	1.32	1.14	50.35	-1.60
EWCRP		14.09	11.94	1.18	1.26	48.25	0.48

Panel B: Barclay CTAs funds

$K = 7,\ \&\ L = 10$

Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
$\mathbf{I}^K$		38.10	44.23	0.86			
	5	23.31	14.37	1.62	0.50	49.65	38.26
$\mathbf{MV}^K$	10	19.77	11.32	1.75	0.56	52.45	15.82
	15	17.59	9.84	1.79	0.60	55.24	13.99
	5	14.40	12.06	1.19	0.74	53.85	2.24
GO-GARCH(1,1)	10	14.27	11.98	1.19	0.75	53.85	2.26
	15	14.22	11.96	1.19	0.75	53.85	2.27
EWCRP		13.67	11.21	1.22	0.75	53.15	0.98

Panel C: Equity Hedge funds

$k = 7,\ \&\ L = 10$

Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
$\mathbf{I}^K$		31.19	19.26	1.62			
	5	25.10	12.57	2.00	0.92	57.14	1.61
$\mathbf{MV}^K$	10	21.54	9.91	2.18	1.04	57.14	1.82
	15	19.42	8.61	2.26	1.11	56.39	1.83
	5	12.32	8.03	1.53	1.70	59.40	6.02
GO-GARCH(1,1)	10	12.32	8.02	1.53	1.69	58.65	5.96
	15	12.33	8.02	1.53	1.69	58.65	6.11
EWCRP		14.09	11.09	1.27	2.11	56.39	0.56

Panel D: Event-Driven funds								
$K = 7,$ & $L = 10$								
Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
I^K		20.90	13.05	1.60			
	5	18.24	7.76	2.35	0.51	47.83	2.93
MV^K	10	17.00	6.30	2.70	0.58	53.62	2.23
	15	16.30	5.51	2.96	0.61	53.62	1.74
	5	12.83	10.93	1.17	3.84	54.35	-0.68
GO-GARCH(1,1)	10	12.81	10.90	1.17	3.79	54.35	-0.69
	15	12.81	10.89	1.17	3.78	54.35	-0.69
EWCRP		12.71	6.55	1.93	1.27	55.07	0.29

Panel E: Relative Value funds							
$K = 7,$ & $L = 10$							
Strategy	λ	AAY(%)	ASD(%)	SR	mSR	$\mathcal{S}$(%)	tc^{be}
I^K		21.44	26.06	0.82			
	5	18.26	13.33	1.37	0.25	53.62	-5.68
MV^K	10	17.62	8.44	2.09	0.22	55.80	9.67
	15	17.09	6.63	2.58	0.23	54.35	5.28
	5	11.29	4.44	2.53	0.47	68.84	-0.50
GO-GARCH(1,1)	10	11.23	4.36	2.56	0.47	68.84	-0.51
	15	11.20	4.33	2.57	0.48	68.84	-0.51
EWCRP		11.23	2.88	3.88	0.44	67.39	0.48

Table 3.6 – *Robustness assessment of the fund selection process. We test the effect of drawing various samples of funds, on the dominance relationships between our four investment strategies. The investigation is addressed by drawing* 150 *times, a set of* 16 *HFR Equity Hedge funds from a sample of* 97*. P-values (in brackets) are computed at* 95% *confidence level. Small P-values indicate that the distribution free allocation strategy performs better than the optimizers listed in column* 1 *according to the Annualized average yields.*

Strategy	**t-statistics**	**M-W tests**
MV^K	**13.4**	**10714**
	(1.125658e-27)	(1.309928e-21)
EWCRP	**39.04**	**11325**
	(1.948257e-80)	(1.161420e-26)
GO-GARCH(1,1)	**43.44**	**11325**
	(8.481405e-87)	(1.161420e-26)

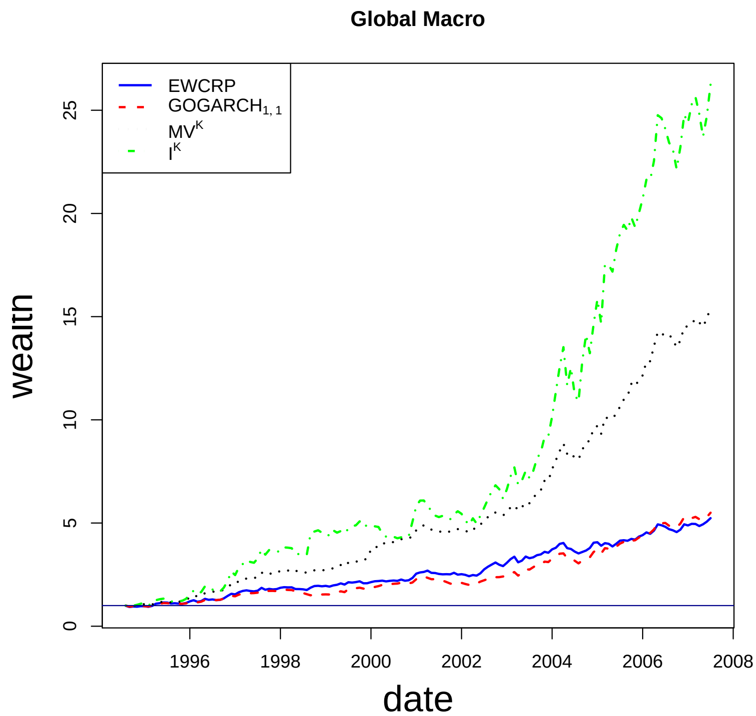

Figure 3.1 – *Wealth achieved by each investment strategy amongst* 16 *HFR Global Macro funds.*

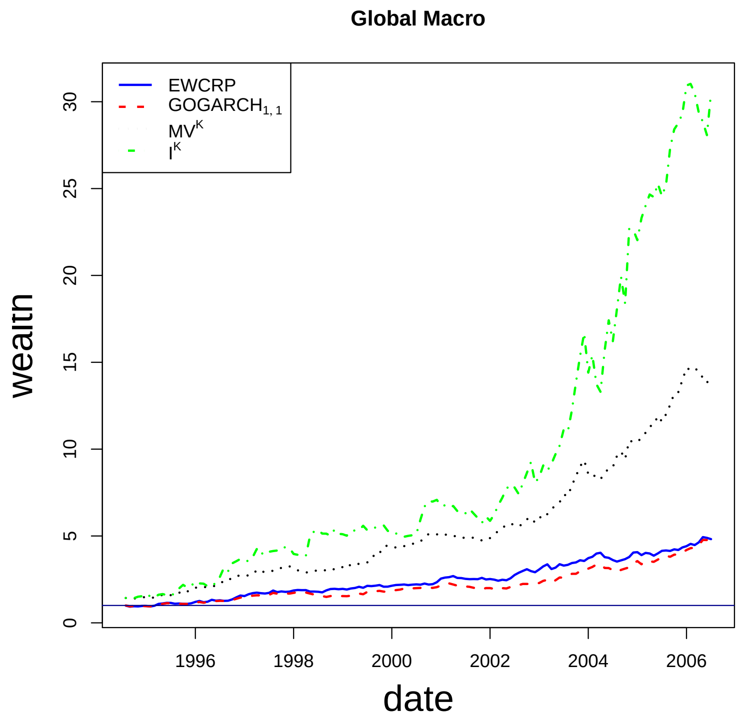

Figure 3.2 – *Wealth achieved by each investment strategy amongst* 16 *HFR Global Macro funds. Experiments are implemented under various hedge fund redemption restrictions.*

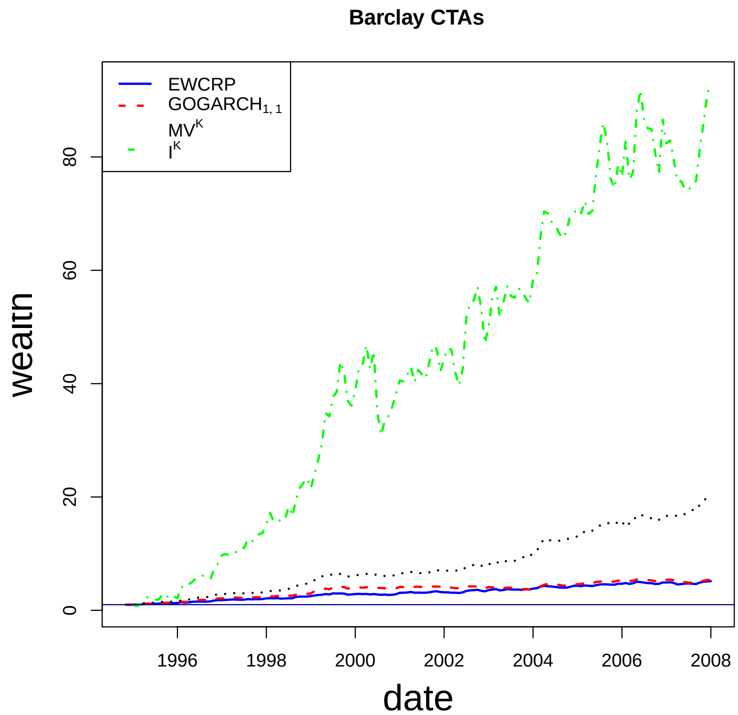

Figure 3.3 – *Wealth achieved by each investment strategy amongst 16 Barclays CTAs funds.*

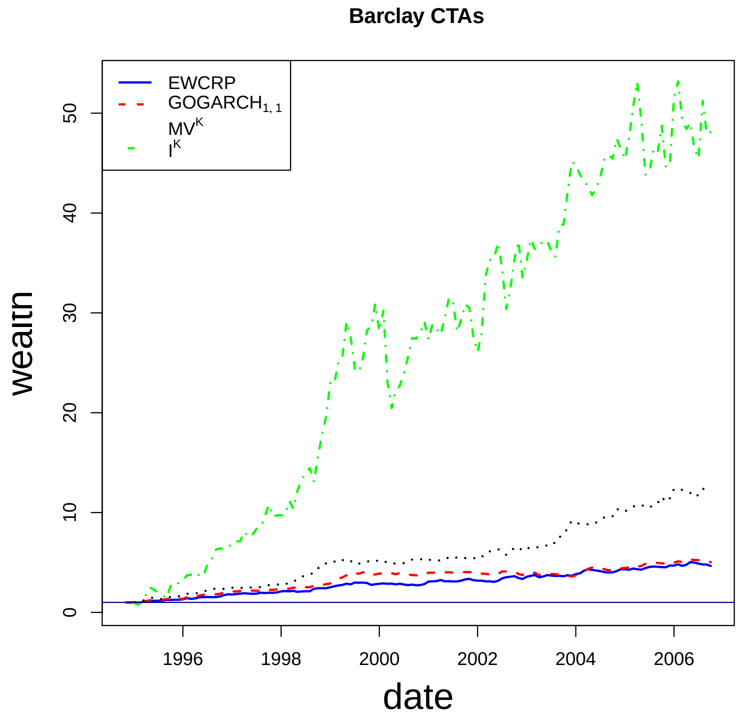

Figure 3.4 – *Wealth achieved by each investment strategy amongst 16 Barclays CTAs funds. Experiments are implemented under various hedge fund redemption restrictions.*

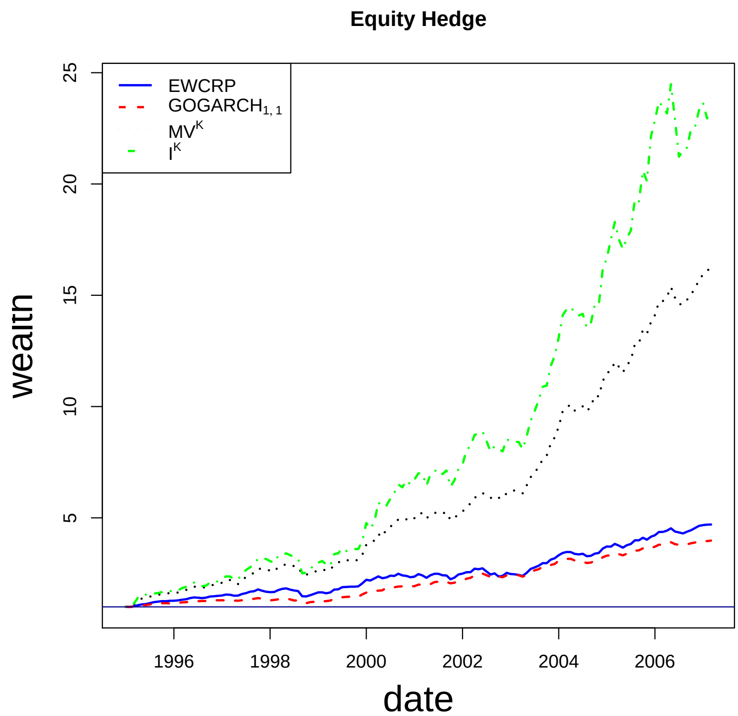

Figure 3.5 – *Wealth achieved by each investment strategy amongst* 16 *HFR Equity Hedge funds.*

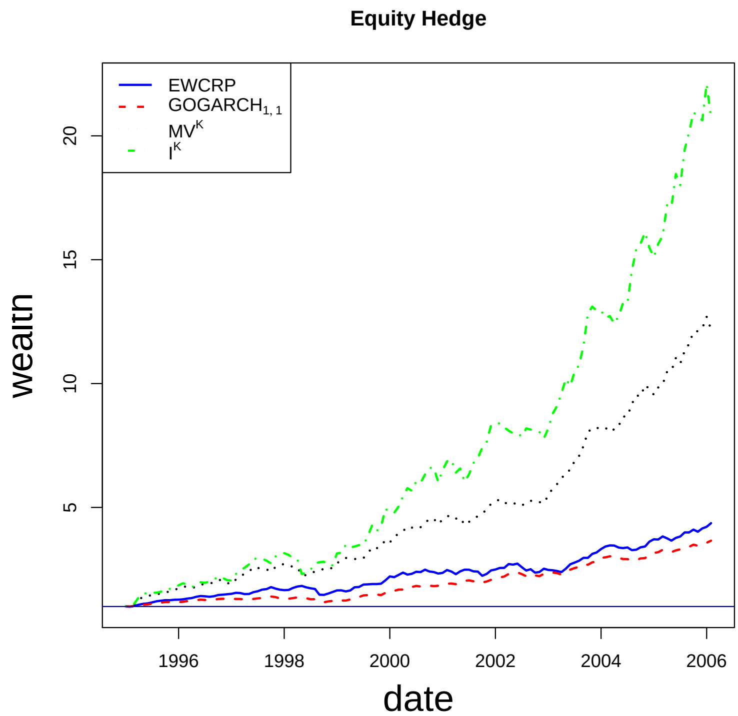

Figure 3.6 – *Wealth achieved by each investment strategy amongst* 16 *HFR Equity Hedge funds. Experiments are implemented under various hedge fund redemption restrictions.*

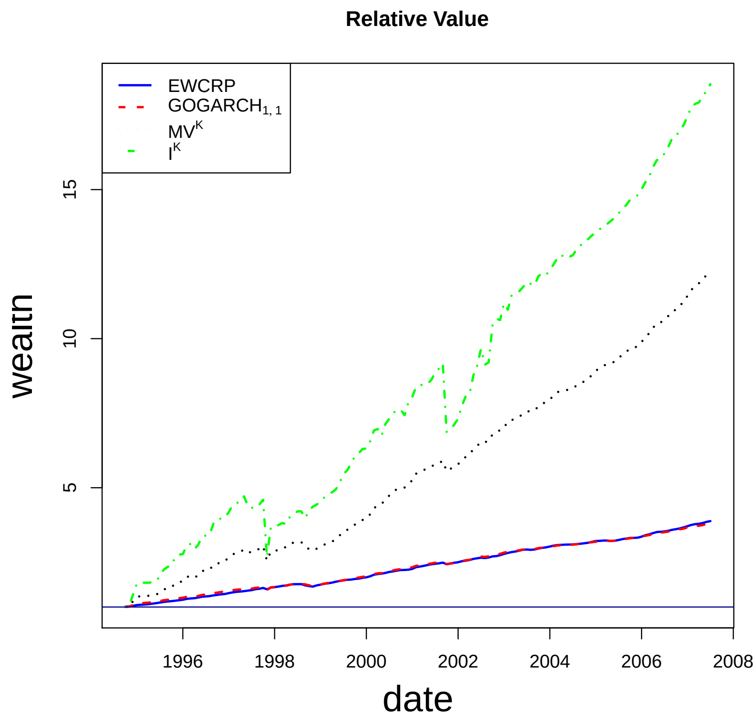

Figure 3.7 – *Wealth achieved by each investment strategy amongst 16 HFR Relative Value funds.*

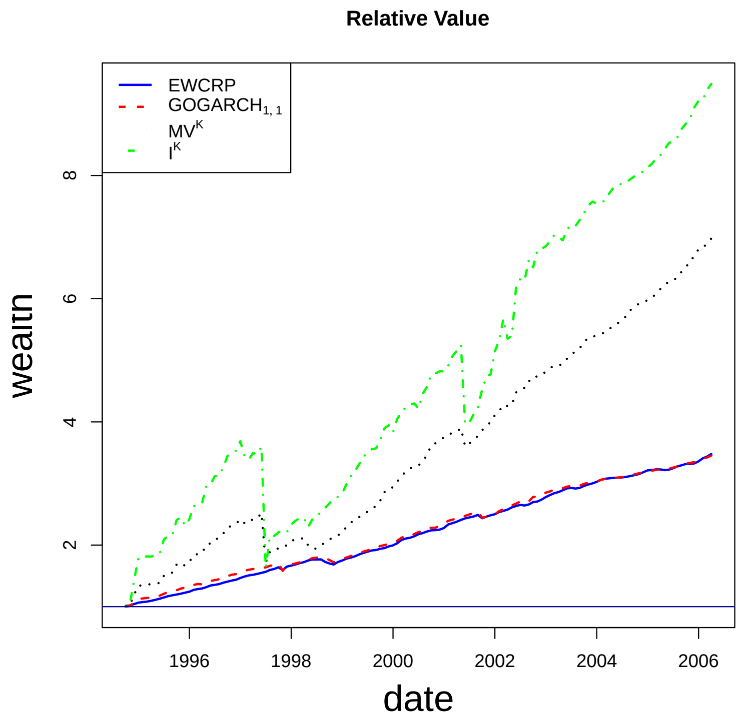

Figure 3.8 – *Wealth achieved by each investment strategy amongst 16 HFR Relative Value funds. Experiments are implemented under various hedge fund redemption restrictions.*

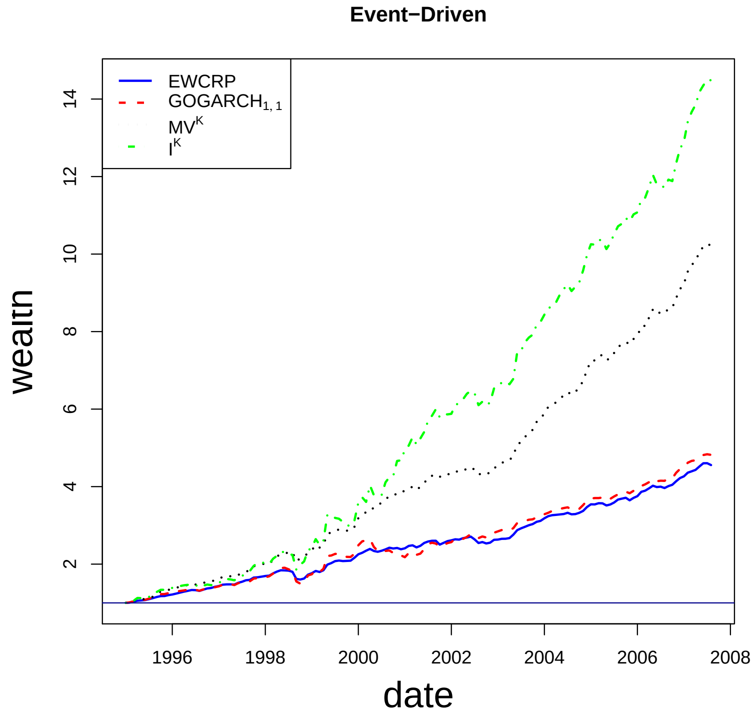

Figure 3.9 – *Wealth achieved by each investment strategy amongst 16 HFR Event-Driven funds.*

