ETC FACULTATEA DE ELECTRONICĂ ŞI TELECOMUNICAŢII

enst Bretagne

HEDGE FUNDS- NEED FOR NEW METHODS FOR CLUSTERING AND FILTERING

Daria Ioana Bâtiu

Lucrarea corespunde prin conţinut şi prin formă de prezentare
cerinţelor pentru obţinerea titlului de

INGINER cu DIPLOMĂ

în specialitatea ELECTRONICĂ APLICATĂ / TELECOMUNICAŢII
la Facultatea de Electronică şi Telecomunicaţii
a Universităţii *Politehnica* din Timişoara

2007

(Semnătura)

Prof. Dr. ing. Ioan Nafornită
Conducător ştiinţific

ENST Bretagne

CS 83818 - 29238 Brest CEDEX 3

Hedge funds ~ Need of new methods for clustering and filtering ~

September 2007

Author:
Daria Ioana BATIU

Guidance: Rim ENNAJAR-SAYADI
Jean Marc LE CAILLEC
Gilles COPPIN

Collaboration with: Arnaud CLEMENT-GRANDCOURT
Française des placements investissement
Paris, France

Contents

Chapter 1

Introduction

Summary

1.1. Introduction

The hedge fund universe consists of a great variety of completely different investment and trading strategies. Despite having some common features (e.g. an unregulated organizational structure, flexible investment strategies, sophisticated investors, etc.), hedge funds remain an extremely diverse asset class. A consistent classification system is important for numerous reasons – it will help improve investment-choices of market participants, and funds of funds will refer to it in the construction of their portfolio to avoid undiversified exposures. A grouping of funds based on return characteristics can furthermore help evaluate the discriminatory power of different styles. In this context, a consistent classification system contributes to an improved performance attribution.

1.2. Definition of the research question

The purpose of this paper is to shed some light on the stylistic differences across hedge funds by analysing their evolution using the filtering approach (in order to eliminate the noise existent in the different hedge fund series), and to try to find a structure in this collection of unlabeled data (clustering). The great variety of hedge funds, as well as their non-linear manifestation during time poses both a challenge and an opportunity to their analysis. The challenge is comprehending and bench marking managers whose operations are essentially opaque, whose instruments vary widely, and who in many cases eschew predictable passive factor exposures. The opportunity lies in the diversification that the varieties of hedge funds present.

We ask a few simple questions.

- First, in light of the extraordinary variety of hedge fund strategies, are there a few basic styles that they pursue?
- Second, are these styles meaningful to investors – that is, do they explain differences in performance?
- Third, are there any significant trends in these styles that investors and analysts should know about?

The research question is therefore: "Can filtering analysis help to better define the different strategies of hedge funds, to distinguish some basic styles and rules in their evolution in time and to classify them with a unified approach?" The following sub questions, defining the research problem, should be articulated.

- First, how and what type of filtering-analysis should be chosen, in order to give an appropriate solution in estimating the hidden hedge funds time-series state in a way that minimizes the error? Here the specificity of the hedge funds time series must be taken into account.

- Second, what type of clustering-method should be chosen, in order to better distinguish and classify the different hedge funds evolutions in time, based on the existent measurements (the TASS an HFR hedge funds databases are used during the experiments)?

- Third, can an alternative, filter-based model and an appropriate clustering method be formulated for hedge funds time evolution estimation and prediction and does this model perform well compared to existing competing models?

- Fourth, can wavelet-analysis decompose the hedge funds returns time-series into multiple levels, such that each level captures specific useful information? Can this analysis help to tailor risk by finding an appropriate resolution for each time horizon, through use of a multi-level decomposition? The answer to this question will remain as a future research.

1.3. Methodology

In order to operationalize the research problem, I took the following steps. As a first step I consulted the available literature on hedge funds – their characteristics, benefits, risks and the available strategies – in order to understand the general idea and its role in finance and economics.

As a second step I made an overview of the most commonly available literature for the problem of estimating the hidden states of a system in an optimal and consistent fashion, given a set of noisy or incomplete observations (the case of hedge funds returns time-series). I found that the optimal solution to this problem is given by the recursive Bayesian estimation algorithm which recursively updates the posterior density of the system state. I dove into the Gaussian Approximate Bayesian Estimation theory to understand the mathematics behind it. I started my study with the simple Kalman Filter. In order to acquire intuition for the theory, I worked out numerical examples, first with simple models, later on with more complex ones. During my experiments I found that, unfortunately for most real-world problems, the optimal Bayesian recursion is intractable and approximate solutions must be used. The most simple of this category is the Extended Kalman filter (EKF). Unfortunately, the EKF is based on a sub-optimal implementation of the recursive Bayesian estimation framework applied to Gaussian random variables. This can seriously affect the accuracy or even lead to divergence of any inference system that is based on the EKF or that uses the EKF as a component part. So, I continued my research in the literature, to find a better solution, in parallel with working out numerical examples. Great benefits for our non-linear estimation problem can be obtained by algorithmic alternatives to the EKF, based on derivativeless statistical linearization, called Sigma-Point Kalman Filters which are deeply explained in Chapter 3.

The third step was to make an overview and comparison of the available software for the Gaussian Approximate Bayesian Estimation – Kalman Filter Framework. Using this overview, I chose an appropriate software package in order to acquire intuition for the theory and to conduct experiments. The software package used is the ReBEL toolkit, which is a Matlab toolbox designed to facilitate the sequential estimation in general state space models. ReBEL is developed and maintained by Rudolph van der Merwe.

As a fourth step I continued my study with another category of filters, the Particle filters. Instead of a Gaussian approximation of the aposteriori state, they use Markov-Chain Monte-Carlo simulations and are therefore also used in the case of Non-Gaussian noises. The time evolution of hedge funds return is a non-linear time series, usually affected by non-gaussian noise, so a suitable approach was needed. Whereas the standard EKF and the sigma-point filters, discussed above, make a Gaussian assumption to simplify the optimal recursive Bayesian estimation, particle filters make no assumptions on the form of the probability densities in question, that is full nonlinear, non-Gaussian estimation. I studied the mathematics and the behaviour of these filters; then I searched an appropriate software package in order to acquire intuition for the theory and to conduct experiments.

As a fifth step, in order to fulfil my research I made an overview of the most commonly available wavelet literature and dove into wavelet theory to understand the mathematics behind it. Wavelet transformation is a powerful signal processing tool that is well suited to process non-linear and non-stationary dynamical processes. Due to the joint time-frequency nature of the wavelets, wavelet analysis is able to yield features that describe properties of a time series, both at various locations and at varying time granularities.

As a sixth step I consulted the available literature on clustering methods – their characteristics, requirements and the available algorithms – in order to understand the general idea and their role and use in finance and economics. A *cluster* is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. By using clustering procedures, we are able to identify and separate the hedge funds returns time-series into different groups based on similarities found in both the underlying trends and the localized transient patterns of these time series. An accurate clustering analysis, together with the good choice of attributes can bring some light of the extraordinary variety of hedge fund strategies.

The seventh step was to make an overview and comparison of the available software for the K-means, fuzzy C-means clustering algorithms. Using this overview, I chose an appropriate software package in order to acquire intuition for the theory and to conduct experiments. The software package used is the Fuzzy Clustering and Data Analysis Toolbox, which is a collection of Matlab functions. Its purpose is to divide a given data set into subsets, under different initial assumptions.

The eighth step was to conduct simulations and experiments with the proposed models on artificially generated data and real-life data from the TASS and HFR hedge fund databases. I studied the results and checked whether the models gave an accurate estimation of the non-linear and non-stationary dynamical processes of hedge funds returns time-series. The simulations were divided in two categories: simple clustering (testing different methods and obtaining the optimal number of clusters) and filtering (testing the different algorithms presented in the theory). The filtering + clustering approach (in order to analyse how the filtering affects the data, if this process affects the initial clusters and if it diminishes the error) will be treated in a future work. Each simulation and validity test is repeated a significant number of times in order to get a reliable notion of the performance.

As a final step I compared the results obtained from the simulations, in order to make a classification, and to propose a different way of analysing the hedge fund returns time-series, using the filtering and clustering approach.

The writing of the final thesis was a recurrent process. It entailed the performing of a literature study to gain a general understanding, experimenting with the theory to get intuition for the theory and reporting the results. This process was for me the logical order to structure this complex problem into smaller solvable problems.