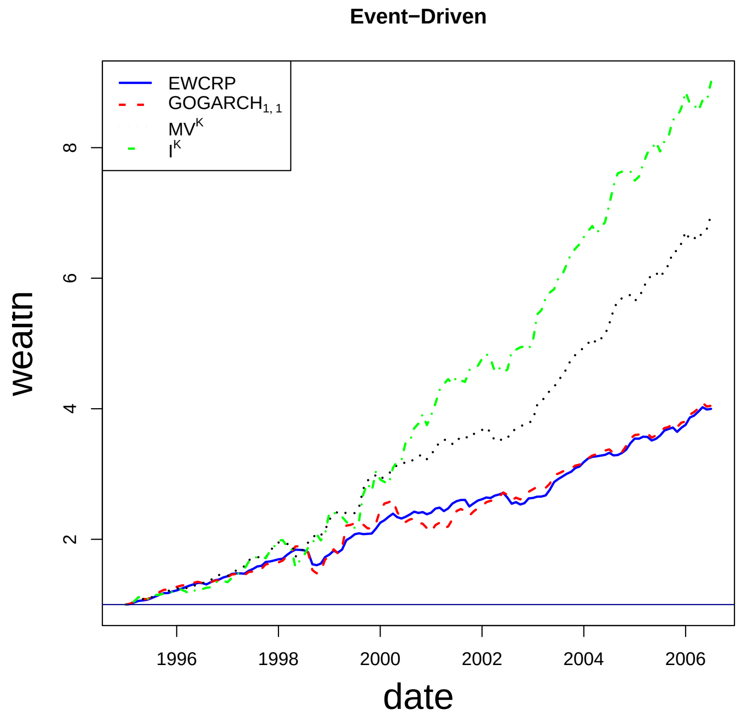

Figure 3.10 – *Wealth achieved by each investment strategy amongst 16 HFR Event-Driven funds. Experiments are implemented under various hedge fund redemption restrictions.*

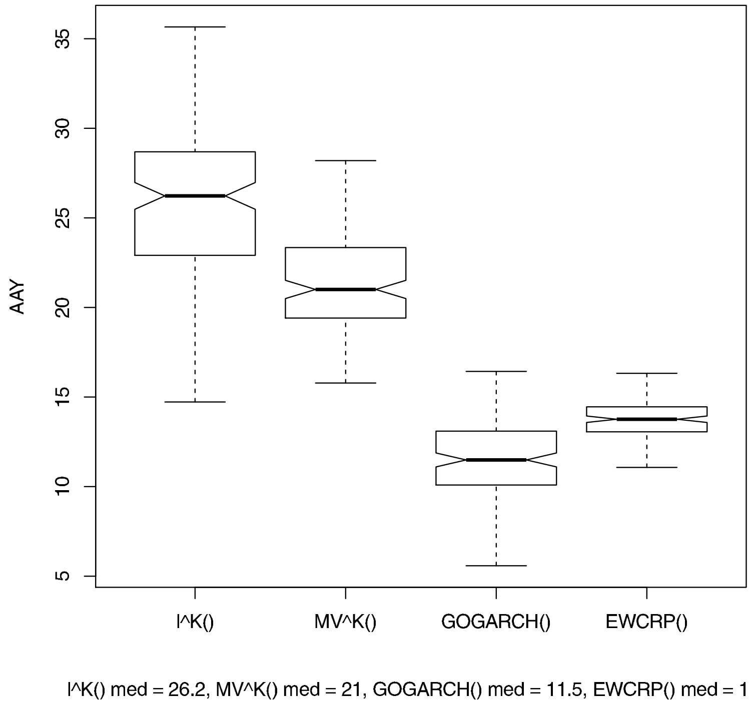

Figure 3.11 – *Boxplots of AAY for* 150 *resamplings HFR individual Equity Hedge funds. Median values for each strategy are displayed in the subtitle below the graphic.*

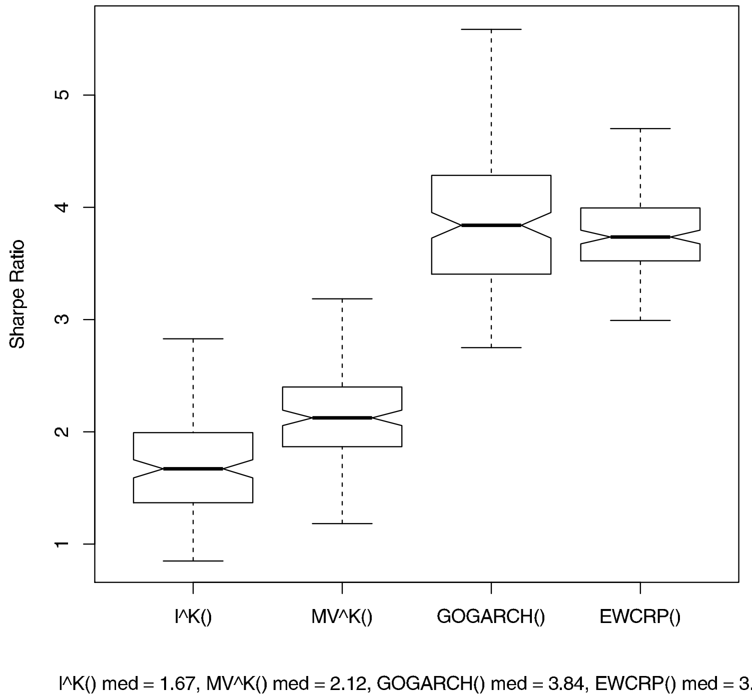

Figure 3.12 – *Boxplots of annualized Sharpe ratios for* 150 *resamplings HFR individual Equity Hedge funds. Median values for each strategy are displayed in the subtitle below the graphic.*

4 Selection of Balanced Portfolios to Track the Main Properties of a Large Market

Abstract

Index-based investment products are becoming increasingly popular among passive managers. So far, empirical studies have focused on complex heuristic-related optimization techniques. In this article, we adopt a different perspective and apply a survey sampling framework in the context of stock market tracking. We describe a novel and automatic method that enables us to construct a small portfolio to track the total market capitalization (TMC). The constructed portfolio is randomly selected using a new method of balanced sampling. Empirical studies are performed on the constituents of S&P500. Our findings suggest that balanced sampling portfolios efficiently track the S&P500 TMC.[1]

Keywords: Market Tracking, Survey sampling, Balanced sampling, Capitalization, Heuristic

4.1 Introduction

In 2010, the Exchange-Traded Funds (ETFs) community celebrated the 20th anniversary of the first ETF. Launched on 9 March 1990, the first ETF was listed on the Toronto Stock Exchange (TSX). Called the Toronto 35 Index Participation Fund (TIP), the goal of this fund was to provide the investment community with a passively managed instrument that could track the TSX 35 Index. The exponential growth of the sector is closely related to the structural shift of investors towards index-based allocation.

[1]This chapter is a reprint of: Tafin Djoko and Tillé (2013). Selection of Balanced Portfolios to Track the Main Properties of a Large Market. Accepted for publication in Quantitative Finance.

With index-related derivative products, index funds and ETFs, the effects of indexing in the financial industry have been deep and permanent.[2]

According to various sources in the performance evaluation literature, it seems difficult for actively managed funds to outperform the market over a certain time period. As wisely pointed by Sharpe (1991): "Properly measured, the average actively managed dollar must underperform the average passively managed dollar, net of costs. Empirical analyses that appear to refute this principle are guilty of improper measurement". Along the same line, Malkiel (1995) examining equities mutual fund performance from 1971 to 1991, found no consistent evidence that active management provides abnormal returns to investors. The authors concluded that: "most investors would be considerably better off by purchasing a low expense index fund, than by trying to select an active fund manager".

Barras, Scaillet, and Wermers (2010) draws similar conclusions. They investigated a database of 2,100 funds from 1975 to 2006 and estimated that only 0.6% of funds exhibited positive alpha (statistically not different from zero). That means 99.4% of funds revealed zero or negative alpha. More recently, Fugazza, Guidolin, and Nicodano (2010) discuss why and how the result of DeMiguel and Uppal (2009) that a passive equally weighted portfolio may easily outperform active ones, as it all depends on the type of activism, the utility function, and the investment horizon. Kirby and Ostdiek (2009), Kirby and Ostdiek (2012) and Tu and Zhou (2011) advocate that active strategies may outperform passive ones. However, Investments in index funds[3] and/or ETFs that track the broad market have turned out to be widely popular. In this context, the fund manager has to replicate the target index as closely as possible. This is the well-known Index Tracking (IT) problem.

Unfortunately, it is not possible to directly invest in indexes. There exist three main IT methods: Total Clone (TC), Synthetic Clone (SC), and Partial Clone (PC). The TC method is the least appealing and is technically infeasible. Under this framework, all the underlying constituents of the index have to be purchased with the appropriate weights. Although the strategy guarantees exact replication of the target index, it also implies high transaction costs and requires a very large amount of capital to achieve the investment characteristics. The second popular IT method is the SC. The index fund or ETF managers hold[4] a subset of the index and obtain full replication via a swap contract with compensation. At the

[2] According to The Economist, January 26 (2013), the ETFs industry has enjoyed an average annual growth rate of 29.6% over the last decade. By January 2013, the magazine estimated that the global exchange-traded funds total assets under management reached about US $2 trillion.

[3] In the 2013 Investment Company Institute fact book, the association estimated that US index funds managed total net assets of US $1.3 trillion. Around 33% of assets were invested in S&P 500 index funds. See Reid, Collins, Holden, and Steenstra (2013)

[4] However, recent developments in the ETF industry, particularly in Europe, have promoted the so-called funded swap model. Under this structure, the ETF is not required to

end of the trading period, the performance of the tracking portfolio is exchanged for the performance of the index. The swap represents a limited proportion of the value of the portfolio ($< 10\%$). For the PC or sampling clone, the target index is reconstructed on the basis of a small sample of assets. The choice of selection algorithm is fundamental to insure an acceptable and stable deviation from the reference benchmark. For an investor, the main advantage is that the performance of the broad market index can be replicated with only a small number of assets. Therefore the amount of capital required and the management process associated with the investment is limited.

The issue of replicating a market using a subset of its constituents has been addressed by various researchers. In this respect, computational algorithms to address the tracking problems include continuous linear and quadratic optimization as well as mixed-integer linear (quadratic) optimization. Mixed-integer formulation is used if cardinality constraints are imposed. The cardinality constraint refers to the fact that the replicator can only use a limited number of n securities out of the market N ($n \ll N$) to track its benchmark. The approximate mixed-integer solutions are well-illustrated by Chen, Batson, and Dang (2010). Typically in the optimized sampling approach, the objective is to minimize a function measuring the distance between performance achieved by the target benchmark and that achieved by the replicating portfolio, using a small number of assets. The goal of the researcher is to construct a portfolio that is as close to the market as possible. Generally, the tracking quality of the developed algorithm is based on the tracking error (TE). The TE measures the time-varying difference between the returns of the tracking portfolio and the returns of the index. In the academic literature, TE is defined in several variants. Roßbach and Karlow (2011) studied the stability of various measures of tracking quality and listed three basic alternatives: the Mean Squared Error, the Mean Absolute Deviation and the Tracking Error Variance. While the first two measure respectively the mean of the square differences and the mean of the absolute differences between the returns over a given time period, the third one is expressed as the variance of performance differences. Along this line, Canakgoz and Beasley (2009) presented a mixed-integer linear programming formulation with cardinality constraints on the number of assets in the replicating portfolio, addressing both the issues of market tracking and enhancing.

Markowitz (1987) formulated the market tracking problem, also known as dimension reduction problem, as a mean-variance optimization, making fundamental assumptions on the distribution of the replicated benchmark. However the main difficulty with this approach is the computational burden associated with the estimation of the variance-

hold the basket of securities of the index used in the synthetic tracker. See Bioy, Johnson, and Rose (2011) for additional details.

covariance matrix. Derigs and Nickel (2003) extend the Markowitz-based formulations and derived a meta-heuristic decision support for portfolio management and optimization. Using a linear factor model based on macro-economic variables, they estimated the covariance matrix and then solved the tracking optimization problem.

Focardi and Fabozzi (2004) illustrated a time-series clustering approach. They argue that clustering might be an optimal methodology for building a sparse tracking portfolio when full replication of an aggregate basket of assets is not feasible. By suitably defining Euclidean distances between the time series of asset prices available in a given market, a hierarchical clustering "discovers" the correlation structure of the target benchmark. Francesco and Marcellino (2006) innovated by introducing the idea of using a dynamic linear factor-based model for the purpose of index tracking. The basic assumption here is that asset prices can be decomposed as in the Arbitrage Pricing Theory into the sum of either observable economic variables or "synthetic" variables. Hence the objective is to build a tracking portfolio with the same factor structure of the index. In a second step, the replica is refined in order to minimize the TE. Another econometric based approach was developed by Alexander and Dimitriu (2005b). They opt for a co-integration strategy to replicate market fluctuations. Co-integration is a statistical technique that seeks to exploit a long-term relationship between a portfolio and the target index, ensuring that the two are connected in the long term. They also show that when tracking becomes more difficult, a co-integration indexation strategy could enhance performance. Inspired by models of biological evolution, evolutionary algorithms iteratively apply genetic-based operations selection to a population of potential candidate solutions. In this respect, Maringer and Oyewumi (2007) implement a Differential Evolution (DE) technique for the IT problem. The basic idea of DE is to randomly generate an initial population of candidate solutions, then iteratively perform a subsidiary local search to produce a new solution.

Ruiz-Torrubiano and Suárez (2009) described a hybrid strategy that combines quadratic optimization with a genetic algorithm to solve the combinatorial issue of identifying the appropriate number of assets. The resulting approach is similar to the one investigated by Shapcott (1992). Both frameworks propose to separate handling of the problem of selecting the optimal subset of assets and the quadratic optimization that follows. They use the Random Assorting Recombination algorithm to perform asset set recombination, which can be appropriated to avoid premature convergence. However Ruiz-Torrubiano and Suárez (2009) empirical experiments show that the presence of highly correlated assets makes convergence of the algorithm difficult. Optimization routines tend to construct highly concentrated replicas. Indeed, if, in a given time period, a group of companies in the same sector are similar to the market index, the optimization program might then select these securities as constituents of the replicating portfolio. The resulting allocation will then undermine the primary property of indexing, diversification, and can lead to increased volatility and risk for a risk-adverse investor or passive manager who is interested in diversified investment strategies. The diversification reduction issue is a direct consequence of the dimension reduction objective of the IT problem, and it is very difficult to preserve the diversification benefit inherent to index investing with an optimization approach. On the other hand, heuristic methods can deal with the issue and provide a relatively good solution. However, the technique provides no guarantee of optimality.

In this article, we describe a novel automatic method for constructing a small portfolio in order to track the total capitalization of a large market of assets. This small portfolio is selected randomly by means of a new method of balanced sampling. In order not to be mistaken, the term "balanced" here does not allude to its financial connotation, but instead refers to the statistical sampling meaning. Statistically speaking, a sample is said to be balanced if the estimator of the mean matches or is approximately equal to the population mean. Hence, balanced sampling is implemented to construct a tracking portfolio whose capitalization is proportional, at any trading period, to the total market capitalization. By balancing the randomly selected replication portfolios on principal components, the proposed sampling design ensures that the trackers constantly reconstruct the time series evolution of the market, while preserving its diversification property. The diversification principle is automatically incorporated in the efficient balanced sampling procedure without imposing additional constraints .The performances of our strategy seem to support the fact that the constrained IT problem can be solved by a survey sampling approach.

This article makes several contributions. We move from a hard to solve complex optimization problem, to a simpler and more intuitive survey sampling framework. The issue of efficiently selecting a subset from a large population is a natural sampling topic. There are obviously various ways of randomly selecting a sample. While the use of sampling techniques was already proposed by Meade and Salkin (1989), we advocate a new method of balanced sampling with unequal selection probabilities. We show that this method can easily generate random portfolios that efficiently track the market and maintains the diversification property. In addition to the balanced sampling design, the article discusses three alternative random selection strategies and a simple heuristic for constructing a tracking portfolio consisting of the 50 stocks with the largest market capitalizations. One of the computational benefits of the method is the possibility to run a large number of simulations to evaluate the replication strategy without increasing the computational burden. Finally, we show that a balanced portfolio of assets can only be selected if its dimensionality is reduced. The restriction of the matrix of market capitalizations to a small number of components of the singular value decomposition is enough to approximately reconstruct the original array. Indeed the singular values rapidly decreases to zero.

The paper is structured as follows. We first introduce the basic concepts in survey sampling methods, define the notations and state the aim of the study in Section 4.2. In Section 4.3, we illustrate the option that enables us to derive the inclusion probabilities, which will be set proportional to asset capitalization. A computational approximation is proposed for large assets, for which an algorithm is developed in Section 4.4. A short review of survey sampling theory is introduced in Section 4.5. Dimensionality reduction through singular value decomposition is outlined in Section 4.6. In Section 4.7, we discuss the survey sampling approach used to address the IT problem. An empirical experiment performed on the S&P500 component stocks is illustrated. The experiment is tested on two different phases of the market. Further on, we compare the results of our sampling strategy with three alternative random sampling approaches and a simple heuristic involving the construction of a tracking portfolio consisting of the 50 stocks with the largest market capitalizations. A tentative conclusion is drawn in Section 4.8.

4.2 Background

4.2.1 Survey sampling

In survey sampling, one is in general concerned to study a variable of interest from a population. The population in most of the cases denoted by U. The units of the population can be households, businesses, forests. Each unit might be identified by a label. Let use the notation of Tillé (2006):

$$U = \{1, \ldots, i, \ldots, N\}.$$

The goal of the researcher is to estimate a parameter of the variable of interest y, taking value y_i for statistical unit i. The common target estimated parameters are the population total

$$Y = \sum_{i \in U} y_i,$$

and the variance

$$\sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{Y})^2$$

The beauty of sampling methods is to be able to estimate various characteristics on a population by selecting only a subset from the population, that is, sample. A sample does not has to be a 'clone' of the population, it is just an instrument to perform estimation of the parameters of interest without covering the full population, therefore, saving time, effort, and money.

In practice there are different ways to obtain a sample. Depending of the situation at hand, the procedure may be without or with replacement, statistical units can be selected wit equal or unequal probabilities. For what follow, the author will consider random sampling without replacement, which is widely preferred in practice. Let denote by $\mathbf{s}$, the set of all possible samples without replacement from the population U, I_i an indicator variable for the unit i, which is defined by

$$I_i = \begin{cases} 1 & \text{if unit } i \in \mathbf{s} \\ 0 & \text{if unit } i \notin \mathbf{s}, \end{cases}$$

and π_i be the probability of selecting unit i, with

$$\pi_i = \Pr(k \in \mathbf{s}).$$

The expectation of I_i is equal to π_i. If all the selection probabilities are positive, the Horvitz and Thompson (1952) estimator of the total given by

$$\widehat{Y} = \sum_{i \in U} \frac{y_i}{\pi_i} I_i \tag{4.1}$$

is unbiased for Y, since

$$\mathrm{E}(\widehat{Y}) = \mathrm{E}\left(\sum_{i \in U} \frac{y_i}{ß_i} I_i \right) = \sum_{i \in U} \frac{y_i}{ß_i} \mathrm{E}(I_i) = Y.$$

Along the approach proposed in this article to efficiently select a sample and estimate the character of interest, we will also investigate two additional sampling designs: simple random sampling without replacement designs (*srswor*) and random systematic (*rs*) method with unequal probability and fixed sample size.

Random systematic (*rs*) is a design of fixed size n in which for any inclusion probability $0 < \pi_i < 1$, $i \in U$, with $\sum_{i \in U} \pi_i = n$, we define $V_i = \sum_{l=1}^{i}$ for all $i \in U$, with $V_0 = 0$. Samples are then selected according to an exact method as described in Algorithm 2

Simple random sampling design without replacement (*srswor*) is also a design of fixed size in which the probability of selecting sample s from a population of size N is

Algorithm 2: Random systematic sampling

Generate a uniform random number $u \in [0, 1]$

- selected first unit i_1 if $V_{i_1 - 1} \leq u \leq V_{i_1}$,
- selected second unit i_2 if $V_{i_2 - 1} \leq u + 1 \leq V_{i_2}$ and
- selected jth unit i_j if $V_{i_j - 1} \leq u + j - 1 \leq V_{i_j}$.

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{if card}\{s\} = n \\ 0 & \text{otherwise .} \end{cases}$$

In the following, we will relax the notation outlined here to focus on the market tracking problem. In our study, statistical units are stock market capitalizations and the function of interest is the total market capitalization.

4.2.2 Notation and aim

Consider a market composed of N assets. Let C_{it} denotes the total capitalization of asset i at time t, for $t = 1, \ldots, T$. The total value of the market at time t is given by

$$C_t = \sum_{i=1}^{N} C_{it}. \tag{4.2}$$

We assume the usual hypothesis that any asset is infinitely divisible. Any fraction between $[0, 1]$ of each capitalization C_{it} can thus be selected in the portfolio.