1.4. Structure of the paper

To support the methodology as presented, this paper is structured as follows and is shown in the next figure: Chapter 1, this chapter, is the introduction. Chapter 2 explains the characteristics and strategies of the hedge funds and the various studies existent in the literature. In chapter 3 the approximate Bayesian estimation theory is systematically investigated. Following the simplest case, the celebrated Kalman filter is briefly derived, followed by the discussion of optimal nonlinear filtering. Chapter 4 discusses a popular numerical approximation technique - Monte Carlo approximation and sequential sampling method - which results in various forms of particle filters. Chapter 5 explained the fundamentals of wavelet analysis and the basis for time-scale decomposition. In chapter 6 the cluster analysis techniques are presented; here were given different methods of clustering, which will be used in Chapter 7 during the experiments, in order to compare, group and find structures in the various models for the hedge funds returns. Chapter 7 sets up experiments, first using artificially generated data and then real-life continuously compounded returns obtained from the TASS and HFR databases. Chapter 8 concludes this paper. Chapter 9 is the bibliography. Chapter 10 corresponds to the appendix.

Chapter 2

Hedge funds

Summary

2.1. Definition

In financial terminology, the meaning of hedge is protecting oneself against unfavourable changes in prices. The hedge funds have known a powerful growth in recent years, and became more and more popular. This is due to their ability to outperform the overall market through individual stock and security selection and by taking market neutral positions in order to protect financial capital in times of market volatility.

The term *hedge fund* dates back to the first such fund founded by Alfred Winslow Jones in 1949. His innovation consisted in combining a leveraged long stock position with a portfolio of short stocks in an investment fund, thus some of the market risk was hedged. Many hedge fund characteristics have changed since then, but also many important features have remained the same. Nowadays the hedge funds are spread in many places around the world, not only in the U.S.

The hedge fund universe consists of a great variety of completely different investment and trading strategies. Despite having some common features (like flexible investment strategies, sophisticated investors, unregulated organizational structure), hedge funds remain an extremely diverse asset class (Ackermann, 1999), so it is difficult to give them a unique definition. However, I tried to synthesize here what a hedge fund represents and its main characteristics:

A **hedge fund** is usually used by wealthy individuals and institutions, which is allowed to use aggressive strategies that are unavailable to mutual funds, including selling short, leverage (borrowing), program trading, swaps, arbitrage, and derivatives. It that can take both long and short positions, buy and sell undervalued securities, trade options or bonds, and invest in almost any opportunity in any market where it foresees gains at *reduced risk*. The primary aim of most hedge funds is to ***reduce volatility and risk*** while attempting to ***preserve capital and deliver positive returns under all market conditions***.

Success is measured by tracking the *"**absolute return**"* of a fund, which means that the return is not related to the overall direction of any particular investment market. So, unlike conventional share market funds, for example, hedge funds can profit irrespective of the overall market direction.

2.2. Characteristics, benefits and risks

Hedge funds are exempt from many of the rules and regulations governing other mutual funds, which allow them to accomplish aggressive investing goals. Legally, hedge funds are most often set up as private investment partnerships that are open to a limited number of investors (no more than 100 investors per fund) and require a very large initial minimum investment (ranging anywhere from $250,000 to over $1 million). Investments in hedge funds are illiquid as they often require investors keep their money in the fund for a minimum period of at least one year.

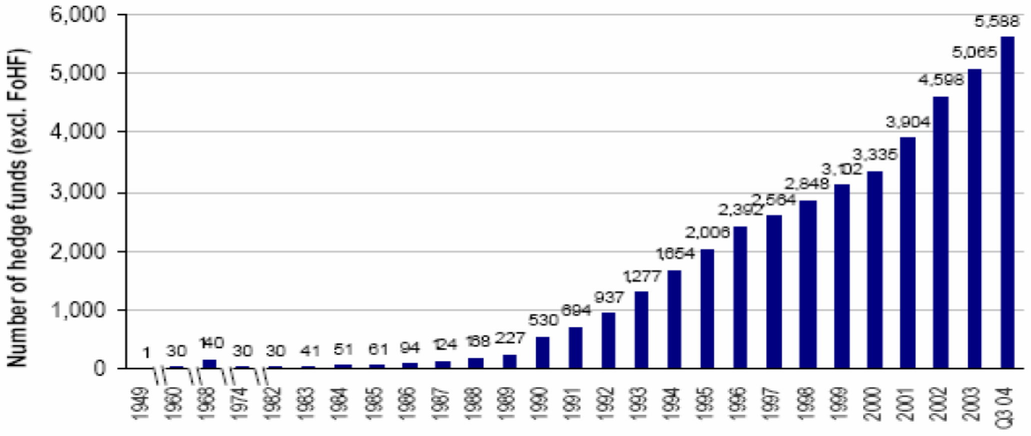
Other characterististics of hedge funds are presented above:

- Hedge funds utilize a variety of financial instruments to reduce risk, enhance returns and minimize the correlation with equity and bond markets. Many hedge funds are flexible in their investment options (can use short selling, leverage, derivatives such as puts, calls, options, futures, etc.).
- Hedge funds vary enormously in terms of investment returns, volatility and risk. Many, but not all, hedge fund strategies tend to hedge against downturns in the markets being traded.
- Many hedge funds have the ability to deliver non-market correlated returns.
- Many hedge funds have as an objective consistency of returns and capital preservation rather than magnitude of returns.
- Most hedge funds are managed by experienced investment professionals who are generally disciplined and diligent.

- Pension funds, insurance companies, private banks and high net worth individuals and families invest in hedge funds to minimize overall portfolio volatility and enhance returns.
- Most hedge fund managers are highly specialized and trade only within their area of expertise and competitive advantage.
- Hedge funds benefit by heavily weighting hedge fund managers' remuneration towards performance incentives, thus attracting the best brains in the investment business. In addition, hedge fund managers usually have their own money invested in their fund.

The growing evolution of the number of hedge funds between 1949 - 2004 is presented in Fig.1.

Chart 10: Number of hedge funds (1949-2004)



Source: UBS (1990-2004 from Hedge Fund Research, 1982-1989 from Quellos, prior to 1982 from Elden [2001] and Caldwell [1995].

Fig.1: Number of hedge funds (1949 – 2004)

Benefits

- Many hedge fund strategies have the ability to generate positive returns in both rising and falling equity and bond markets.
- The inclusion of hedge funds in a balanced portfolio reduces overall portfolio risk and volatility whilst increasing returns and diversification.
- The huge variety of hedge fund investment styles – many uncorrelated with each other – provides investors with a wide choice of hedge fund strategies to meet their stated investment objectives.
- Academic research suggests hedge funds have higher returns and lower overall risk than traditional investment funds.
- Hedge funds provide an ideal long-term investment solution, eliminating the need to correctly time entry and exit from markets.

- Adding hedge funds to an investment portfolio provides diversification not otherwise available in traditional investing.

Risk in hedge funds

Hedge funds make *uncorrelated* returns because they take *different* risks. Analyzing their risks is not just a good idea; it is the beginning of any investment operation.

The following is a list (not an exhaustive one) of risks categories specific to hedge funds:

- Lack of transparency. Hedge funds are businesses, and as such, they often choose not to disclose their most precious asset: their strategy. It is typically impossible to get a hedge fund to report the positions it holds in its investment portfolio. However, they do offer a subscription document or offering memorandum, which is a legal document that binds the manager to a certain set of activities, and therefore the manager has limits to what she can do. Audited financial statements are typically available, and they should be consulted prior to any investment.

- Fraud. We mentioned above that the manger is bound to limit their activity to certain legitimate activities. Fraud will occur when they don't. Fraud can also occur when they misquote performance or valuations.

- Counterparty risk. Although not specific to hedge funds, they are especially sensitive to this risk type because of the unregulated and specialized nature of their transactions. Counterparty risk (or credit risk) refers to losses that the fund can incur into when the counterparty to some of its financial transactions does not honour their obligations (default). This term also refers to situations when, without default or bank-ruptcy, the counterparty undergoes a credit downgrade, hence affecting the market value of the securities in the fund.

- Portfolio liquidity and redemption orders. Hedge funds often restrict fund redemptions to quarterly terms (or more), and usually with some advance notice. Liquidation orders, therefore, can take time to process. Moreover, funds can choose to suspend redemption orders when they estimate that liquidation would be detrimental to the remaining investors in the fund: a hedge fund forced to sell securities to meet redemption orders is an easy prey to its counterparties or competitors in the financial markets.