The aim is to select a small portfolio of n assets at time τ that is a subset of the market portfolio. We would like the portfolio to track the market as closely as possible. To sum up, in our development, we will select the portfolio with inclusion probabilities proportional to the capitalization of the assets at the end of the estimation period. This methodology is very usual in survey sampling theory where the statistical units are commonly selected with selection probabilities proportional to their sizes (Nedyalkova and Tillé, 2008). Next, we will allocate the same amount to each asset in the tracking portfolio, except for units with a very large capitalization. This sampling strategy provides unbiased portfolios.

Let a_i be the amount of asset i in the portfolio, and I_i be the indicator variable that takes value 1 if asset i is selected in the portfolio and 0 otherwise. We thus have

$$\sum_{i=1}^{N} I_i = n. \tag{4.3}$$

Variables I_i are supposed to be random with a Bernoulli distribution. We suppose that π_i is the probability of selecting unit i in the portfolio. In the following lines, the capitalizations C_{it} and the amount a_i are not supposed to be random. The only source of randomness is I_i. The expectations and variances are then computed with respect to I_i.

If the portfolio is constructed at time τ, therefore its value is defined as

$$P_\tau = \sum_{i=1}^{N} I_i a_i,$$

and for any other t, the factor by which amount (a_i) invested in the i-th asset grows during the tracking period is given by $C_{it}/C_{i\tau} \geq 0$. Therefore, the portfolio value at time t is defined as

$$P_t = \sum_{i=1}^{N} a_i I_i \frac{C_{it}}{C_{i\tau}}.$$

If all the constituents within the portfolio have the same allocation, i.e. $a_i = a$, then

$$P_\tau = a \sum_{i=1}^{N} I_i = an,$$

and for any other time t

$$P_t = a \sum_{i=1}^{N} I_i \frac{C_{it}}{C_{i\tau}}. \tag{4.4}$$

The portfolio will be constructed randomly by means of a sampling technique. To satisfy our tracking objective at time τ, that is, to build a portfolio that is as close as possible to the complete market during the period $\tau - q, \ldots, \tau$, we would like

$$\frac{P_t}{P_\tau} \approx \frac{C_t}{C_\tau}, \text{ for all } t = \tau - q, \ldots, \tau. \tag{4.5}$$

Equations (4.5) imply q linear constraints on P_t. However, in the market tracking issue, the replicator is bound by the limited number of n securities out of the market N $(n \ll N)$ that can be selected to track the target benchmark. Since the number of observations q is larger than n, we cannot satisfy more constraints than the number of units in the tracking portfolio. For this reason, we should reduce the number of constraints that must be satisfied. In what follow, period $\tau - q$ to τ is referred to as the portfolio estimation period, while the period that follows is referred to as the investment period.

4.3 Probability of selecting an asset

The assets are selected randomly. Let π_i be the probability of selecting asset i. Thus, we can write $\mathrm{E}(I_i) = \pi_i$, and given Equation (4.3)

$$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} \mathrm{E}\left(I_i\right) = \mathrm{E}\left(\sum_{i=1}^{N} I_i\right) = \mathrm{E}(n) = n, \tag{4.6}$$

the sum of the selection probabilities must thus be equal to the number of assets in the tracking portfolio.

If all the units in the portfolio have the same allocation a at time τ, then a convenient alternative for choosing the inclusion probabilities is to construct the π_i proportional to the capitalization at time τ, that is,

$$\pi_i \propto C_{i\tau}\,, i = 1, \ldots, N.$$

To simplify the reasoning, assume that $C_{i\tau} \leq C_\tau/n$, for all $i = 1, \ldots, N$. This assumption will be relaxed in the next section. Given Equation (4.6), the selection probabilities can be computed by using

$$\pi_i = \frac{nC_{i\tau}}{C_\tau}, \tag{4.7}$$

In this case, in expectation, the portfolio value remains always proportional to the total market capitalization. Indeed, from Expressions (4.4) and (4.7), we have

$$\begin{aligned}
\mathrm{E}\left(P_t\right) &= \mathrm{E}\left(a\sum_{i=1}^{N} I_i \frac{C_{it}}{C_{i\tau}}\right) = a\sum_{i=1}^{N} \frac{C_{it}}{C_{i\tau}} \mathrm{E}\left(I_i\right) = a\sum_{i=1}^{N} \frac{C_{it}}{C_{i\tau}} \pi_i \\
&= a\sum_{i=1}^{N} \frac{C_{it}}{C_{i\tau}} \frac{nC_{i\tau}}{C_\tau} = \frac{na}{C_\tau} \sum_{i=1}^{N} C_{it} = \frac{na}{C_\tau} C_t.
\end{aligned}$$

where C_τ is defined as in Equation (4.2).

This reasoning proofs that when the assets are selected with selection probabilities proportional to their capitalization, then their allocation in the portfolio must be equal in order to have an unbiased portfolio under the sampling design. This strategy increases the likelihood that some stocks with large capitalizations will be selected. The randomness of the selection scheme ensures that all constituents are given the same opportunity to enter the tracking portfolio and not only those that are "close" to the market. The diversification is thus automatically ensured. This procedure also tends to avoid selection of a large set of stocks with a small capitalization. The aim is thus not only to track the total market capitalization but also to represent it in all its diversity.

However, a computational problem appears when one or more assets have a very large capitalization. A capitalization is said to be very large when

$$C_{i\tau} \geq \frac{C_\tau}{n}.$$

In this case, Expression (4.7) can be larger than 1. Therefore, units with very large market capitalizations must necessarily be included in the tracking portfolio, with an allocation bigger than other assets.

4.4 Computation of selection probabilities with very large capitalization assets

In order to process the issue related to large capitalization stocks, we propose the following simple Algorithm (3) to compute the selection probabilities.

Algorithm 3: Computation of the selection probabilities

- Let $U^- = \{1, \dots, N\}$ and $U^+ = \emptyset$.
- Repeat until U^+ is stable
  - Compute $\pi_i = \dfrac{[n - \text{card}(U^+)]\, C_{i\tau}}{\sum_{j \in U^-} C_{j\tau}}$, $i = 1, \dots, N$.
  - Compute $U^- = \{i | \pi_i < 1\}$ and $U^+ = \{i | \pi_i \geq 1\}$, where $\text{card}(U^+)$ is the cardinality or the number of elements of U^+.

The result of Algorithm (1) provides two sets of stocks U^- and U^+ such that:

- U^- is the set of assets with $\pi_i \in [0, 1]$ such that

$$\sum_{j \in U^-} \pi_i = n - \text{card}(U^+),$$

- U^+ is the set of stocks with π_i larger than 1.

The units in U^+ are automatically in the portfolio. A sample of assets is then selected only in U^-. The units in U^+ receive an allocation equal to $a\pi_i$, for all $i \in U^+$, while the units in U^- that are selected in the portfolio receive the same allocation a, for all $i \in U^-$. In the next sections, without loss of generality, we consider that $\pi_i < 1$ for all assets because the selection of units U^+ is not random. They are thus always in the portfolio.

For illustration, suppose that the capitalization $C_{i\tau}$ of $N = 10$ stocks are respectively given as 5, 10, 15, 20, 40, 40, 80, 100, 200, 1000. We would like to select $n = 5$ entities. At the first step of the algorithm $U^+ = \emptyset$, thus we compute the first values for $\pi_i = 5C_{i\tau} / \sum_{i=1}^{10} C_{i\tau}$

$$0.017, 0.033, 0.050, 0.066, 0.132, 0.132, 0.265, 0.331, 0.662, 3.311.$$

The 10^{th} unit has a value larger than 1, thus we set $U^- = \{1, \dots, 9\}$ and $U^+ = \{10\}$. Next we recompute the following selection probabilities according to $\pi_i = 4C_i / \sum_{i=1}^{9} C_{i\tau}$, which gives

$$0.039, 0.078, 0.118, 0.157, 0.314, 0.314, 0.627, 0.784, 1.569, 7.843.$$

Again, the 9^{th} unit has a value bigger that 1. Thus we set $U^- = \{1, \dots, 8\}$ and $U^+ = \{9, 10\}$. Therefore, the selection probabilities are calculated according to $\pi_i = 3C_i / \sum_{i=1}^{8} C_{i\tau}$, yielding

$$0.05, 0.1, 0.15, 0.19, 0.39, 0.39, 0.774, 0.97, 1.94, 9.68.$$

Finally, 3 assets amongst units 1 to 8 are randomly selected in the portfolio with their corresponding probabilities. If one of these units is selected, the allocated capital is equal to a. Units 9 and 10 are always included in the portfolio and the amounts invested are respectively equal to $1.94 \cdot a$ and $9.68 \cdot a$.

4.5 Balanced sampling

In this section, we present a set of results developed in the framework of survey sampling theory in order to provide a very efficient estimation of a total. Deville and Tillé (2004) have proposed a sampling algorithm, called the cube method, which enables selection of a sample from a register with approximately the same means in the sample as in the population for all the variables of the register. This sample can be selected with equal or unequal selection probabilities.

The aim is to estimate through a sample, a total, Y of the variable of interest y_i. The values y_i are not supposed to be random. The only source of randomness is the process of selecting the sample. A sample is a subset of the population that is selected randomly. Random samples can be selected by several sampling designs such as simple random, stratified, and balanced sampling. The statistical units can be selected with equal or unequal selection probabilities. If all the selection probabilities π_i are positive, the unbiased Horvitz and Thompson (1952) estimator for Y is defined in Equation (4.1).

In order to construct an efficient sampling design, one generally uses auxiliary information that is known for all the population units, for instance a register. Let $\mathbf{x}_i$ be a vector of $\mathbb{R}^p$ containing the p values taken by the p auxiliary variables on statistical unit i. The $\mathbf{x}_i$ are supposed to be known on all the units of the population. A sample is said to be approximately balanced on a set of auxiliary variables if

$$\sum_{i=1}^{N} \frac{\mathbf{x}_i}{\pi_i} I_i \approx \sum_{i=1}^{N} \mathbf{x}_i. \tag{4.8}$$

This means that the Horvitz-Thompson estimator of the total is very close to the true population total for all the variables that are known at the population level. Unfortunately, a sample can rarely be exactly balanced, because the selection of a sample is an integer number problem (a unit is selected or not in the sample). It is thus rarely possible to exactly satisfy the p balancing equations given in Expression (4.8). This is called the 'rounding problem'.

In this paper, our aim is to select a portfolio using this technique of balanced sampling. We will thus randomly select a subset of companies in such a way to reproduce as well as possible the sum of the capitalization of all the listed companies in a given stock market. The statistical units are thus the stock prices or market capitalizations of these companies, and the auxiliary information is the set of values taken by the stocks during a period. In order to construct the portfolio, we refer to the following result.

Result 4.1 *For any vector of selection probabilities* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_i, \ldots, \pi_N)^t$ *such that*

$$0 < \pi_i \leq 1 \text{ and } \sum_{i=1}^{N} \pi_i = n \in \mathbb{N}$$

then there is a random vector $\mathbf{s} = (s_1, \ldots, s_i, \ldots, s_N)^t \in \mathbb{R}_N$ *such that*

(i) $\displaystyle\sum_{i=1}^{N} \frac{\mathbf{x}_i}{\pi_i} s_i = \sum_{i=1}^{N} \mathbf{x}_i.$

(ii) $\mathrm{E}(s_i) = \pi_i$, *for* $i = 1, \ldots, N$,

(iii) $0 \leq s_i \leq 1$, *for all* $i = 1, \ldots, N$,

(iv) $\mathrm{card}\{I | 0 < s_i < 1\} \leq p$, *for all* $i = 1, \ldots, N$.

where $\mathrm{card}(A)$ *is the cardinality of set* A *and* p *is the dimension of the* $\mathbf{x}_i$.

The proof of Result 4.1 directly follows from Deville and Tillé (2004)'s Result 4.1 and Section 4 of the same paper, where an algorithm is proposed to generate a balanced sample. This algorithm is called the 'flight phase of the cube method'. A faster algorithm is proposed in Chauvet and Tillé (2006) and a complete development on the methods of balanced sampling is given in Chauvet and Tillé (2006). A generator of balanced vectors is implemented in the R language, in the 'sampling' package (function fastflightcube).

The interpretation of Result 4.1 is as follows. The strict equality of Expression (4.8) cannot be obtained by the I_i's that only take the values 0 and 1. Nevertheless, it is possible to obtain a strict equality given as Result 4.1 *(i)* if we accept that at most p values for the s_i are not equal to 0 or 1 but can take a value between 0 and 1. Consequently, the balancing equations will be more difficult to achieve if the number of auxiliary variables is large. In our financial case study, the number of auxiliary variables is the number of observations during the period used to construct the portfolio. This number can thus be very large. A dimensionality reduction technique is therefore necessary.

In the next sections, vector $\mathbf{s} = (s_1, \dots, s_i, \dots, s_N)^\top$ is called a balanced vector on the $N \times p$ matrix

$$\mathbf{A} = \left(\frac{\mathbf{x}_1}{\pi_1} \cdots \frac{\mathbf{x}_i}{\pi_i} \cdots \frac{\mathbf{x}_N}{\pi_N} \right)^\top$$

with probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_i, \dots, \pi_N)^\top$ and will be denoted by

$$\mathbf{s} \sim BV(\mathbf{A}, \boldsymbol{\pi}).$$

Random vector $\mathbf{s}$ is such that $\mathrm{E}(\mathbf{s}) = \boldsymbol{\pi}$ and the number of values in $\mathbf{s}$ that are not equal to 1 or 0 is less or equal to the number p of columns of $\mathbf{X}$. Result 1 *(i)* can then be written with a matrix notation:

$$\mathbf{A}^\top \mathbf{s} = \mathbf{A}^\top \boldsymbol{\pi}.$$

If the rows of matrix $\mathbf{A}$ correspond to the stocks, and the columns of matrix $\mathbf{A}$ correspond to the observations, random vector $\mathbf{s}$ is a portfolio that reproduces the characteristics of the assets described in matrix $\mathbf{A}$. Unfortunately, the number of observations is generally very large. Vector $\mathbf{s}$ will thus contain a very small number of integer values. In order to apply balanced sampling, we thus need to reduce the dimensionality of the data as described in Section 4.6.

4.6 Reduction of dimensionality

In order to construct the tracking portfolio, suppose we use the market capitalization values from time $\tau - q$ to time τ that is the estimation period. Let $\mathbf{C}$ denotes the matrix of capitalization for N assets:

$$\mathbf{C} = \begin{pmatrix} C_{1\tau-q} & \cdots & C_{i\tau-q} & \cdots & C_{N\tau-q} \\ \vdots & & \vdots & & \vdots \\ C_{1\tau} & \cdots & C_{i\tau} & \cdots & C_{N\tau} \end{pmatrix}$$

matrix $\mathbf{C}$ has dimension $(q+1) \times N$. The length of the estimation period $(q+1)$ is generally too large to apply a balanced sampling procedure. We will thus reduce the dimensionality of the problem.

Suppose we have a vector of selection probabilities whose components are strictly in the interval $(0,1)$ and whose sum is equal to $n \in \mathbb{N}$. This vector is computed in order to be proportional to the capitalization $C_{1\tau}, \cdots, C_{i\tau}, \cdots, C_{N\tau}$ during the last known time τ of the estimation period. Therefore, the reduction of dimensionality can be obtained as follows.

Firstly, we construct matrix $\mathbf{D} = (d_{jt})$ of dimension $N \times (q+1)$ where

$$d_{jt} = \frac{\pi_j}{n} \sum_{i=1}^{N} C_{it} = \frac{\pi_j C_t}{n}, \text{ for } t = \tau - q, \ldots, \tau, i = 1, \ldots, N.$$

The sums of the rows of matrix $\mathbf{D}$ are thus the same as those of matrix $\mathbf{C}$. The procedure continues by computing matrix $\mathbf{F} = \mathbf{C} - \mathbf{D}$, and deriving the singular value decomposition of $\mathbf{F}$, i.e.

$$\mathbf{F} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U} = (u_{\alpha t})$ is a $(q+1) \times (q+1)$ real matrix, $\boldsymbol{\Sigma}$ is a $(q+1) \times N$ diagonal matrix with nonnegative singular values σ_α on the diagonal that are ordered decreasingly, and $\mathbf{V}^\top = (v_{\alpha i})$ is the transposition of $\mathbf{V}$ that is an $N \times N$ real matrix. The capitalizations can then be decomposed as follows.

$$C_{it} = \frac{\pi_i}{n} C_t + \sum_{\alpha=1}^{\min(N,q+1)} \sigma_\alpha u_{\alpha t} v_{\alpha i}.$$

Finally, a restriction of the dimension can be obtained by applying the sum only of the first r components. One can construct a matrix

$$\mathbf{C}^{(r)} = \begin{pmatrix} C_{1\tau-q}^{(r)} & \cdots & C_{i\tau-q}^{(r)} & \cdots & C_{N\tau-q}^{(r)} \\ \vdots & & \vdots & & \vdots \\ C_{1\tau}^{(r)} & \cdots & C_{i\tau}^{(r)} & \cdots & C_{N\tau}^{(r)} \end{pmatrix},$$

where

$$C_{it}^{(r)} = \frac{\pi_i}{n} C_t + \sum_{\alpha=1}^{r} \sigma_\alpha u_{\alpha t} v_{\alpha i}.$$

A few dozen components of the singular value decomposition are enough to almost perfectly reproduce a matrix of capitalizations that contains several hundred rows and columns. Indeed the singular values σ_α very quickly decrease to zero.

Instead of selecting a balanced portfolio on matrix $\mathbf{C}$, we can thus use a matrix that contains $\boldsymbol{\pi}$ and the first r components of matrix $\mathbf{F}$

$$\mathbf{X}^{(r)} = \begin{pmatrix} \pi_1 & v_{11} & \cdots & v_{\alpha 1} & \cdots & v_{r1} \\ \vdots & \vdots & & \vdots & & \vdots \\ \pi_i & v_{1i} & \cdots & v_{\alpha i} & \cdots & v_{ri} \\ \vdots & \vdots & & \vdots & & \vdots \\ \pi_N & v_{1N} & \cdots & v_{\alpha N} & \cdots & v_{rN} \end{pmatrix}.$$

Next define matrix $\mathbf{A}^{(r)}$ by

$$\mathbf{A}^{(r)} = \mathrm{diag}^{-1}(\boldsymbol{\pi})\,\mathbf{X}^{(r)} = \begin{pmatrix} 1 & v_{11}/\pi_1 & \cdots & v_{\alpha 1}/\pi_1 & \cdots & v_{r1}/\pi_1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & v_{1i}/\pi_i & \cdots & v_{\alpha i}/\pi_i & \cdots & v_{ri}/\pi_i \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & v_{1N}/\pi_N & \cdots & v_{\alpha N}/\pi_N & \cdots & v_{rN}/\pi_N \end{pmatrix}$$

where $\mathrm{diag}(\boldsymbol{\pi})$ is the diagonal matrix containing vector $\boldsymbol{\pi}$ on its diagonal.

4.7 Selecting a balanced sampling portfolio

The purpose of this section is to discuss an approximation of the balanced sampling design. This is important in the context of our analysis and allows us to apply the theoretical results of our methodology to real financial data.

4.7.1 Implementation details

The main objective of Algorithm (4) is to enable the application of a balanced sampling framework to a limited number of dimensions. At the beginning of the algorithm, the method is applied to the largest possible dimension that is equal to the minimum between the total number of units minus one $(N-1)$, and the length of the estimation period q. At each iteration, the dimension is reduced in order to progressively round all the components of $\mathbf{s}$ to 0 or 1. Finally a portfolio of n assets is selected in such a way that it is as balanced as possible. In market tracking terminology, this means that the selected portfolio tracks the total market capitalization almost exactly.

Note that $\mathbf{s}_{(j)}$ is a martingale because $\mathrm{E}(\mathbf{s}^{(j+1)}|\mathbf{s}^{(j)}) = \mathbf{s}^{(j)}$. By using iterated conditioning, we obtain that $\mathrm{E}(\mathbf{s}^{(j)}) = \boldsymbol{\pi}$ for all $j = 1, \ldots, p$. Therefore, Algorithm (4) satisfies the given inclusion probabilities. At step 1, $\mathbf{s}^{(1)}$ contains p non integer values. At step 2, $\mathbf{s}^{(2)}$ contains $p-1$ non integer values. At step j, $\mathbf{s}^{(j)}$ contains $p-j$ non integer values. At step p, all the components of $\mathbf{s}^{(p)}$ are either equal to 0 or to 1. At the end of

Algorithm 4: Selecting the portfolio

- Initialize by setting $\mathbf{s}^{(0)} = \boldsymbol{\pi}$
- Define $p = \min(N-1, q)$
- For $j = 0$ to p do $\mathbf{s}^{(j+1)} = BV(\mathbf{A}^{(p-j)}, \mathbf{s}^{(j)})$

the algorithm, the final vector $\mathbf{s} = \mathbf{s}^{(p+1)}$ contains only 0 and 1 and is an approximately efficient balanced sample.