- Capacity risk. Hedge funds make money on fees; while management fees are considered generous by investment standards (1% or more), it is with the performance fees that most funds make their money. Hence, they are encouraged to close the fund to new investments if they see no new opportunities for return and sense that they have reached capacity. Those who don't limit their fund-raising activity may raise assets beyond their natural capacity, which may lead to a decrease in their future returns.

- Style drift. Individual hedge funds are normally selected to be part of an investment portfolio for a good reason, which is usually the particular trading style they employ; when a fund changes its style and adopts another one, it may create imbalances inside the portfolio; if sufficiently many funds undergo style drift over a period of time, the risk pattern of the portfolio can be changed drastically and give rise to unintended risk concentration.

- Data. Imagine a hedge fund that trades between New York and Tokyo. When it calculates the daily value of the assets, does it use NY closing time of 4 pm EST? Or Tokyo closing time? Or does it use NY close for some positions and Tokyo for others? Questions as simple as this (and of course much more complex) can introduce huge differences in valuation of the firm's assets and hence of the price settlement when new investors join the fund or the price paid to investors exiting the fund. It can also lead to a smoothing effect in the fund's performance numbers.

- Legal risk; in 2003, after some illegal activity involving mutual funds and some hedge funds, market timing activities (one small but profitable hedge fund style) became under general legal scrutiny. Investors in certain funds rushed for redemptions, driving the value of the assets remaining in the fund down dramatically. Changes in law affect all activities in life and in particular in the investment sector, but when they mix with highly complex, illiquid, investments such as the ones inside a hedge fund, the result can be dramatic. Tax laws can be particularly sensitive for certain hedge fund activity.

And of course, hedge funds are exposed to investment risks in general:

- Market Risk. The risk in reducing the value of the portfolio's positions due to changes in markets.

- Credit Risk. The risk in reducing the value of the portfolio's assets due to changes in the credit quality of the counterparties.

- Liquidity Risk. The risk of losses because of travel-time delays of assets.

- Common factor risk: industry specific, geographical risk, etc.

- Operational Risk. Internal systems, people, physical events.

- Corporate event risk: earnings revisions, mergers, etc.

- Model risk.

- Legal and Regulatory Risk.

2.3. Strategies

It is important to understand the differences between the various hedge fund strategies because **all hedge funds are not the same** -- *investment returns, volatility and risk* **vary enormously among the different hedge fund strategies**. Some of them, which are not correlated to equity markets, are able to deliver consistent returns with extremely low risk of loss, while others may be as or more volatile than mutual funds. A wide range of hedging strategies is available to hedge funds. For example:

- Selling short - selling shares without owning them, hoping to buy them back at a future date at a lower price in the expectation that their price will drop.
- Using arbitrage - seeking to exploit pricing inefficiencies between related securities - for example, can be long convertible bonds and short the underlying issuer's equity.
- Trading options or derivatives - contracts whose values are based on the performance of any underlying financial asset, index or other investment.
- Investing in anticipation of a specific event - merger transaction, hostile takeover, spin-off, exiting of bankruptcy proceedings, etc.
- Investing in deeply discounted securities - of companies about to enter or exit financial distress or bankruptcy, often below liquidation value.
- Many of the strategies used by hedge funds benefit from being non-correlated to the direction of equity markets
- Distressed securities funds look for shares or fixed-interest investments issued by companies which have gone into bankruptcy or are otherwise in trouble, in the hope that the investment will gain in value when the company emerges from its difficulties.
- Macro funds look for global trends, in the hope of profiting from changes in interest rates or currency values.
- Special situations funds react to news - good or bad - which is expected to result in a rapid change in the value of shares or fixed-interest investments.

There are many other strategies, and some managers combine more than one.