4.7.2 Empirical experiment: the S&P 500 case

The computational analysis is performed on the United States S&P 500 data. The sample consists of asset capitalizations calculated on adjusted daily close prices downloaded from the Chicago-based center for research in security prices (CRSP). The components of the dataset are companies listed in the S&P 500 index at the trading day of 2005 to 2010 without missing values. We have tested the strategy on a basket of stocks available for trading during four periods of four years each. In each period, two years are used as estimation period and the next two years as the investment period. The periods are presented in Table 4.1.

Table 4.1 – *Description of the tested periods*

	Estimation period		Investment period	
Period 1	2005-01-01	2006-12-31	2007-01-01	2008-12-31
Period 2	2006-01-01	2007-12-31	2008-01-01	2009-12-31
Period 3	2007-01-01	2008-12-31	2009-01-01	2010-12-31
Period 4	2008-01-01	2009-12-31	2010-01-01	2011-12-31

Each period corresponds to a total of about 1000 points long each with non-missing observations of stocks. Therefore our empirical investigation does not track the total market capitalization of the S&P 500 index but a subset. Among the S&P 500 stock constituents, we fixed n, the size of tracking portfolio to 50. We have knowingly chosen a small simple size in order to evaluate the method. Obviously, the capacity of tracking an index is directly related to the sample size. As any sampling problem, the accuracy is proportional to the inverse of the square root of the sample size. Thus the more is number of assets, the more is the capacity of tracking the portfolio. Nevertheless, the transaction costs are also increasing with the number of assets. So the sample size should be chosen in order to balance the cost and the accuracy of the portfolio. The size of the portfolio mainly should depend of the capital invested in the portfolio because the relative transaction costs decrease with the capitalization of the portfolio. So this method is mainly proposed in order to save transaction costs by limiting the number of assets in the portfolio while the tracking is as best as possible.

The inclusion probabilities are proportional to the total market capitalization at the end of the estimation period. The flexibility of the selection strategy enables us to run a very large number of simulations to assess the statistical quality of the design. For each period, we constructed $10,000$ portfolios. Algorithm (3) gives stocks with a value superior to 1. Their ticker symbols are: BAC, C, GE, MSFT, PG, XOM, for period 1, GE, MSFT, PG, T, XOM, for period 2, CVX, GE, JNJ, MSFT, PG, T, WMT, XOM, for period 3, and AAPL, JNJ, MSFT, PG, WMT, XOM for period 4. These stocks are thus always included in the tracking portfolio as explained in Algorithm (3).

Figures 4.1 and 4.2 display the results of the tracking strategy during the market phases. Figure 4.1 shows for the four periods the time series dynamics of the total market capitalization and, in dashed lines, the 5th and 95th percentiles capitalization values of the 10,000 portfolios selected according to Algorithm (4). In all the figures, the vertical line separates the estimation period from the investment periods. There are always 2 years for the estimation and 2 years for evaluation. At the end of the investment periods, all the portfolios are equal. Indeed, the same amount is allocated to each portfolio. Figure 4.1 highlights how efficiently the selected portfolios tracked the S&P500 total market capitalization.

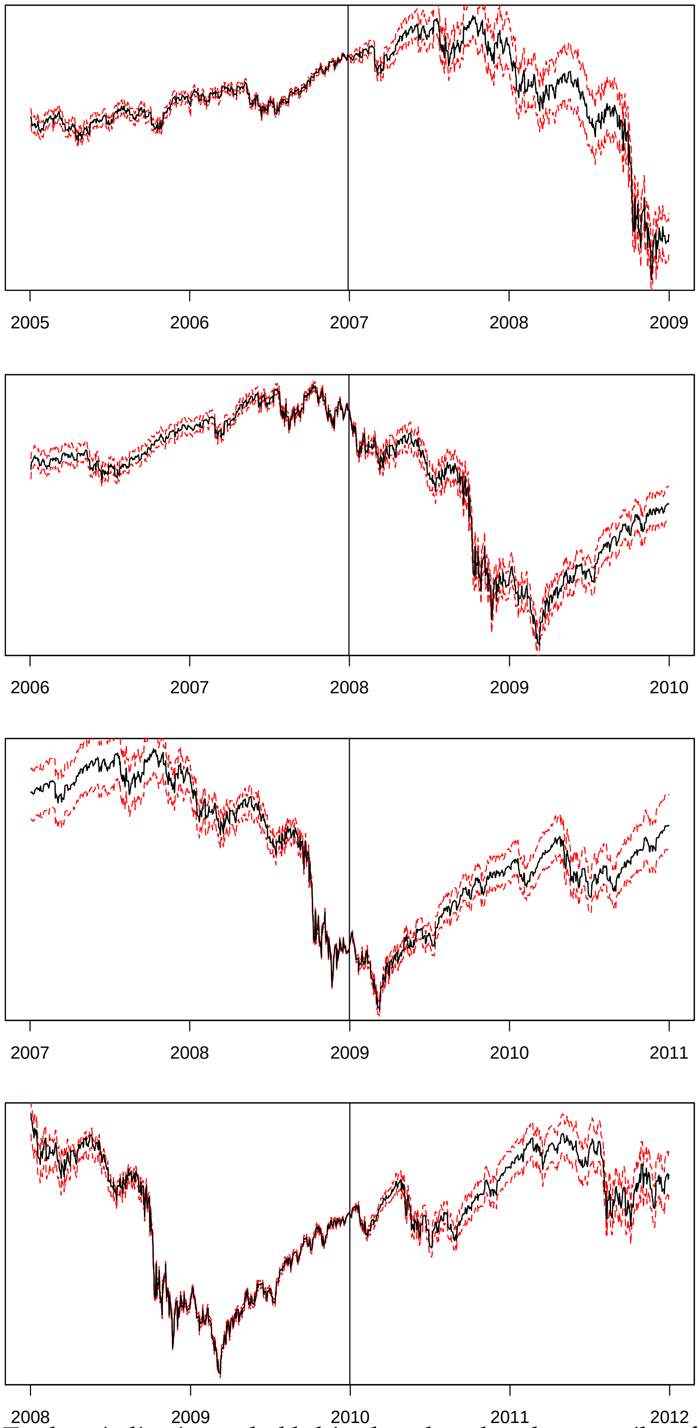

Figure 4.1 – *Total capitalization embedded in the 5th and 95th percentiles of the corresponding values for a selection of 10,000 portfolios derived using the balanced sampling method for the four periods*

Figure 4.2 shows the deciles of the ratios between the 10,000 simulated tracking portfolios and total market capitalization. More precisely, the ratios are defined as:

$$R_t = \frac{P_t}{P_\tau}\frac{C_\tau}{C_t},$$

where τ refers to the portfolios allocation date that is the end of the estimation period. The figure shows that, with only 50 stocks, the basket of trackers remains very "close" to the total market capitalization. The method of balanced sampling works very well in the sense that the deviation during the estimation period is very small. During the in-sample empirical analysis, the interdecile interval is between 0.96 and 1.04 at most, with a deviation close to zero on average. At the end of the estimation period, all the portfolios have the same value and $R_\tau = 1$. During the investment period, the portfolios begin to deviate from market capitalization. The fact that a portfolio remains stuck on the market capitalization during the estimation period does not guarantee that it will remain stuck during the investment period. However, it is interesting to notice that the expectation lies always around 1, which confirms that the portfolios are unbiased.

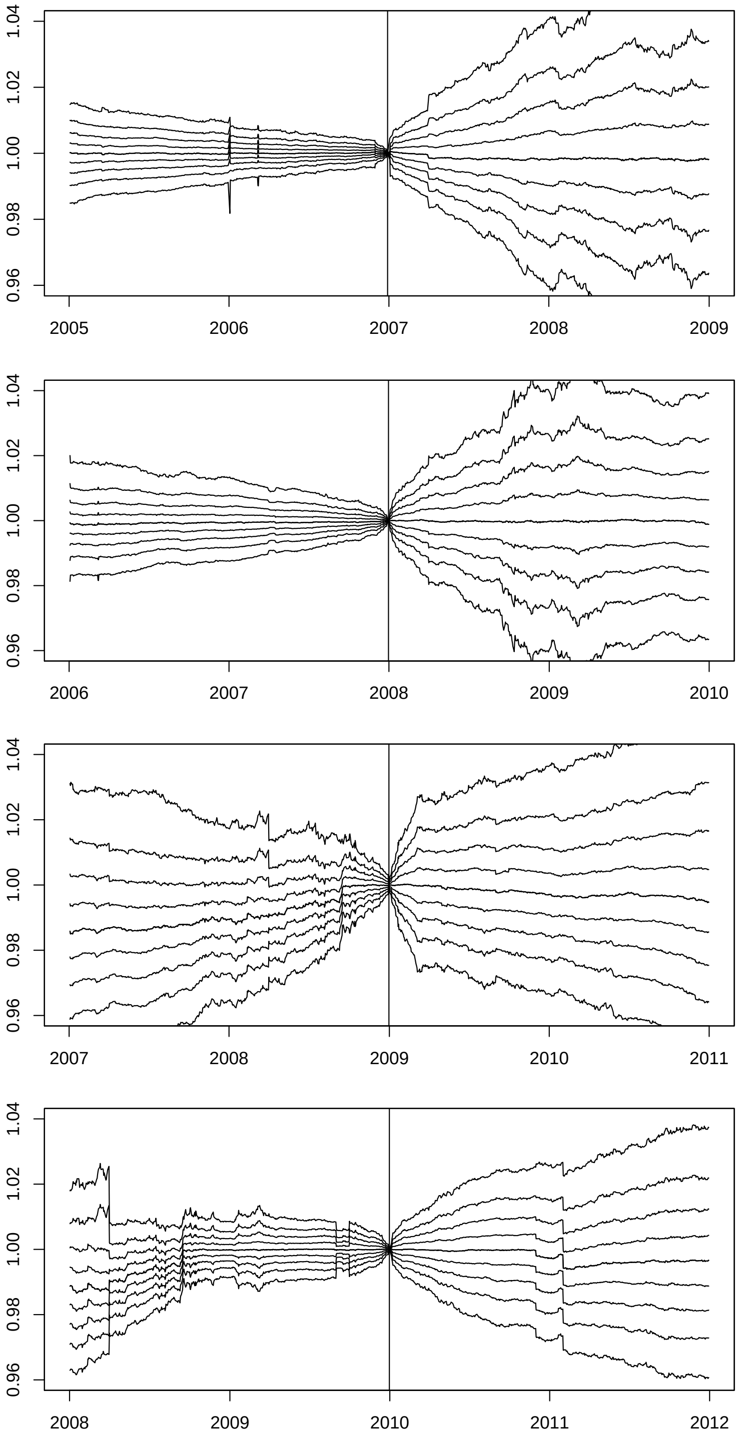

Figure 4.2 – *Deciles of the ratios of* 10, 000 *portfolios selected by balanced sampling over the total capitalization for the four periods*

4.7.3 Comparison with other methods of selection

In this section, we compare the following four strategies.

- Strategy 1: The proposed method of balanced sampling described in this paper.
- Strategy 2: A method of unequal probability sampling with fixed sample size. The selection probabilities are the same as for Strategy 1. A large set of methods of unequal probability sampling are described for instance in Tillé (2006). Most of these methods give equivalent results in term of variance. We have chosen the random systematic method.
- Strategy 3: Simple random sampling without replacement where the same amount is allocated to each selected stock. This strategy is biased under the process of selection of the samples design because it gives too much importance to the stocks with a small capitalization.
- Strategy 4: Simple random sampling without replacement where an amount proportional to the capitalization is allocated to each selected stock. This strategy is unbiased under the sampling design but can provide a portfolio with very unequal weights.
- Strategy 5: This strategy would involve the construction of a tracking portfolio consisting of the 50 stocks with the largest market capitalizations. The assets will be included in the tracker with market weights.

Figure 4.3 contains for strategies 1 to 4 the Relative Means Square Errors (RMSE)

$$\mathrm{RMSE}(P_t) = \mathrm{E}\left(\frac{C_\tau \frac{P_t}{P_\tau} - C_t}{C_t}\right)^2. \tag{4.9}$$

Expectations defined in Expressions (4.9) and (4.10) are approximated with respect to the random selection of portfolios and is approximated by 10,000 simulations for each of the four strategies. For all the portfolios, $\mathrm{RMSE}(P_\tau) = 0$ because a same amount is allocated at the date of creation of the portfolio for all the strategies. The empirical analysis is implemented in the four market phases previously defined. Figure 4.3 clearly shows the benefits of selecting assets with unequal inclusion probabilities proportional to the capitalization. The methods that use simple random sampling and the largest 50 stocks badly work. In the market phase 2007-2010, the performance of the tracker consisting of the largest 50 stocks is undermined by its high volatility. This is probably due to its strong concentration amongst big players, lacking diversification. Balanced sampling surpasses all other methods.

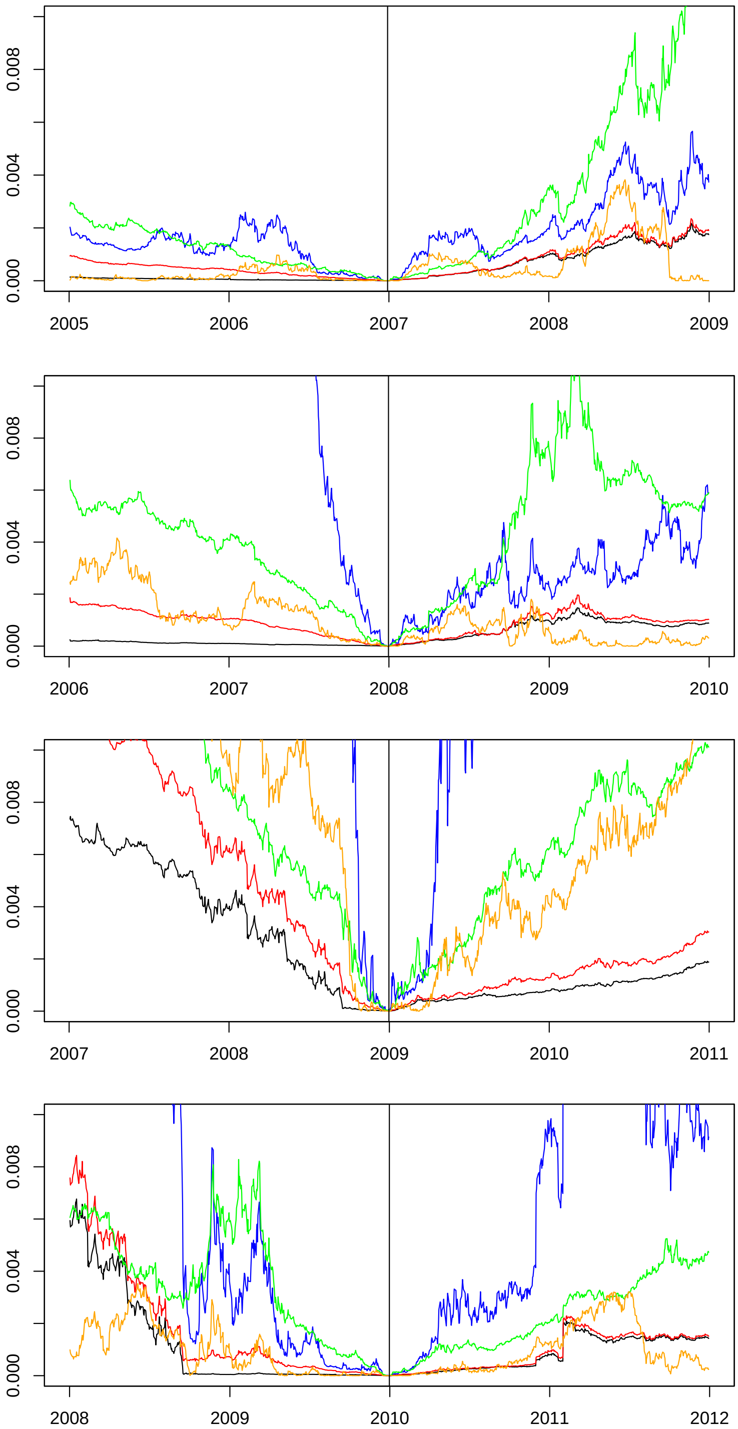

Figure 4.3 – *Relative mean square errors of four different strategies during the four periods: portfolio consisting of the 50 stocks with the largest capitalizations (in orange), balanced sampling (in black), unequal probability sampling (in red), simple random sampling with the same amount (in blue), simple random sampling with an amount proportional to the capitalization (in green)*

Figure 4.4 displays a more accurate comparison of the accuracy of strategies 1 and 2. The figures illustrate the ratio of the RMSE (in black) and the RMAD (in red) of Strategy 1 over Strategy 2. The Relative Median Absolute Deviation (RMAD) is given by

$$\mathrm{RMAD}(P_t) = \mathrm{E}\left|\frac{C_\tau \frac{P_t}{P_\tau} - C_t}{C_t}\right|. \tag{4.10}$$

For the estimation period the gain in accuracy is very apparent. The RMSE of balanced sampling is sometimes 5 times smaller than under unequal probability sampling. The balanced portfolio selection remains very close to the market. At the end of the estimation period, the ratio of RMSE's is indeterminate because the two RMSE's equal 0. During the investment period, there is always a gain particularly in the weeks that follow the creation of the portfolio compared to unequal probability sampling. Balanced sampling thus improves the tracking quality.

We unfortunately have to notice that the fact of being stuck on the market capitalization during the estimation period only partially reduces the deviation from the market during the out-of-sample analysis. This observation is in line with Bajgrowicz and Scaillet (2012)'s assertion that it is hard for an investor to predict what the best performing investment strategies will be in the future. This observation is probably also true for tracking methods based on optimization methods. However, these methods cannot be evaluated by simulations because they provide only one portfolio.

In Figure 4.4, we compare two ways of selecting randomly the assets: balanced sampling and unequal probability sampling. The Figures show that balanced sampling performs better than unequal probability sampling in the investment period (2009-2010) when the portfolio is constructed during a crisis period (2007-2008). Our interpretation is the following: during the crisis period, the series are more volatile and thus contain more information compared to a non-crisis period during which the series are smoother and changes are less sudden. Balanced sampling enables us to reproduces in the portfolio the main characteristic of the market during the estimation period. If the estimation period contains more information, the portfolio is balanced on more information and can more efficiently track the market during the investment period.

This can look contradictory with the recommendations given by Roßbach and Karlow (2011) who have concluded that the size of tracking portfolios should be increased in turbulent markets to achieve a target tracking quality. Nevertheless, Roßbach and Karlow (2011) derives their recommendation from the fact that the variance of the portfolio is proportional to the inverse of the square root of the number of assets. The larger the number of assets, the more accurate is the portfolio in term of tracking.

This is a general rule of statistical inference. During a crisis period, the volatility is larger and thus it could be better to increase the sample size to track more efficiently the market. The recommendation of Roßbach and Karlow (2011) and our interpretation of Figure 4.4 are not contradictory but complementary. Roßbach and Karlow (2011) do not refer to balanced sampling. The volatility increases during the crisis period and the number of assets can be increased, but balanced sampling can be another option to control this increasing of volatility.

Figure 4.5 presents boxplots for the realizations of the $10,000$ tracking portfolios for each of the 5 strategies on the end of January of the first year of the investment period along with mean values for each. It can be seen from Figure 4.5 that Strategy 3 (simple random sampling with equal allocation) and Strategy 5 (stocks with largest capitalizations) are biased. Strategy 4 (simple random sampling with allocation proportional to the capitalization) is unbiased but very unstable, which confirms the interest of unequal probability sampling with equal allocation. The gain of Strategy 1 (balanced sampling) over Strategy 2 (unequal probability sampling) is slightly visible and is confirmed on Figure 4.4.

4.8 Discussion

The paper examined the stock market tracking problem with cardinality constraints in the survey sampling framework. We formulate a novel and completely different approach that can be used to construct a portfolio, containing relatively few securities to track the dynamic in aggregate market capitalization. We cast the issue in a balanced sampling selection process and easily derive a large number of potential tracking portfolios. This new perspective enables researchers to considerably reduce the computational complexity inherent to the cardinality constraint portfolio choice, both in terms of processing time and model assumptions. Furthermore, our methodology is market capitalization-based. Defined as the product between the stock price and the number of shares outstanding, market capitalization is a precise picture of the shares of a company. Additionally, the capitalization-related principle is in line with the passive investment strategy.

In sampling theory, there is not just one way of randomly selecting a sample. We presented four different random selection strategies and a simple heuristic consisting of a tracking portfolio with the 50 largest capitalized stocks. Our empirical experiment has shown that it is more efficient to construct the tracker with unequal sampling probabilities. During the estimation period, the portfolio tracks the change in total market capitalization. With balanced sampling, the portfolios are automatically diversified because the main characteristics of the market are reproduced.

Obviously, our balanced sampling design does not provide a unique

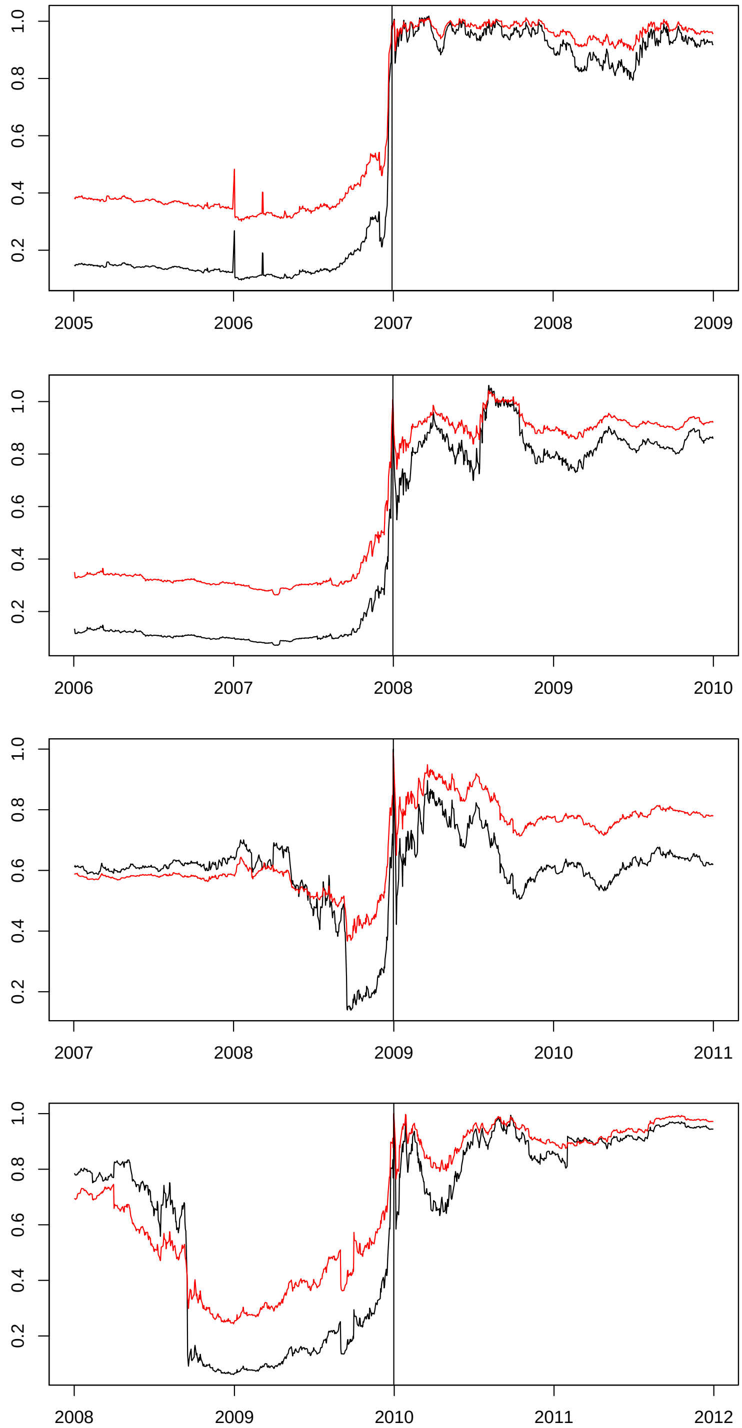

Figure 4.4 – *Ratios of the RMSE (in black) and of RMAD (in red) of Strategy 1 (balanced sampling) over Strategy 2 (unequal probability sampling) during for the four periods*

portfolio, but can randomly generate a large number of different track-

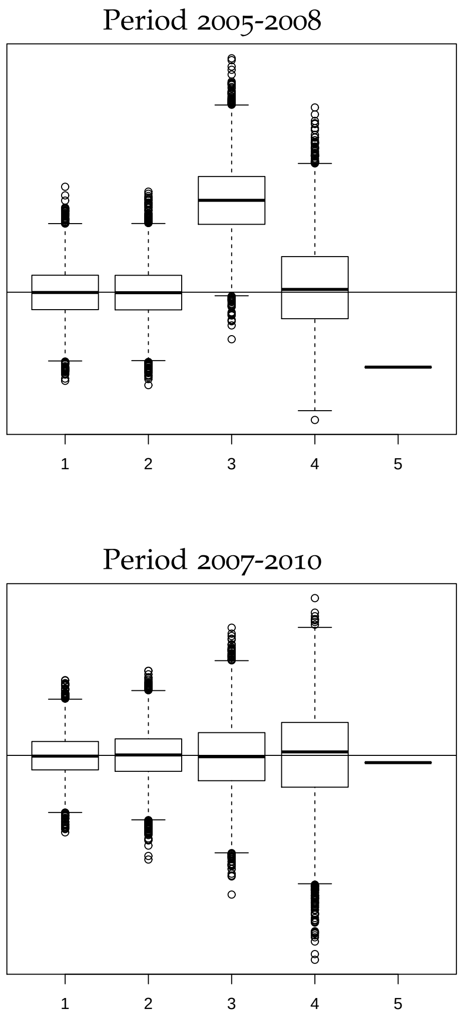

Figure 4.5 – *Boxplots of the replicated capitalizations for the* 10,000 *realizations of each tracking portfolio for strategies 1 to 4 and the realized value of strategy 5. The top plot highlights the values at the end of January of the first year of the investment period. The horizontal line indicates the raw value of the total market capitalization as of the specified trading day.*

ers, all balanced with respect to the target population. This flexibility enables the researcher to perform a consistent statistical evaluation of the methodology over a large range of alternative portfolios. In contrast, the optimization routine yields a unique tracking portfolio. Hence, it is statistically impossible to evaluate by simulations the performance of the optimization-based tracker. All the selected portfolios are equivalent, in the sense they reproduce the time varying evolution of the total market capitalizations and they are diversified. There is no particular reason (statistically speaking) to prefer one over another. It is up to the investor to choose a portfolio in line with her/his investment appetite.

Several extensions to this paper may be considered. Future research might address the issue of how the balanced sampling tracking strategy is affected by the cardinality constraint. In this paper, we set the dimension of tracking portfolios to 50, i.e. around 10% of the complete market. Next, the out-of-sample characteristics of the tracking portfolios could be improved by introducing a portfolio rebalancing strategy, based on a tolerance deviation threshold. This means updating the tracking portfolio constituents when the deviation to a reference benchmark reaches a given threshold. Thirdly, it would be interesting to identify new balancing variables in order to further reduce the dimensionality of the data set. These variables could be additional auxiliary information such as variables given by company balance sheets. In our simulations, we only used the capitalizations of the companies during the observation period. We noticed that the use of asset return series did not provide additional improvements. Another natural extension would be to introduce a second step in the portfolio selection routine as an additional filtering phase. Such an analysis could possibly improve the statistical properties of the tracker.

Acknowledgements

The authors are grateful to anonymous referees for their insightful comments. Thank you to the organizers and participants of the Steyr, Austria, 12 International Research Conference Finance Perspectives for their important suggestions. We are also grateful to Xiqian Zhang who helped us to download and format the data.

5 Tracking Hedge Fund Performance: A Balanced Sampling Strategy

Abstract

This paper discuss the application of statistical, survey sampling technique to hedge fund tracking problems. I describe a strategy that allows an investor or a fund of hedge funds manager, to construct a small tracking portfolio that replicates the time series changes of the total relative Net-Asset-Value (NAV) of a large basket of funds. The trackers are constructed using a method of balanced sampling, in which components are selected randomly with unequal inclusion probabilities. Empirical studies are performed on directional hedge fund styles: Commodities trading advisors, Global macro and Equity hedge. In all cases, empirical results show that the proposed strategy replicated efficiently the total fund's relative NAVs using only ten percent of the sample. The constructed portfolios are stable in the long run, allowing the investor to implement a passive investment strategy in the alternative investment universe. I also consider a larger sample of funds, mixing the aforementioned category of hedge funds. The market tracking ability of balanced sampling remains statistically significant.[1]

Keywords: Hedge fund, Fund tracking, Survey sampling, Balanced sampling

5.1 Introduction

While investing in hedge funds (HF) yields several advantages such as double digit returns and diversification opportunities, the selection of funds in which to invest is a highly challenging issue. This challenge has important implications for the ex-post investor's risk-return structure.

[1]This chapter is a reprint of: Tafin Djoko (2013b) Tracking Hedge Fund Performance: A Balanced Sampling Strategy. Accepted for publication in *Journal of Statistical and Econometric Methods*.

In this article, the attention will be directed solely to the selection problem. How to design a strategy to reproduce closely the behavior of a large universe of HF with a subset of his constituent. The goal hence is not to maximize the investor wealth, but to reproduce faithfully the time series changes of HF full sample performances.

Driven by the strong growth of indexing in the traditional equity universe, a large range of models have been implemented to address the market tracking issue. Markowitz (1987) formulated the market tracking problem, also known as dimension reduction problem, as a mean-variance optimization, making fundamental assumptions on the distribution of the replicated benchmark. However the main difficulty with this approach is the computational burden associated with estimating the variance-covariance matrix for the returns of all assets within the index. Derigs and Nickel (2003) extend the Markowitz-based formulations and derived a meta-heuristic decision support for portfolio management and optimization. Using a linear factor model based on macro-economic variables, they estimated the covariance matrix and then solved the tracking problem. In addition to the traditional tracking error minimization approach, Alexander (1999) advocated a cointegration-based model. Basically, this implementation relies on the fact that the difference between the benchmark and the tracking portfolio is by construction supposed to be stationary. That is, in the long run the deviation of the tracker to his benchmark will be small. Furthermore, Focardi and Fabozzi (2004) illustrated a time-series clustering approach. They argue that clustering might be an optimal methodology for building a sparse tracking portfolio when full replication of an aggregate basket of assets is not feasible. By suitably defining Euclidean distances between the time series of asset prices available in a given market, a hierarchical clustering 'discovers' the correlation structure of the target benchmark.

The purpose of this manuscript is to extend the market tracking principle to the HF world. There are four main reasons to study HF indexing models. First, in practice an investment manager is exposed to a relatively small number of individual HF. According to Lhabitant and Learned (2003), a portfolios of approximately 10 funds are sufficient to construct a diversified portfolio of HF. Second, a portfolio of HF presents specific characteristics such as very high transaction costs, liquidity constraint (notice periods, lockup periods and redemption restriction), and slow rebalancing cycles. As a consequence, an allocation in HF must be inspected as a long term investment. HF indexation may ensure that the tracking portfolio is tied to the fund universe in the long run. Thirdly, the selected tracker should result in a stable portfolio, filtering out the noisy part of the joint variation in the sample. Fourth, if the fund benchmark performances can be reconstructed, we considerably reduce the capital incentive and qualitative analysis burden associated with managing a large portfolio of funds.

Most previous work on tracking financial markets with a subset of its constituents focuses on hard-to-solve optimization techniques and econometric models. However, optimization routines tend to construct highly concentrated replicas, while econometric models such cointegration require consistent estimation of error term and/or structural loadings with direct specification of the data generating process. In this article, I adopt a different perspective and implement an automatic survey sampling approach to select small portfolio of funds in order to track the total HF universe relative NAVs. This small portfolio is selected randomly by means of a method of balanced sampling. Tafin Djoko and Tillé (2013) developed and implemented the balanced sampling replication strategy in the case of the S&P 500 total market capitalization. The methodological frame adopted here is inspired by the later article. Statistically speaking, a sample is said to be balanced if the estimator of a characteristic of interest matches or is approximately equal to the population charateristic. Hence, balanced sampling is implemented to construct a tracking portfolio whose total relative NAV is proportional, at any trading period, to the total relative NAV of a large basket of funds. The diversification principle is automatically incorporated in the efficient balanced sampling procedure without imposing additional constraints.

This article makes several contributions. I move from a hard to solve complex optimization problem, to a simpler and more intuitive survey sampling framework. The issue of efficiently selecting a subset from a large population is a natural sampling topic. There are obviously various ways of randomly selecting a sample. I show that this method can easily generate random portfolios that efficiently track the total performance of HF and maintains the diversification property. The proposed methodology is fully nonparametric and design-based. In addition to the balanced sampling design, the article discusses three alternative random selection strategies and a simple heuristic for constructing a tracking portfolio consisting of funds with the biggest asset under management (AUM). One of the computational benefits of the method is the possibility to run a large number of simulations to evaluate the tracking strategy without increasing the computational difficulty. Furthermore, instead of tracking the return series, we focus on the relative NAV.

The outline of the manuscript is as follows. In Section 5.2, I define the notations and state the aim of the study. Section 5.3 formulates our approach for modeling relative NAVs within the statistical survey sampling framework and for investigating HF indexing. I discuss the computation underlying the inclusion probabilities, which is made proportional to individual fund relative NAV. Then, A short review of survey sampling theory is introduced. Section 5.4 presents the empirical results. I report both a multivariate analysis based on selecting $10,000$ portfolios, and a univariate characterization of the methodology according to the minimum variance portfolio. Further on, I compare the results of our sampling strategy with three alternative random sampling approaches and a simple heuristic involving the construction of a tracking portfolio consisting of funds with the biggest asset under management. Finally, I provide a robustness check of our main results. Section 5.5 concludes the article.

5.2 Notation and objective

To crystalize the context of the analysis, let denote by $R_t = NAV_t/NAV_{t-1}$ the vector of relatives net-asset-value at time t. We have available d funds, constituents of the HF universe with $R_t = (R_t^1, \cdots, R_t^d)$, $t = 1, 2, \cdots, T$, where the jth component $R_i^{(j)}$ of R_i denotes the relative NAV of fund jth on the ith investment period. If T denote the current time, we want to choose the portfolio of HF to track the HF universe at time $T+1$. The total value of the HF universe at time t may be expressed as

$$r_t = \sum_{i=1}^{d} R_t^i \tag{5.1}$$

We assume the usual hypothesis that any fund is infinitely divisible. Any fraction between $[0, 1]$ can thus be selected in the portfolio.

The aim is to select a small portfolio of n funds at time τ that is a subset of a large sample of HF. We would like the portfolio to track the HF universe as closely as possible. To keep the model theoretically simple, we will allocate the same amount to each fund in the tracking portfolio. It is not difficult to generalize the following theory to situations where all the amounts allocated are not equal. In fact during the empirical investigation, the allocation is adjusted according to the inclusion probabilities. Additionally, in the sequel I use the term market to denote the the HF universe.

Let us assume that w_i is the level of capital invested in fund i within the tracking portfolio, and I_i be the indicator variable that takes value 1 if HF i is selected in the portfolio and 0 otherwise. We thus have

$$\sum_{i=1}^{d} I_i = n \tag{5.2}$$

Variables I_i are supposed to be random with a Bernoulli distribution. We suppose that π_i is the probability of selecting unit i in the tracking portfolio. In the following lines, R_t^i and the capital allocated in each fund w_i are not supposed to be random. The only source of randomness is I_i. The expectations and variances are then computed with respect to I_i.

5.3 Methodology

This section presents the methodology for solving the fund selection problem in the context of market tracking. I first characterize the tracking portfolio value at any trading time t and the implications in the random selection process. I then describe how to compute the probability of selecting a fund i in our replicating portfolio. The selection probabilities are defined proportionally to the total market relative NAV. Next, I present the balanced sampling design. For a complete description of sampling algorithms, see Tillé (2006). I approximate the cube method as in Deville and Tillé (2004) to select portfolios that are unbiased, with small variances. Finally, I present the algorithm designed to select a balanced sampling portfolio.

5.3.1 Tracking portfolio value

Given our our interest in constructing a tracking portfolio of HF, we build a model that provides a clear description of the tracker values, V_t. If the portfolio is constructed at time τ, therefore its value is defined as

$$V_\tau = \sum_{i=1}^{d} I_i w_i$$

and for any other t, the factor by which every amount (w_i) invested in the i-th fund grows during the tracking period is given by R_t^i / R_τ^i. Therefore, the portfolio value at time t is expressed as

$$V_t = \sum_{i=1}^{d} w_i I_i \frac{R_t^i}{R_\tau^i}$$

If all the constituents within the portfolio have the same allocation, ie, $w_i = w$, then

$$V_\tau = w \sum_{i=1}^{d} I_i = wn$$

and for any other time t

$$V_t = w \sum_{i=1}^{d} I_i \frac{R_t^i}{R_\tau^i} \tag{5.3}$$

The portfolio will be constructed randomly by means of a sampling technique. To satisfy our tracking objective at time τ, that is, to build a portfolio that is as close as possible to the complete market during the period $\tau - q, \ldots, \tau$, we would like

$$\frac{V_t}{V_\tau} \approx \frac{r_t}{r_\tau}, \text{ for all } t = \tau - q, \ldots, \tau \tag{5.4}$$

Equations (5.4) imply q linear constraints on V_t. However, in the market tracking issue, the replicator is bound by the limited number of n securities out of the market d $(n \ll d)$ that can be selected to track the target benchmark. Since the number of observations q is larger than n, we can not satisfy more constraints than the number of units in the tracking portfolio. For this reason, we should reduce the number of constraints that must be satisfied using singular value decomposition as in Tafin Djoko and Tillé (2013). In what follow, period $\tau - q$ to τ is referred to as the portfolio estimation period, while the period that follows is referred to as the investment period.

5.3.2 Random selection

The selection mechanism is said to be random when all units in the population of interest have a positive probability of being selected. Random selection has been for decades, source of intense debate in the scientific community. As opposed to non-probability sampling in which the selection probabilities are unknown or/and some statistical units have zero probability of being selected, random sampling process does not suffer of self selection bias and allows the researcher to perform inferences on the parameter of interest.

When holding index-related products, the primary interest of a rational investor is diversification. Therefore, an index tracking portfolio should not only solve the size reduction issue, but predominantly preserve the diversification attribute. Random selection is the only procedure able to reduce the dimension and infer from the sample, characteristics of the population. Consequently, the tracking portfolio diversification is guaranteed under the random selection process.

5.3.3 Selection probability

The key and powerful characteristic of our tracking strategy is randomness. Each fund is selected randomly according to a given probability mass. Let π_i be the probability of selecting individual HF i. Thus, we can write

$$\pi_i = \mathrm{Pr(selecting\ fund\ i\ in\ portfolio\ j)} = \mathrm{E(I_i)},$$

and given Equation (5.2)

$$\sum_{i=1}^{d} \pi_i = \sum_{i=1}^{d} \mathrm{E}\left(I_i\right) = \mathrm{E}\left(\sum_{i=1}^{d} I_i\right) = \mathrm{E}(n) = n \tag{5.5}$$

the sum of the selection probabilities is thus equal to the number of funds in the tracking portfolio.

Since each unit in the portfolio have the same allocation w at time τ, a convenient alternative for computing the inclusion probabilities is to construct the vector of π_i proportional to the relative NAVs at time τ, that is,

$$\pi_i \propto R^i_\tau \,,\, i = 1, \ldots, d$$

Given Equation (5.5), the selection probabilities can easily be derived according to the following expression

$$\pi_i = \frac{nR^i_\tau}{r_\tau} \tag{5.6}$$

Consequently, it is evident that the portfolio value at any given trading period, is proportional to the total market relative NAV. Effectively, a short decomposition of Equations (5.3) and (5.6), gives

$$\begin{aligned} \mathrm{E}\left(V_t\right) &= \mathrm{E}\left(w \sum_{i=1}^{d} I_i \frac{R^i_t}{R^i_\tau}\right) = w \sum_{i=1}^{d} \frac{R^i_t}{R^i_\tau} \mathrm{E}\left(I_i\right) = w \sum_{i=1}^{d} \frac{R^i_t}{R^i_\tau} \pi_i \\ &= w \sum_{i=1}^{d} \frac{R^i_t}{R^i_\tau} \frac{nR^i_\tau}{r_\tau} = \frac{nw}{r_\tau} \sum_{i=1}^{d} R^i_t = nw \frac{r_t}{r_\tau} \end{aligned}$$

where r_τ is expressed as in Equation (5.1).

Indeed, if the allocation decision consists of investing the same amount to each fund, the selection probabilities are proportional to the total relative NAV in the HF universe. This strategy is unbiased under the sampling design and increases the likelihood that some funds with large relative NAV will be selected. The randomness of the selection scheme ensures that all constituents are given the same opportunity to enter the tracking portfolio and not only those that are 'tied' to the market. The diversification is thus automatically guarantied. This procedure also avoids selecting a disproportionate number of funds whose performs badly. The aim is thus not only to track the total market relative NAV but predominantly to keep its intrinsic diversification. In other words, keeping in balance the relative structure between highly and modestly performing funds in the tracking portfolio.