Hedge Fund Styles

- **Very high risk strategies**

Emerging Markets: Invests in equity or debt of emerging (less mature) markets that tend to have higher inflation, volatile growth and the potential for significant future growth. Examples include Brazil, China, India, and Russia. Short selling is not permitted in many emerging markets, and, therefore, effective hedging is often not available. This strategy is defined purely by geography; the manager may invest in any asset class (e.g., equities, bonds, currencies) and may construct his portfolio on any basis (e.g. value, growth, and arbitrage). *Expected Volatility:* **Very High**

Short Selling: In order to short sell, the manager borrows securities from a prime broker and immediately sells them on the market. The manager later repurchases these securities, ideally at a lower price than he sold them for, and returns them to the broker. In this way, the manager is able to profit from a fall in a security's value. Short selling managers typically target overvalued stocks that are characterized by prices they believe are too high given the fundamentals of the underlying companies. It is often used as a hedge to offset long-only portfolios and by those who feel the market is approaching a bearish cycle. *Expected Volatility:* **Very High**

Macro: Aims to profit from changes in global economies, typically brought about by shifts in government policy that impact interest rates, in turn affecting currency, stock, and bond markets. Rather than considering how individual corporate securities may fare, the manager constructs his portfolio based on a top-down view of global economic trends, considering factors such as interest rates, economic policies, inflation, etc and seeks to profit from changes in the value of entire asset classes. For example, the manager may hold long positions in the U.S. dollar and Japanese equity indices while shorting the euro and U.S. treasury bills. Uses leverage and derivatives to accentuate the impact of market moves. The leveraged directional investments tend to make the largest impact on performance. *Expected Volatility:* **Very High**

- ## High risk strategies

Aggressive Growth: A primarily equity-based strategy whereby the manager invests in companies, with smaller or micro capitalization stocks, characterized by low or no dividends, but experiencing or expected to experience strong growth in earnings per share. The manager may consider a company's business fundamentals when investing and/or may invest in stocks on the basis of technical factors, such as stock price momentum. Managers employing this strategy generally utilize short selling to some degree, although a substantial long bias is common. This includes sector specialist funds such as technology, banking, or biotechnology. *Expected Volatility:* **High**

Market Timing: The manager attempts to predict the short-term movements of various markets (or market segments) and based on those predictions, moves capital from one asset class to another in order to capture market gains and avoid market losses. While a variety of asset classes may be used, the most typical ones are mutual funds and money market funds. Market timing managers focusing on these asset classes are sometimes referred to as mutual fund switchers. Unpredictability of market movements and the difficulty of timing entry and exit from markets add to the volatility of this strategy. *Expected Volatility:* **High**

- ## Moderate risk strategies

Special Situations: The manager invests both long and short, in stocks and/or bonds which are expected to change in price over a short period of time due to an unusual event. Examples of event-driven situations are mergers, hostile takeovers, reorganizations, or leveraged buyouts. It may involve simultaneous purchase of stock in companies being acquired, and the sale of stock in its acquirer, hoping to profit from the spread between the current market price and the ultimate purchase price of the company. Generally the results do not dependent on the direction of market. *Expected Volatility:* **Moderate**

Value: A primarily equity-based strategy whereby the manager invests in securities perceived to be selling at deep discounts to their intrinsic or potential worth. The manager takes long positions in stocks that he believes are undervalued, i.e. the stock price is low given company fundamentals such as high earnings per share, good cash flow, strong management, etc. Possible reasons that a stock may sell at a perceived discount could be that the company is out of favour with investors or that its future prospects are not correctly judged by Wall Street analysts. Securities may be out of favour or under-followed by analysts. Long-term holding, patience, and strong discipline are often required, until the ultimate value is recognized by the market. The manager can take short positions in stocks he believes are overvalued. *Expected Volatility:* **Low - Moderate**

Funds of Hedge Funds: The manager invests in other hedge funds (or managed accounts programs) rather than directly investing in securities such as stocks, bonds, etc. These underlying hedge funds may follow a variety of investment strategies or may all employ similar approaches. Because investor capital is diversified among a number of different hedge fund managers, funds of funds generally exhibit lower risk than do single-manager hedge funds. Funds of funds are also referred to as multi-manager funds. It’s a diversified portfolio of generally uncorrelated hedge funds and it’s a preferred investment of choice for many pension funds, endowments, insurance companies, private banks and high-net-worth families and individuals. Returns, risk, and volatility can be controlled by the mix of underlying strategies and funds. *Expected Volatility:* **Low - Moderate - High**

- ## Variable risk strategies

Opportunistic: Rather than consistently selecting securities according to the same strategy, the manager's investment theme changes from strategy to strategy as opportunities arise to profit; sudden price changes are often caused by an interim earnings disappointment, hostile bids, and other event-driven opportunities. Characteristics of the portfolio, such as asset classes, market capitalization, etc., are likely to vary significantly from time to time. The manager may also employ a combination of different approaches at a given time. *Expected Volatility:* **Variable**

Multi Strategy: The manager typically utilizes many specific, pre-determined investment strategies, e.g., Value, Aggressive Growth, and Special Situations in order to better diversify their portfolio and/or to more fully use their range of portfolio management skills and philosophies and also in order to realize short or long term gains. This style of investing allows the manager to overweight or underweight different strategies to best capitalize on current investment opportunities. Although the relative weighting of the chosen strategies may vary over time, each strategy plays a significant role in portfolio construction. *Expected Volatility:* **Variable**

- **Low risk strategies**

Distressed Securities: The manager invests in the debt and/or equity of companies having financial difficulty. Such companies are generally in bankruptcy reorganization or are emerging from bankruptcy or appear likely to declare bankruptcy in the near future. Because of their distressed situations, the manager can buy such companies' securities at deeply discounted prices. The manager stands to make money on such a position should the company successfully reorganize and return to profitability. Also, the manager could realize a profit if the company is liquidated, provided that the manager had bought senior debt in the company for less than its liquidation value. "Orphan equity" issued by newly reorganized companies emerging from bankruptcy may be included in the manager's portfolio. The manager may take short positions in companies whose situations he deems will worsen, rather than improve, in the short term. Generally the results do not dependent on the direction of market. *Expected Volatility:* **Low - Moderate**

Income: Invests with primary focus on yield or current income rather than solely on capital gains, though it may also utilize leverage to buy bonds and (sometimes) fixed income derivatives in order to profit from principal appreciation and interest income. Other strategies (e.g. distressed securities, market neutral arbitrage, and macro) may heavily involve fixed-income securities trading as well. *Expected Volatility:* **Low**

Market Neutral - Securities Hedging: The manager invests similar amounts of capital in securities both long and short, generally in the same sectors of the market, maintaining a portfolio with low net market exposure. Long positions are taken in securities expected to rise in value while short positions are taken in securities expected to fall in value. Due to the portfolio's low net market exposure, performance is insulated from market volatility. Market risk is greatly reduced, but effective stock analysis and stock picking is essential to obtaining meaningful results. Leverage may be used to enhance returns. It sometimes uses market index futures to hedge out systematic (market) risk. *Expected Volatility:* **Low**

Market Neutral - Arbitrage: The manager seeks to exploit specific inefficiencies in the market by trading a carefully hedged portfolio of offsetting long and short positions. By pairing individual long positions with related short positions, market-level risk is greatly reduced, resulting in a portfolio that bears a low correlation to the market. For example, long convertible bonds and short underlying issuer's equity. For example, can be long convertible bonds and short the underlying issuer's equity. It may also use futures to hedge out interest rate risk. These relative value strategies include fixed income arbitrage, mortgage backed securities, capital structure arbitrage, and closed-end fund arbitrage. *Expected Volatility:* **Low**

2.4. Hedge Fund Database Providers and Classification

Four primary databases are popular among researchers and in the investment industry. Providers of these databases offer different services to the industry. The Zurich Capital Markets (WCM/Hedge) database provides a comprehensive coverage of global hedge funds. The Hedge Fund Research (HFR) database contains more equity-based hedge funds. TASS is the information and research subsidiary of Credit Swiss First Boston Tremont Advisers.

Various database providers classify hedge funds, but in different ways. All the four databases have their own indices based on the categories in the database. The index composition is also different for different databases. Hedge fund categories are based on the self-reported style classification of hedge fund managers that are listed in a particular database. None of the database provides information on the complete hedge fund universe. The databases differ in the definition of the 'hedge fund'. For example, TASS is the only database that includes the managed future funds. Unlike hedge funds, managed future funds limit their activities to the futures market.