5.3.4 Balanced sampling design

Given my interest in tracking the total relative NAV of funds in the HF universe with a subset of its constituents, I build a model based on sampling methods. Sampling can be defined as a process of selecting statistical units from a population to estimate unknown quantities of the population. The author implemented a balanced sampling strategy. Balanced sampling design has the property that the Horvitz and Thompson (1952) estimator of the mean is unbiased. Consequently, the variances of all the parameters of interest are reduced. In general, balanced sampling strategy uses auxiliary information that is known for all the population units, to construct an efficient sampling design. Moreover, sample is said to be approximately balanced on a set of auxiliary variables if its estimator of mean $\bar{X}$ will be equal to the known population mean of these variables. Balanced samples can be selected with equal or unequal selection probabilities.

The sampling algorithm used in this article has been developed by Deville and Tillé (2004). The authors proposed a general procedure to select balanced samples named the Cube method. The method is non-rejective and is composed of two phases, called the flight and the landing phase. During the flight phase, the balanced constraints are exactly satisfied. The later phase focuses on managing the non satisfied balancing equations. In case of the total, balancing equations are

$$\sum_{i=1}^{d} \frac{\mathbf{x}_i}{\pi_i} I_i \approx \sum_{i=1}^{d} \mathbf{x}_i \tag{5.7}$$

where $\mathbf{x}_i$ is a vector of $\mathbb{R}^p$ containing the p values taken by the p auxiliary variables on statistical unit i. For more details on sampling algorithms see Tillé (2006).

In this paper, our aim is to select a portfolio of funds using this technique of balanced sampling. We will thus randomly select a subset of individual HF to reconstruct as close as possible, in the long run the time series changes total relative NAVs in the universe of funds. Since our principal objective is to implement a passive investment strategy in the alternative investment world, our tracker has to be manageable in terms of size, transaction costs, and rebalancing cycles. In other words, the balanced sampling replicator should result in a steady portfolio structure. Under these specifications, statistical units are fund's relative NAV, and the auxiliary information is the set of values taken by each fund during a period.

This model has several appealing features. First, it is distribution free. There is no assumption on the data generating process of HF's relative NAV and the optimal tracker is derived without specifying any structural form between the total market value and its individual constituents. Under the balanced sampling design, the selection technique is statistically efficient. In other words, the selected portfolios are guarantied to be unbiased with a small variance. Third, the model yields tracking portfolios that are automatically diversified without specific heuristic hypothesis, allowing the investor to perform allocation in an optimal way. Fourth, the proposed replication strategy is highly flexible, providing an alternative insight to the computational intensive constraint portfolio choice.

5.4 Empirical Analysis

5.4.1 Data description

To attest that the proposed methodology provides general results, I used three different HF categories to evaluate the statistical significance balanced sampling tracking portfolios. We focus on live funds from two main HF data providers: Hedge Fund Research (HFR), and Barclay Commodity Trading Advisors (CTAs) databases. The former data provider is composed of returns reported on different frequencies (mainly monthly) and additional qualitative/quantitative information such as main strategy, sub-strategy, asset under management, etc. on 5230 individual funds and 2720 Funds-of-Funds that are still active on September 30, 2008, while the later data system is widely recognized by both practitioners and academicians as the largest, most comprehensive, available Commodity Trading Advisors sample. Generally, CTAs are funds primarily trading listed commodity and financial futures contracts. The database consists of 981 reporting funds as of September 2008. CTAs, also denominated Managed futures are by no means homogeneous investment vehicles. CTAs managers employ a large range of strategies and asset classes. Combined, the two databases provide a rather completed and detailed picture of the hedge funds universe.

The alternative investment industry data are subject to different biases resulting from the data collection process. Fund managers *voluntarily* provide information to databases and the industry lacks an uniform reporting standard. Since I confine the analysis to funds in the live database, I recognized that it suffers from survivorship bias. However, the relevance of such a bias for the current investigation is limited by the fact that it will affect both the tracker and the full population of funds. The tracking portfolio is just a reconstruction of what we have in the large sample.

For the scope of our empirical analysis, we impose a set of filters on both databases. First we selected funds that reported in U.S. dollar net of all fees on a monthly basis. Then we required that each fund has at least 12 years of reported returns. Additionally, for each strategy, we extract the group of funds with the longest consecutive stretch of non-missing NAVs and AUM. The analysis is performed on HF directional strategies: CTAs funds from Barclay; Global Macro and Equity Hedge funds from HFR. After imposing these restrictions, the data comprises 134 CTAs covering the period from October 1996 to June 2008, for a total of 141 observations; 101 Global Macro funds from July 1997 to September 2007, giving 141 observations; and 108 Equity Hedge funds from May 1995 to September 2007 for a total of 149 monthly observations. To avoid in-sample spurious findings, these samples period are divided in two sub-periods: The first 60 months are used for backtesting the quality of the model (also called estimation period), while the second samples (longer or equal to the first one) are used for the out-of-sample investigation (called investment period).

Table 5.1 reports several summary statistics for the fund relative NAVs under study. The statistics are computed over the entire data period for each HF and then averaged across funds within each HF category. The monthly average relative NAVs are all bigger than one and significant, ranging between 1.008 to 1.012 with small standard deviations amongst funds. The average monthly volatilities range between 6% and 5%. Equity hedge funds tends to exhibit high expected relative NAVs with low risk, whereas Global Macro funds have low expected relative NAVs with low risk. Barclay CTAs funds have high average relative NAVs and high risk.

Skewness measures are all positive across HF categories, suggesting that periods of booms are more present than periods of crashes. Kurtosis measures are equal to 5.6 for Barclay CTAs funds, 6.4 for Global Macro and 6.1 for Equity Hedge funds, values that are not in line with the normality assumption. Consequently, we reject the normality assumption with strong confidence for all HF categories in the sample. Except for Equity Hedge funds, in average, I don't find strong evidence for serial correlation and ARCH effect in fund's relative NAVs.

Table 5.1 – *This table reports monthly average summary statistics of HF relative NAVs for the entire sample period (estimation and investment periods). The mean and t-statistic value, the standard deviation, the standardized skewness and kurtosis with their respective t-statistics, the Jarque-Bera normality test statistic (JB), the Ljung-Box test statistic for no serial correlation [LB(6)], the first-order serial correlation of relative NAVs [$\rho(R)$], the Lagrangian multiplier test statistic for no serial correlation in relative NAVs [LM(6)]. The critical values at 5% are 5.99 for JB, 12.59 for LB(6), $\rho(R)$, and LM(6).*

	Barclay CTAs		**Global Macro**		**Equity Hedge**	
Sample Size	134		101		108	
	Mean	SD	Mean	SD	Mean	SD
Mean	1.011	0.006	1.008	0.004	1.012	0.005
Mean t-stat	2.421	1.132	2.085	1.047	3.738	2.939
SD	0.058	0.03	0.048	0.022	0.048	0.024
Skewness	0.447	0.754	0.303	1.235	0.025	0.879
Skew t-stat	2.165	3.655	1.372	5.592	0.125	4.379
Kurtosis	5.586	4.551	6.372	8.66	6.058	4.778
Kur t-stat	6.268	11.031	7.633	19.604	7.618	11.904
Min	0.845	0.109	0.865	0.116	0.844	0.092
Max	1.223	0.141	1.166	0.077	1.184	0.125
JB	178.024	826.616	471.629	3289.655	217.447	949.74
LB(6)	10.129	7.363	8.693	6.861	20.741	77.413
$\rho(R)$	7.922	4.734	6.866	4.391	6.934	4.822
LM(6)	7.879	7.856	6.171	5.385	14.29	11.043

5.4.2 Balanced sampling tracking portfolio

We now turn to the analysis of the performance of the market tracking ability of the strategy described above. The goal here is to construct a long term and stable buy-and-hold strategy in the alternative investment world.

The balanced sampling portfolio selection strategy implemented for 3 HF categories under study is performed as follows. I divide the sample in two non-overlapping sub-periods. The first 60 months are used to select the tracking portfolios and perform a backtest of the method. The second sub-period is dedicated to the out-of-sample analysis. In the sequel I refer to the first sub-period as estimation period and to the second as investment period. In all the figures, the vertical line separates the two sub-periods. The number of funds selected in each tracking portfolio, $n = 10$. This figure represents less than 10% of the population of interest across HF categories.

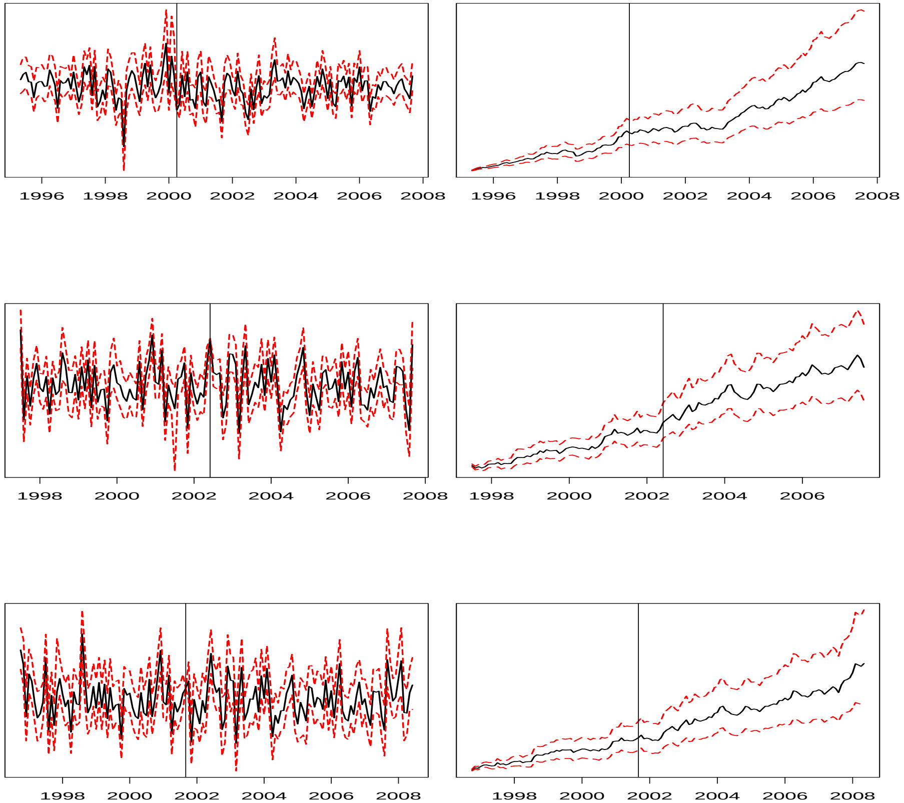

Figure 5.1 – *The first-column entries are the total HF relative NAVs (black line) embedded in the 5th and 95th percentiles (red line) of the corresponding values for a selection of 10,000 portfolios derived using the balanced sampling tracking strategy. The second-column entries represented the wealth achieved investing in an equal weighted portfolio of the full sample of funds (black line) embedded in the 5th and 95th percentiles (red line) of the corresponding values for a selection of 10,000 portfolios with the proposed strategy. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of $n = 10$ funds.*

In finance, there is general consensus on how difficult is to separate skill from luck. This issue is even more stringent for hedge fund investors. The opacity inherent from the HF industry makes hard to distinguish knowledgeable fund picking strategies from lucky ones. The flexibility of our method allows us to avoid that by running a large number of simulations. This multivariate analysis is useful to assess the statistical quality of the sampling design. I fixed the number of simulations to 10,000, that is, we construct 10,000 tracking portfolios for each HF category. For each HF style, the inclusion probabilities are proportional to the total relative NAVs in the last month of the estimation period (60th month).

The first column-entries of Figure 5.1 display the dynamics of the total relative NAVs and, in dashed lines, the 5th and 95th percentiles relative NAVs of the 10,000 portfolios selected according to Algorithm (**??**). The second column-entries of Figure 5.1 shows the evolution of the wealth achieved investing 1 USA $ in an equal weighted portfolio constituted of the entire sample of funds with the corresponding 5th and 95th percentiles from 10,000 balanced sampling trackers. Inspection of these figures highlight how efficiently the selected portfolios tracked the total relative NAVs. The performance is significant in both estimation and investment sub-periods. The inter-percentile range is narrowed in the two sub-periods and the gap is relatively stable in the long run.

The graphs in Figure 5.2 display the dynamics the deciles of the ratios between the 10,000 simulated tracking portfolio values and total HF relative NAVs (first column) for the three categories of alternative investment under study. More precisely, the ratios are defined as:

$$Ratio_t = \frac{V_t}{V_\tau}\frac{r_\tau}{r_t} \tag{5.8}$$

where $\tau = 60$ refers to the portfolio estimation date. The corresponding ratio for the wealth achieved is displayed in the second column of Figure 5.2. Several comments are of interest. First, with only 10 funds, the sample of trackers remains very 'close' to the portfolio investing in all the funds in the universe. The method of balanced sampling works well in the sense that the deviation during the estimation period is very small. During the in-sample empirical analysis, the ratio dynamics of relative NAVs range between 0.97 and 1.04 (except for one large swift for Global Macro funds), with a deviation close to zero on average. At allocation time (60th month), all the portfolios have the same value and $Ratio_\tau = 1$. After this month, the portfolios start to deviate from the total market value, but stay in a narrow range. For the wealth achieved, we observe comparable dynamics. It is interesting to notice that in expectation, the tracking portfolio ratios stay well around 1, which indicates that the portfolios are unbiased. Furthermore, by increasing the number of replication, I decrease the variance while increasing the power of the balanced tracking strategy. This important characteristic is further illustrated in the sequel from Figure 5.3.

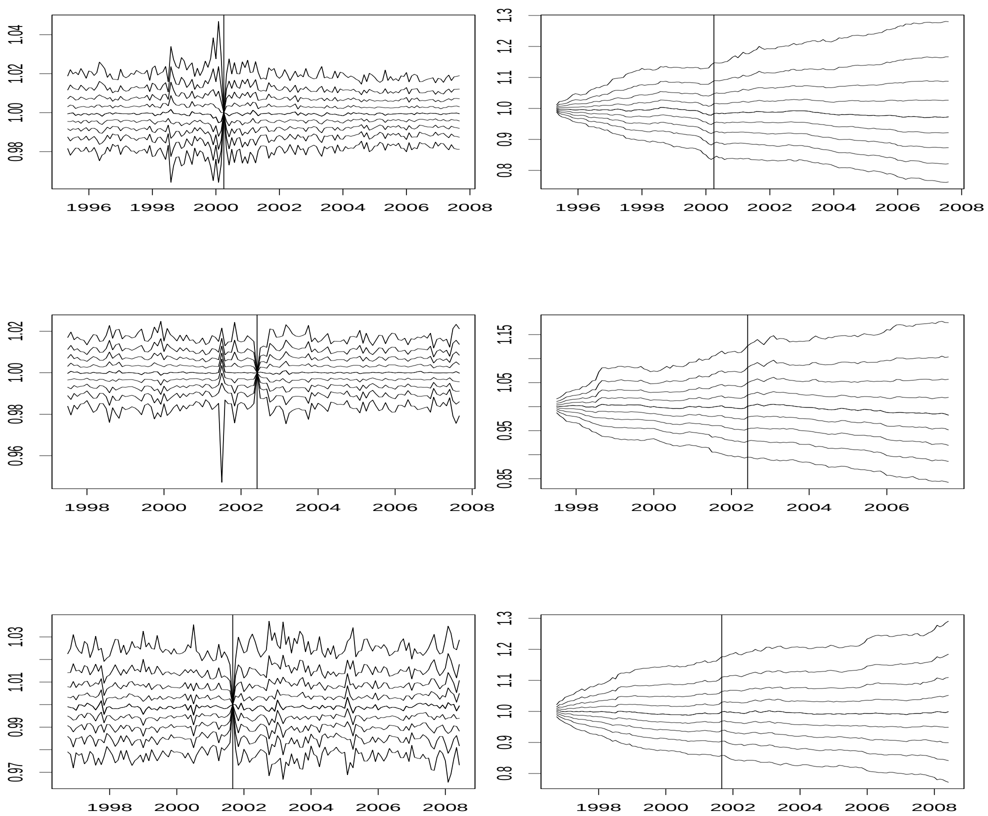

Figure 5.2 – *Deciles of the ratios from* $10,000$ *portfolios selected by balanced sampling over the portfolio constituted of the entire sample of funds. The first column are relative NAVs and the second are wealth achieved. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of* $n = 10$ *funds.*

5.4.3 Comparison with other methods of selection

We now turn to the analysis of the performance of our tracking strategy relative to several alternatives. This control phase is performed against three different random selection strategies and one 'do nothing' strategy consisting of a portfolio with the n biggest funds according to their AUM. Each strategy will provide some insights of the relative value of tracking strategy trough randomization over a simple asset based heuristic strategy. In this section, we compare the following four strategies:

- Strategy 1: The proposed method of balanced sampling described in this paper.
- Strategy 2: A method of unequal probability sampling with fixed sample size. The selection probabilities are the same as for Strategy 1. A large set of methods of unequal probability sampling are described for instance in Tillé (2006). Most of these methods give equivalent results in term of variance. We have chosen the random systematic method.
- Strategy 3: Simple random sampling without replacement where the same amount is allocated to each selected fund. This strategy is biased under the process of selecting the sampling design because it gives too much importance to poor performing funds (small relative NAVs).
- Strategy 4: Simple random sampling without replacement where an amount proportional to the relative NAVs is allocated to each selected HF. This strategy is unbiased under the sampling design but can provide a portfolio with very unequal weights.
- Strategy 5: This strategy would involve the construction of a tracking portfolio consisting of the $n = 10$ funds with the biggest AUM ('do nothing' strategy). The funds will be included in the tracker with weights corresponding to their size in the HF universe.

Figure 5.3 presents boxplots for the realizations of the $10,000$ tracking portfolios for each of the 5 strategies on a given trading day during the estimation and investment periods, along with mean values for each. It can be seen from Figure 5.3 that a dominance relationship exists between our balance sampling design and the four alternative tracking strategies. Our replication strategy is unbiased and much less volatile in both the estimation and investment sub-samples. While random sampling strategies 2, and 4 are effectively unbiased but more volatile than the proposed balanced sampling design, strategy 3 is biased and also less efficient. The beauty of a large number of simulations is the ability to discriminate among concurrent strategies. However, it is difficult to compare our proposed methodology to the 'do nothing' portfolio. The graphs in Figure 5.3 show its value is different to the true total relative NAVs. In what follow, I will try to solve this issue by moving from a multivariate analysis to a univariate one. I will compare the cumulative relative mean square error of the big funds portfolio to the best tracker amongst the $10,000$ constructed portfolios.

5.4.4 Univariate analysis

Statistically speaking, all the selected portfolios are equivalent, in the sense they reproduce the time varying evolution of the total HF relative NAVs and they are automatically diversified between highly and poorly performing funds. There is no particular reason to prefer one over another. It is up to the investor to choose a portfolio in line with his investment appetite. Moreover, since the portfolio is constructed during first the sub-sample (60 months length), there is no guaranty that the best tracking portfolio will keep performing well in the second sub-sample of data.

For practical reason, I select one of the $10,000$ constructed portfolios to directly assess its tracking performance relative to the portfolio consisting of the n biggest funds. Since in this manuscript I consider an investor who allocates his HF portfolio of n funds to minimize

the distance to the total performance of funds in the universe, I select the fund with the minimum Relative Mean Square Error (RMSE) during the estimation period (first sub-sample). According to Equation (5.4), the RMSE for strategy 1 is defined as

$$\mathrm{RMSE}(V_t) = \mathrm{E}\left(\frac{r_\tau \frac{V_t}{V_\tau} - r_t}{r_t}\right)^2 \tag{5.9}$$

while the corresponding measure for the 'naive' strategy 5 is expressed as

$$\mathrm{RMSE}(V_t) = \left(\frac{r_\tau \frac{r_t}{r_\tau} - r_t}{r_t}\right)^2 \tag{5.10}$$

Figure 5.4 shows the evolution of the total relative NAVs and wealth achieved of the minimum variance portfolio along the corresponding values for all the funds in the HF universe.

Several comments are of interest. First for the 3 HF categories under investigation, the best tracker is closely tied to the total market values. Second, it is of great interest to notice that the best in-sample selected portfolio performs very well in the out-of-sample analysis. Third the performance quality of the best tracking strategy is stable in the long term perspective. Unreported results of two-sample Wilcoxon test and paired samples

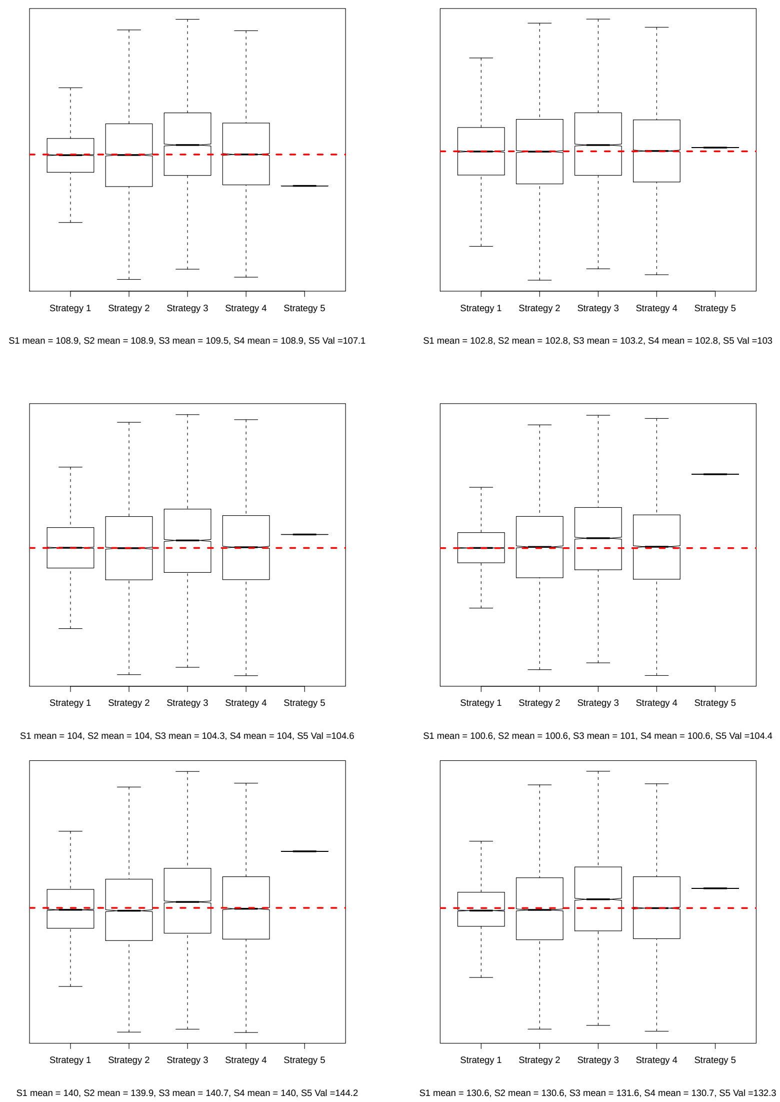

Figure 5.3 – *Boxplots of the replicated relative NAVs for the* 10,000 *realizations of each tracking portfolio for strategies 1 to 4 and the realized value of strategy 5. The first-column entries highlight the values at a given trading day of the estimation period. The second-column entries are the numbers occurring in a random day of the investment period. The horizontal red dotted line indicates the raw value of the total relative NAVs as of the specified trading day. Mean values for each strategy appear in the subtitle below the figure. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of* $n = 10$ *funds.*

t-test strongly reject the hypothesis that the two series are different at any confidence level. The second-column entries of Figure 5.4 show that our balanced sampling tracker and the portfolio consisting of all the funds in the market achieved approximately identical wealth.

Finally, I illustrate the tracking quality of the minimum variance portfolio relative to the big funds portfolio. Recording, the big funds portfolio is AUM-weighted portfolio. Figure 5.5 displays the time plots of the cumulative relative mean square errors for the two strategies as expressed in Equations 5.9 and 5.10. Not surprisingly, the balanced sampling strategy strongly over-performs the 'do nothing' strategy. As it appears, the gain of adopting the proposed tracking design is both statistically and economically significant, as it guaranties a diversify and unbiased portfolio, achieving approximately the true dynamic changes of the total relative NAVs in the HF universe.

5.4.5 Robustness analysis

To further evaluate the robustness of the balanced tracking strategy, I perform a last empirical analysis. The experiment is based on the same data. I mixed the 3 samples of HF to have a consistent alternative investment universe. The final data set results in 343 funds covering the period from July, 1997 to September, 2007, which includes the dotcom bubble burst and the summer 2007 subprime crisis. I keep the same empirical framework using the first 60 months of data to construct the $10,000$ tracking portfolios and left the allocation constant for the rest of the period. The portfolios are not rebalanced, mainly because I am interested to the long term behavior of the selected portfolios. The indexing portfolios contains $n = 35$ funds, which represents approximately 10 percent of the total number of funds in the benchmark. The graphs in Figure 5.6 show that the experiment with the new large benchmark performs well despite the low frequency associated to HF data. In particular, we notice from the second-row entry in Figure 5.6 that variability amongst the $10,000$ constructed portfolios is reduced. The deciles of the ratio defined in Equation (5.8) range between 0.99 and 1.01. Since the proposed tracking procedure is unbiased, a smaller variance means that the estimate tracker is closer to the true value of the benchmark. In this large sample context, the performance of the balanced sampling strategy improves. The superiority of our proposal is further highlight in the graphs of Figure 5.7, where we can see that the distribution values of the selected tracking portfolios is unbiased and less volatile in both sub-periods than the 4 alternative strategies. The evolution of the cumulative relative mean square error, as depicted in the lower graph of Figure 5.7 shows the dominance relationship among strategies 1 to 5. It can be seen that a stochastic dominance relationship exists again indicating that our balanced sampling tracking strategy is to be preferred on the basis of its performance.

We now turn to a comparison analysis between moments of the total relative NAVs of the balanced sampling and 'do nothing' tracking strategies and the original funds from which the tracking portfolios are derived. Table 5.2 presents the comparison summary statistics. The results are striking in both estimation and investment sub-periods. During the first sub-period, the average moments amongst the 10,000 portfolios (mean, standard deviation, skewness and kurtosis) are similar to those of their benchmark. This is not the case for the big AUM funds, which figures differ significantly from the benchmark. Expect for the skewness, I have comparable numbers in the investment period.

In order to understand the relationship between the 10,000 selected tracking portfolios according to the balanced sampling strategy and the target benchmark, a simple ordinary least squared regression analysis is implemented. I regress the total relative NAVs of each selected fund to their benchmark counterpart and the average betas estimated, t-statistics and adjusted-R^2 are reported in Table 5.2. If a tracker efficiently replicates the benchmark, the regression slope should be equal approximately to 1. According to the table, the average betas and adjusted-R^2 are close to 1.

Table 5.2 – *This table reports summary statistics on the total relative NAVs for balanced sampling* 10,000 *trackers, the best tracker (in term of relative mean square error), the big AUM, and the benchmark (consisting of the entire sample of funds). The statistics corresponding to the* 10,000 *selected portfolios are sample averages. Panels B.1 and B.2 report average statistics for linearly regressing each of the* 10,000 *tracking portfolios to the benchmark.*

Panel A: Allocation period				
Panel A.1: Moments				
	μ	σ	sk	ku
Benchmark	346.37	6.93	0.42	-0.06
10,000 trackers	346.42	7.15	0.39	-0.06
Best tracker	346.38	7.03	0.39	-0.19
Big AUM funds	348.27	5.58	0.19	0.13
Panel A.2: Beta Sensitivity				
		$\hat{\beta}$	*t.test*	*Adj.Rsq*
		0.99	1535.42	0.99
Panel B: Investment period				
Panel B.1: Moments				
Benchmark	346.01	7.31	-0.01	-0.47
10,000 trackers	346.04	7.57	0.00	-0.36
Best tracker	346.20	8.36	0.08	-0.41
Big AUM funds	347.412	5.84	-0.081	-0.219
Panel B.2: Beta Sensitivity				
		$\hat{\beta}$	*t.test*	*Adj.Rsq*
		0.98	1399.14	0.99

5.4.6 The economic value of balanced sampling strategy: Diversification

The randomness of the selection frame gives equal opportunity to individual HF to be part of the tracker. All constituents have a strictly positive probability of being selected, ensuring that the population of interest is totally covered. Additionally, the balanced sampling design yields representative sample, since it allows estimating exactly the population (HF) total relative NAVs, that is, without bias. The key diversification property is then automatically incorporated in the efficient balanced sampling strategy.

To highlight the diversification characteristic of the balanced sampling strategy, I decomposed the $10,000$ trackers according to which HF category each fund belongs and their size (AUM). While the upper plot of Figure 5.8 depicts the proportion of funds relative to style, the lower plot shows the tracking portfolio distribution based on fund sizes[2] (Small Funds: AUM ≤ 10.5, Middle Funds: AUM $\in (10.5, 154)$, Big Funds: AUM ≥ 154). One can clearly see that the randomness of the selection scheme guaranties a stable portfolio constitution across the $10,000$ trackers. The population of individual HF is well covered from Small to Big funds and amongst styles. On average[3], balanced sampling trackers are formed of 25.78%, 49.53%, and 24.68% for the Small, Middle, and Big funds, respectively; Additionally, 40.35%, 29.42%, and 30.22% for CTAs, Equity Hedge, and Macro funds, respectively.

5.5 Conclusion

In this paper, I investigate the HF indexing problem with cardinality constraints in the survey sampling framework. Whereas most previous work has been devoted to the variance and tracking error minimization, and co-integration based index tracking, I formulate that the issue of selecting a small sample from a large one is a natural sampling problem. I solve the issue in a balanced sampling selection process and easily derive a large number of potential tracking portfolios. For a review on HF indexing, see Alexander and Dimitriu (2005a). The present study contributes to the market indexing literature by providing several additional insights. From a computational point of view, the proposed model enables researchers to considerably reduce the computational complexity inherent to the cardinality constraint portfolio choice, both in terms of processing time and

[2]Fund classifications by size are performed according to the first and third quartile of the population AUM. The figures are in millions US$ and correspond to 10.5, and 154 for the first and third quartile, respectively.

[3]The decomposition of the minimum variance portfolio is of equal magnitude to the average: 20%, 48.57%, and 31.42% for the Small, Middle, and Big funds, respectively; And, 40.43%, 29.35%, and 30.21% for CTAs, Equity Hedge, and Macro funds, respectively.

model assumptions. Despite the necessary large number of simulations to statistically assess the long term quality of the tracker, the estimation of the model remains tractable. From a theoretical perspective, the methodology implements does rely on any assumption on the data generating process and it is free of heuristic constraints. This setting yields a highly consistent, data driven tracking portfolios.

We demonstrate that the portfolios constructed under the balanced sampling design closely track the total relative NAVs of individual funds in the HF universe. The performance of the tracker is robust both in-sample and out-of-sample, and it is stable in the long run. This characteristic is fundamental for a passive investment strategy - specially in the alternative investment world - since it avoids a costly rebalancing mechanism. I consider three different HF categories and a large sample consisting of merging the 3 main styles in an unique global benchmark, and I further confirmed in this case the relevance of constructing HF tracker under a statistical balanced sampling design. This development can be successfully applied for all asset classes and various auxiliary information, from traditional investments to complex funds of HF portfolios.

Acknowledgements

The author is grateful to anonymous referees for their insightful comments. Thank you to professor Yves Tillé important suggestions.

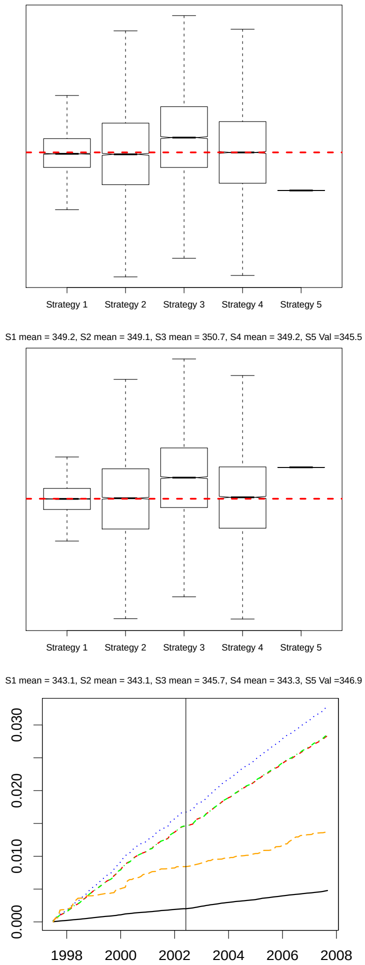

Figure 5.7 – *The first two row-entries display boxplots of the replicated relative NAVs for the* 10, 000 *realizations of each tracking portfolio for strategies 1 to 4 and the realized value of strategy 5. The first graph highlights the values at a given trading day of the estimation period and the second are the numbers occurring in a random day of the investment period. The horizontal red dotted line in both first plots indicates the raw value of the total relative NAVs as of the specified trading day. Mean values for each strategy appear in the subtitle below the figure. The lowest third graph displays the evolution of the cumulative value of the relative mean square errors for strategy 1 (in black), strategy 2 (in red), strategy 3 (in blue), strategy 4 (in green), strategy 5 (in orange), we have the portfolio consisting of the* 35 *funds with the largest AUM and the proposed balanced sampling portfolio is in black line. These results are derived from the large sample context, combining the three HF categories under study to obtain an universe of 343 funds. The number of funds in each tracker is* $n = 35$.

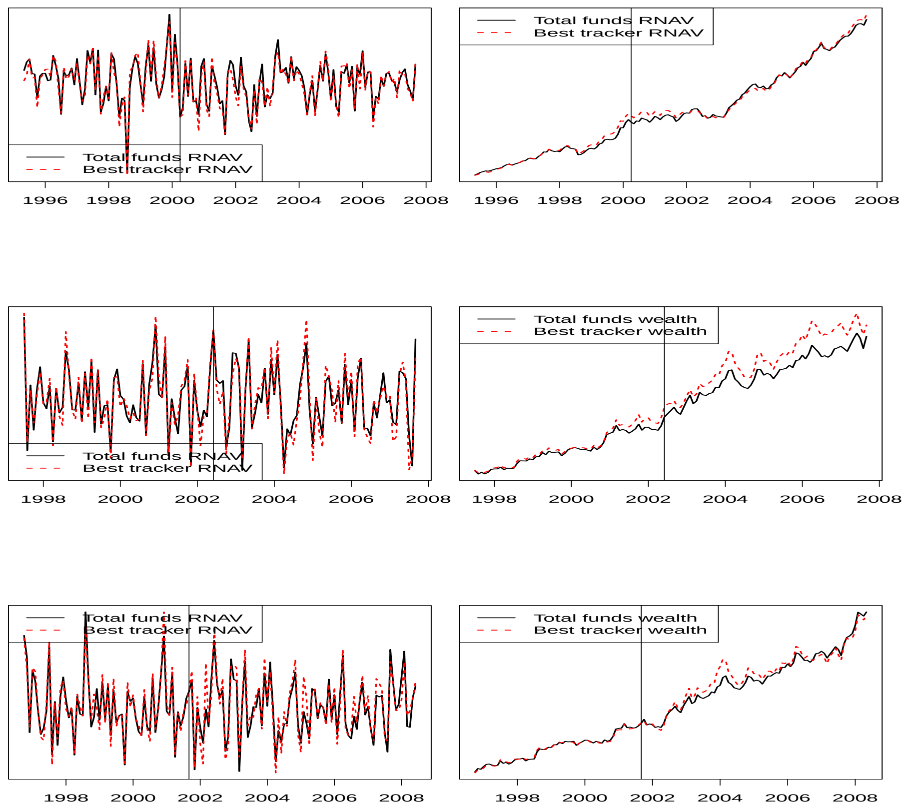

Figure 5.4 – *The evolution of relative NAVs (first-column) and wealth achieved (second-column) from the best tracking portfolio out of the 10,000 selections, as derived by the balanced sampling model. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. The tracker is composed of* $n = 10$ *funds.*

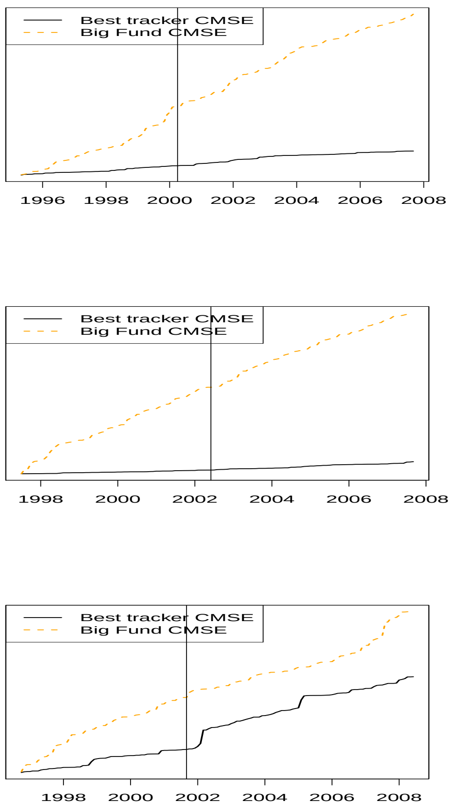

Figure 5.5 – *Cumulative value of the relative mean square errors. Portfolio consisting of the 10 funds with the largest AUM (in orange), balanced sampling (in black). From the up to down, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of* $n = 10$ *funds.*

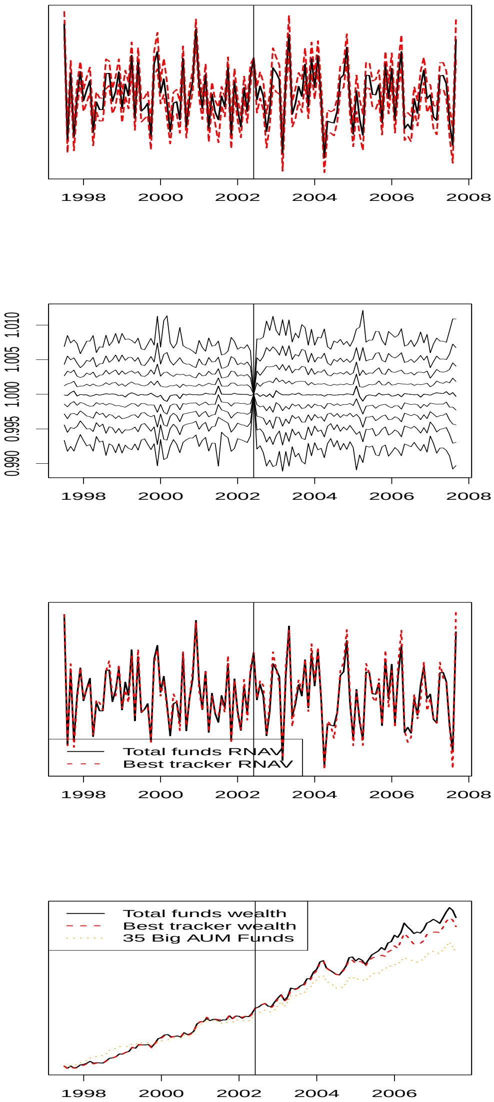

Figure 5.6 – *From upper to lower plot, the total HF relative NAVs (black line) embedded in the 5th and 95th percentiles (red line) of the corresponding values; Deciles of the ratios from* 10,000 *selected portfolios; The minimum variance portfolio; The wealth achieved investing one dollar in the best tracking portfolio and the portfolio consisting of all the funds in HF universe. These results are derived from the large sample context, combining the three HF categories under study to obtain an universe of 343 funds. The number of funds in each tracker is* $n = 35$.

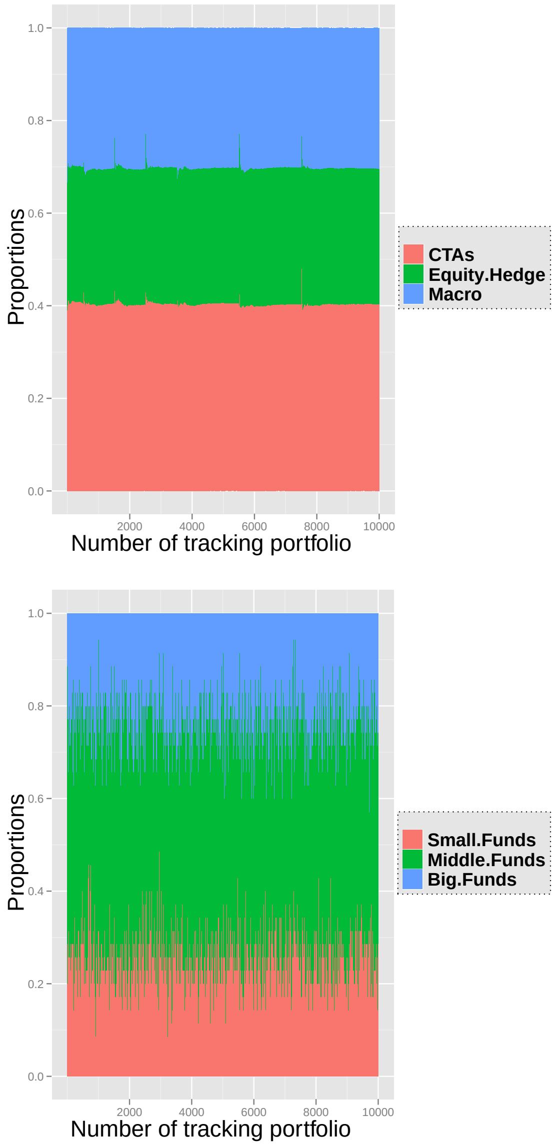

Figure 5.8 – *The upper plot shows the decomposition of each tracking portfolio relative to various HF category. The lower plot highlight the tracking portfolio distribution according to fund AUM. Fund classifications by size are performed according to the first and third quartile of the population AUM in millions of US\$. Small Funds: AUM* ≤ 10.5 *(first quartile), Middle Funds: AUM* $\in (10.5, 154)$*, Big Funds: AUM* ≥ 154 *(third quartile). These results are derived from the large sample context, combining the three HF categories under study to obtain an universe of 343 funds. The number of funds in each tracker is* $n = 35$.

Conclusion and further research

This thesis focuses on the quantitative modelling of one of the fastest growing business models in the financial industry, namely the hedge funds. As mentioned in different sections of this manuscript, the hedge fund or alternative investment sector is characterized by highly complex investment universe and is a key player in the globalized financial arena. Hedge fund investments are dynamic, market-adapting, and highly heterogeneous. A better understanding of hedge fund practices and performance dynamics is therefore important for both investors and regulators. Since August 1998 Russian debt crisis and the Long Term Capital Asset Management event, research on hedge fund has grown remarkably. The majority of the analysis is implemented under the usual assumption of parametric models: stationarity, and/or linearity. The current thesis capitalizes on the rich empirical literature to build a framework more in line with the statistical properties of hedge fund returns. It takes advantage of three main statistical frameworks: Nonstationarity, Nonparametric, and Survey sampling methodology.

The first part of the thesis is dedicated to the issue of hedge fund replication. The empirical literature on the subject is profuse. Recent advances highlight the benefits of macro risk factors in driving hedge fund returns. The framework outlined in Chapter 2 differs from other recent studies by focusing on the dynamic and adapting nature of hedge fund investment strategies. Unlike the purely linear hedge fund replication methodology, our *performance-adaptive* approach allows the detection of dynamic patterns between index returns and market risk factors. Our results draw a more nuanced picture of the predictive ability of commonly used market risk factors of hedge fund returns. While the factor model specification fails to predict hedge fund returns, our *performance-adaptive* approach is capable of explaining and predicting the aggregated style-specific abnormal returns ($alphas$). This conclusion indicates that hedge fund return dynamics are more complex than what is generally assumed, and that a simple linear factor model can not satisfactory describe the data generating process. Future research should move from a macro factor analysis to a micro level, and be able to tackle deeply the heterogeneity associated to the hedge fund industry.

In Chapter 3, we present an extension of the distribution free method, implemented in the previous chapter, for constructing a multistrategy portfolio of hedge funds. The popularity of hedge fund investing takes its grounds on the increased returns and diversification benefits offered to investors. One of the optimal ways to gain access to hedge fund double-digit returns is through funds-of-hedge funds. Chapter 3 documents a robust strategy, originated from the sequential investment theory, that relinquishes the usual global stationary assumption and mean-variance criterion for a locally stationary and nonparametric paradigm. The approach is flexible and allows the manager to reach the maximum rate of growth of capital, with minimal assumption on the underlying process. The main task consists of individuating returns that are 'similar' according to a defined metric. Our results show the superiority of locally stationary models in portfolio allocation. A possible drawback of our approach, the dynamics of the constructed portfolios are more volatile. A direction for further research is to investigate the possibility of incorporating a risk measure in the optimization methodology without loosing its distribution free spirit.

The last part of this thesis deals with applications of the sampling methodology. We intend to construct a bridge between sampling and finance, through the market tracking issue. Motivated by the development of index-related products, we implement a consistent tracking methodology. Selecting of subset of assets to reconstruct the dynamic changes of a large sample is a natural sampling problem. Nevertheless, there are several sampling designs with distinct characteristics and this document focuses on balanced sampling. In Chapter4, we track the total market capitalization of the equity S&P 500 market and Chapter5 addresses the indexing of three main hedge fund categories. Units are selected randomly with unequal and strictly positive inclusion probability to all constituents. The diversification property is automatically ensured by the sampling design. We show that in some cases, it is convenient to overrepresent some assets in the sample. In both cases equity and hedge fund, the tracking ability of the proposed sampling design is strong in-sample and out-of-sample. The constructed tracking portfolios perform better during the recent financial crisis. With a small number of assets (less than 10%), we are able to reconstruct the time evolving changes of a broad market. This approach is novel in the empirical financial literature and opens interesting new avenues for future research (for example, the determination of optimal size constraint parameter to achieve the best tracking portfolio).

A common feature of the approaches proposed in this thesis is the reliance on minimal assumptions on the data generating process when building consistent empirical models. Last but not least, it is worth emphasizing once more that in finance, a solid quantitative framework has to be associated to intelligent qualitative judgements.

Bibliography

Ackermann, C., R. McEnally, and D. Ravenscraft (1999): "The performance of hedge funds: Risk, return, and incentives," *Journal of Finance*, 54, 833–874. (Cited page 23.)

Agarwal, V., N. Daniel, and N. Naik (2009): "Role of managerial incentives and discretion in hedge fund performance," *Journal of Finance*, 64, 2221–2256. (Cited pages 24 and 93.)

Agarwal, V., W. H. Fung, Y. C. Loon, and N. Y. Naik (2009): "Risk and return in convertible arbitrage: Evidence from the convertible bond market," available at `http://ssrn.com/abstract= 885945`. (Cited page 39.)

Alexander, C. (1999): "Optimal hedging using cointegration," *Transactions of the Royal Society Series A*, 357, 2039–2058. (Cited page 156.)

Alexander, C., and A. Dimitriu (2005a): "Hedge Fund Index Tracking. in Hedge Funds: Insights in Performance Measurement, Risk Analysis, and Portfolio Allocation, G .N. Gregoriou, G. Hübner, N. Papageorgiou, and F. Rouah (ed.), pp. 165-179," *John Wiley & Sons, Inc*. (Cited page 176.)

——— (2005b): "Indexing and statistical arbitrage: tracking error or cointegration?," *Journal of Portfolio Management*, 31, 50–63. (Cited page 130.)

Algoet, P. H. (1992): "Universal prediction schemes," *The Annals of Probabilities*, 20, 901–941. (Cited pages 82 and 85.)

Algoet, P. H., and T. M. Cover (1988): "Asymptotic optimality and asymptotic equipartition properties of log-optimum investment," *The Annals of Probabilities*, 16, 876–898. (Cited pages 82 and 84.)

Amenc, N., L. Martellini, J.-C. Meyfredi, and V. Ziemann (2009): "Passive hedge fund-replication-Beyong the linear case," *European Financial Management*, 16, 191–210. (Cited pages 39 and 43.)

Amin, G. S., and H. M. Kat (2003): "Hedge fund performance 1990-2000: Do the money machines really add value?," *Journal of Financial and Quantitative Analysis*, 38, 251–274. (Cited page 92.)

Ang, A., and N. P. Bollen (2010): "Locked up by a lockup: Valuing liquidity as a real option," *Financial Management*, 39, 1069–1096. (Cited page 92.)

Aragon, G. O. (2007): "Share restrictions and asset pricing: Evidence from hedge fund industry," *Journal of Financial Economics*, 83, 33–58. (Cited page 93.)

Asness, C., R. Krail, and J. Liew (2001): "Do Hedge Funds Hedge?," *Journal of Portfolio Management*, pp. 6–19. (Cited page 24.)

Bajgrowicz, P., and O. Scaillet (2012): "Technical trading revisited: false discoveries, persistence tests, and transaction costs," *Journal of Financial Economics*, forthcoming, 1–58. (Cited page 150.)

Barras, L., O. Scaillet, and R. Wermers (2010): "False discoveries in mutual fund performance: Measuring luck in estimated alphas," *Journal of Finance*, 65, 179–216. (Cited page 128.)

Biau, G., K. Bleakley, L. Györfi, and G. Ottucsak (2009): "Nonparametric sequential prediction of time series," *The Journal of Nonparametric Statistics*, 3, 297–317. (Cited pages 46 and 87.)

Bioy, H., B. Johnson, and G. Rose (2011): *Synthetic ETFs Under the Microscope*, Morningstar. (Cited page 129.)

Bollen, N., and V. Pool (2009): "Do Hedge Fund Managers Misreport Returns? Evidence from the Pooled Distribution," *The Journal of Finance*, pp. 2257–2288. (Cited page 23.)

Canakgoz, N. A., and J. E. Beasley (2009): "Mixed-integer programming approaches for index tracking and enhanced indexation," *European Journal of Operational Research*, 196, 384–399. (Cited page 129.)

Chauvet, G., and Y. Tillé (2006): "A fast algorithm of balanced sampling," *Journal of Computational Statistics*, 21, 9–31. (Cited page 139.)

Chen, D.-S., R. G. Batson, and Y. Dang (2010): *Applied Interger Programming: Modeling and Solution*. Wiley, New Jersey. (Cited page 129.)

De Los Rios, A. D., and R. Garcia (2008): "Assessing and valuing the non-linear structure of hedge fund returns," available at `http://ssrn.com/abstract= 890739`. (Cited page 39.)

DeMiguel, V., G. L., and R. Uppal (2009): "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?," *Review of Financial Studies*, pp. 1915–1953. (Cited page 128.)

Derigs, U., and N. H. Nickel (2003): "Meta-heuristic based decision support for portfolio optimization with a case study on tracking error minimization in passive portfolio management," *OR Spectrum*, 25, 345–378. (Cited pages 130 and 156.)

Deville, J.-C., and Y. Tillé (2004): "Efficient balanced sampling: The cube method," *Biometrika*, 91, 893–912. (Cited pages 138, 139, 159, and 162.)

Duarte, J., F. Longstaff, and Y. Fan (2007): "Risk and return in fixed-income arbitrage: Nickels in front of steamroller?," *Review of Financial Studies*, 20, 769–811. (Cited page 39.)

Engle, R. F. (2002): "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business and Economic Statistics*, 20, 339–350. (Cited page 80.)

Focardi, S. M., and F. J. Fabozzi (2004): "A methodology for index tracking based on time-series clustering," *Quantitative Finance*, 4, 417–425. (Cited pages 130 and 156.)

Francesco, C., and M. Marcellino (2006): "Factor based index tracking," *Journal of Banking and Finance*, 30, 2215–2233. (Cited page 130.)

Fugazza, C., M. Guidolin, and G. Nicodano (2010): "1/N and Long Run Optimal Portfolios: Results for Mixed Asset Menus," Federal Reserve Bank of St, Louis, Working Paper No. 2010-003A. (Cited page 128.)

Fung, W., and D. Hsieh (2000): "Performance characteristics of Hedge Funds Commodities Funds: Natural versus spurious biases," *Journal of Financial and Quantitative Analysis*, pp. 291–307. (Cited page 23.)

Fung, W., and D. A. Hsieh (2001): "The risk in hedge fund strategies: Theory and evidence from trend followers," *Review of Financial Studies*, 14, 313–341. (Cited pages 39 and 49.)

——— (2004a): "Extracting portable alphas from equity long-short hedge funds," *Financial of Investment Management*, 2, 57–75. (Cited pages 39 and 49.)

——— (2004b): "Hedge fund benchmarks: A risk-based approach," *Financial Analysts Journal*, 60, 65–80. (Cited page 49.)

——— (2006): "Hedge funds: An industry in its adolescence," *Federal Reserve Bank of Atlanta-Economic Review*, 4th Quater. (Cited page 56.)

——— (2007): "Hedge fund replication strategies: implications for investors and regulators," *Banque de France-Financial Stability Review-Special issue on hedge fund*, 10, 55–64. (Cited page 50.)

Getmansky, M., A. Lo, and I. Makarov (2004): "An Econometric Analysis of Serial Correlation and Illiquidity in Hedge-Fund Returns," *Journal of Financial Economics*, pp. 529–610. (Cited page 24.)

Graham, J. R., and C. R. Harvey (1997): "Grading the performance of market timing newsletters," *Financial Analysts Journal*, 53, 54–66. (Cited page 90.)

Györfi, L., and G. Lugosi (2001): "Strategies for sequential prediction of stationary time series. in Modeling Uncertainty: An Examination of its Theory, Methods and Applications, M. Dror, P. L. Ecuyer, F. Szidarovszky(Eds.), pp. 225-248," *Kluwer Academic Publishers*. (Cited page 45.)

Györfi, L., G. Lugosi, and F. Udina (2006): "Nonparametric kernel-based sequential investment strategies," *Mathematical Finance*, 16, 337–357. (Cited pages 21, 37, 40, 41, 42, 81, 83, 84, and 85.)

Györfi, L., and D. Shäfer (2003): "Nonparametric prediction. in Advances in Learning Theory: Methods, Models and Applications (NATO Science Series: Computer & Systems Sciences), J. Suykens, G. Horvàth , S. Basu, C. Micchelli, J. Vandevalle (Eds.), pp. 341-356," *Amsterdam: IOS Press*. (Cited page 82.)

György, O., and I. Vajda (2007): "An asymptotic analysis of the mean-variance portfolio selection," *Statistics & Decisions*, 99, 63–86. (Cited pages 82, 83, 88, and 102.)

Han, Y. (2006): "Asset allocation with a high dimensional latent factor stochastic volatility model," *Review Financial Studies*, 19, 237–271. (Cited page 90.)

Harvey, C. R., and A. Siddique (2000): "Conditional skewness in asset pricing tests," *Journal of Finance*, 55, 1263–1295. (Cited page 80.)

Hasanhodzic, J., and A. W. Lo (2007): "Can hedge-fund returns be replicated?: The linear case," *Journal of Investment Management*, 5, 5–45. (Cited pages 55, 56, 81, 84, 86, and 93.)

hedgeindex (2012): "Dow Jones Credit Suisse hedge fund index category @ONLINE," . (Cited page 32.)

Horvitz, D. G., and D. J. Thompson (1952): "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685. (Cited pages 133, 138, and 161.)

Hsu, D. A., R. B. Miller, and D. W. Wichern (1974): "On the Stable Paretian Behavior of Stock-Market Prices," *Journal of the American Statistical Association*, 69(345), pp. 108–113. (Cited page 80.)

Jondeau, E., and M. Rockinger (2011): "On the importance of time-variability in higher moments for asset allocation," *Journal of Financial Econometrics*, 10, 84–123. (Cited page 80.)

Kat, H. M., and H. P. Palaro (2006): "Replicating hedge fund returns using futures," Cass Business School, City University, London. (Cited page 39.)

Kelly, J. L. (1956): "A New interpretation of information rate," *Bell System Technical Journal*, 35, 917–926. (Cited page 21.)

Kirby, C., and B. Ostdiek (2009): "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," *Management Science*, 55, 798–812. (Cited page 128.)

——— (2012): "It's All in the Timing: Simple Active Portfolio Strategies that Outperform Naive Diversification," *Journal of Financial and Quantitative Analysis*, pp. 437–467. (Cited page 128.)

Klaus, B., and B. Rzepkowski (2009): "Risk spillover among hedge funds. The role of redemptions and fund failures," *European Central Bank*, Working paper series No. 1112. (Cited page 92.)

Lhabitant, F.-S. (2004): *Hedge funds-Quantitative insights*. John Wiley & Sons Ltd, West Sussex PO19 8SQ, England. (Cited pages 92 and 101.)

Lhabitant, F.-S., and M. Learned (2003): "Hedge fund diversification: How much is enough?," *Journal of Alternative Investments*, 5, 23–49. (Cited pages 92 and 156.)

Liang, B. (1999): "On the performance of hedge fund," *Financial Analysts Journal*, 55, 72–85. (Cited page 93.)

——— (2000): "Hedge Funds: The living and the dead," *Journal of Financial and Quantitative Analysis*, pp. 309–326. (Cited page 23.)

Malkiel, B. G. (1995): "Returns from investing in equity mutual funds 1971 to 1991," *Journal of Finance*, 50, 549–572. (Cited page 128.)

Maringer, D., and O. Oyewumi (2007): "Index tracking with constrained portfolios," *Intelligent Systems in Accounting, Finance and Management*, 15, 57–71. (Cited page 130.)

Markowitz, H. M. (1952): "Portfolio selection," *Journal of Finance*, 7(1), 17–91. (Cited page 88.)

——— (1987): *Mean-variance analysis in portfolio choice and capital markets*. Blackwell, Oxford. (Cited pages 129 and 156.)

Meade, N., and G. R. Salkin (1989): "Index funds-construction and performance measurement," *Journal of the Operational Research Society*, 40, 871–879. (Cited page 131.)

Mitchell, M., and T. Pulvino (2001): "Characteristics of risk and return in risk arbitrage," *The Journal of Finance*, 56, 2135–2175. (Cited pages 39 and 50.)

Nedyalkova, D., and Y. Tillé (2008): "Optimal sampling and estimation strategies under linear model," *Biometrika*, 95, 521–537. (Cited page 134.)

Perron, P., and Z. Qu (2010): "Long memory and level shifts in the volatility of stock market return indices," *Journal of Business and Economic Statistics*, 28, 275–290. (Cited page 80.)

Rahl, L. (2003): *Hedge funds risk transparency*. Risk Waters Group, London. (Cited page 20.)

Reid, B., S. Collins, S. Holden, and S. Steenstra (2013): *2013 Investment Company Fact Book*. 53rd Edition. Investment Company Institute. (Cited page 128.)

Roncalli, T., and J. Teïletche (2008): "An alternative approach to alternative beta," *Journal of Financial transformation*, 24, 43–53. (Cited pages 39 and 43.)

Roncalli, T., and G. Weisang (2011): "Tracking problems, hedge fund replication and alternative beta," *Journal of Financial transformation*, 31, 19–29. (Cited page 39.)

Rossbach, P., and D. Karlow (2011): "The stability of traditional measures of index tracking quality," Working Paper Series 164, Frankfurt School of Finance and Management. (Cited pages 129, 150, and 151.)

Ruiz-Torrubiano, R., and A. Suárez (2009): "A hybrid optimization approach to index tracking," *Annals of Operations Research*, 166, 57–71. (Cited page 130.)

Shapcott, J. (1992): "Index tracking: genetic algorithms for investment portfolio selection," Technical report, EPCC-SS92-24. University of Edinburgh, Parallel Computer Centre. (Cited page 130.)

Sharpe, W. (1992): "Asset allocation: Management style and performance measurement," *The Journal of Portfolio Management*, 18, 7–19. (Cited page 49.)

Sharpe, W. F. (1991): "The Arithmetic of Active Management," *Financial Analysts Journal*, 47, 7–9. (Cited page 128.)

Spellucci, P. (1997): "RDONLP2," As provided by Ryuichi Tamura, available at `spellucci@mathematik.tu-darmstadt.de`, accessed 31 May 2010. (Cited page 95.)

Starica, C., and C. Granger (2005): "Nonstationarities in stocks returns," *The Review of Economics and Statistics*, 87, 503–522. (Cited page 80.)

Switzer, L. N., and A. Omelchak (2011): "Are there benefits from dynamic asset allocation strategies across hedge funds?," *Journal of Portfolio Management*, 37, 116–120. (Cited page 80.)

Tafin Djoko, D. (2013a): "Performance based diversification or how to create a multistrategy hedge fund?," *Journal of Applied Finance & Banking*, pp. 271–325. (Cited page 79.)

——— (2013b): "Tracking Hedge Fund Performance: A Balanced Sampling Strategy," *Journal of Statistical and Econometric Methods*, pp. 81–114. (Cited page 155.)

Tafin Djoko, D., and C. Starica (2013): "Hedge Fund Replication: A Dynamic Performnance-Adaptive Local Linear Regression Approach," Submitted for publication. (Cited page 37.)

Tafin Djoko, D., and Y. Tillé (2013): "Selection of Balanced Portfolios to Track the Main Properties of a Large Market," *Forthcoming in Quantitative Finance*. (Cited pages 127, 157, and 160.)

The Economist, January 26 (2013): "Exchange-traded funds, Twenty years young, The anniversary of a successful financial innovation," Retrieved January 26, 2013, from http://www.economist.com. (Cited pages 20 and 128.)

Tillé, Y. (2006): *Sampling Algorithms*. Springer, New York. (Cited pages 132, 148, 159, 162, and 171.)

Tu, J., and G. Zhou (2011): "Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies," *Journal of Financial Economics*, pp. 204–215. (Cited page 128.)

Weide, R. V. D. (2002): "GO-GARCH: A multivariate generalized orthogonal garch model," *Journal of Applied Econometrics*, 17, 549–564. (Cited pages 82, 89, and 103.)