Following issues are observed about the performance data for various databases:

- A major limitation of most hedge funds databases is that they typically have data only on funds still in existence or that are new and growing.
- Most hedge fund indices do not include performance of closed funds.
- Only those funds that choose to report are included in the database. Not much can be done with this issue due to the industry structure. ZCM/Hedge and TASS have historical performances of all funds that are included in their database. Historical performances are not included (no backfilling) in index construction, but are available for fund analysis.
- HFR, ZCM/Hedge and VanHedge have all inclusive selection criteria; they include all funds in their database that classify them as hedge funds. TASS has its own selection criteria.
- The classification method varies across databases making them difficult to compare.

Hedge fund managers employ a diverse array of strategies. The database providers classify hedge funds based on the voluntary information that they can collect from the hedge fund managers. Style definitions and the number of categories of hedge funds differ among the database providers. The classification of hedge funds by various database providers is briefly described here.

2.4.1. ZCM/Hedge Classification

The ZCM/Hedge database classifies hedge funds into four general classes and ten broad categories of investment styles, as reported by the managers of the hedge fund. The classes are ‘onshore’ hedge fund (HF-US), ‘offshore’ hedge fund (HF-NON), ‘onshore’ fund-of-funds (FOF-US), and ‘offshore’ fund-of-funds (FOF-NON). Some of the categories have further sub-classification. ZCM/Hedge database categories are shown in Fig. 2.

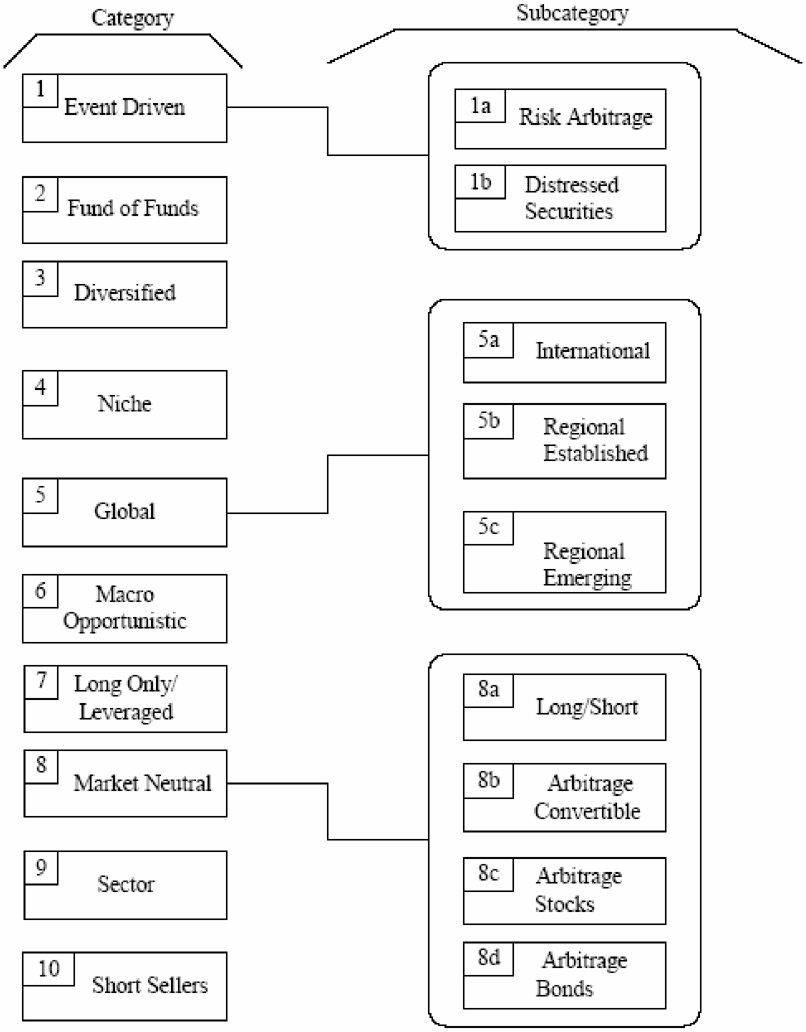


Fig.2. ZCM/Hedge classification of hedge funds

2.4.2. HFR Classification

Hedge Fund Research (HFR) has twenty-six categories of hedge funds. Some of these categories are merely a type of financial instrument or a geographic area for investment. This classification can be reorganized into eleven categories as shown in Fig. 3. Some of the categories have further classification.

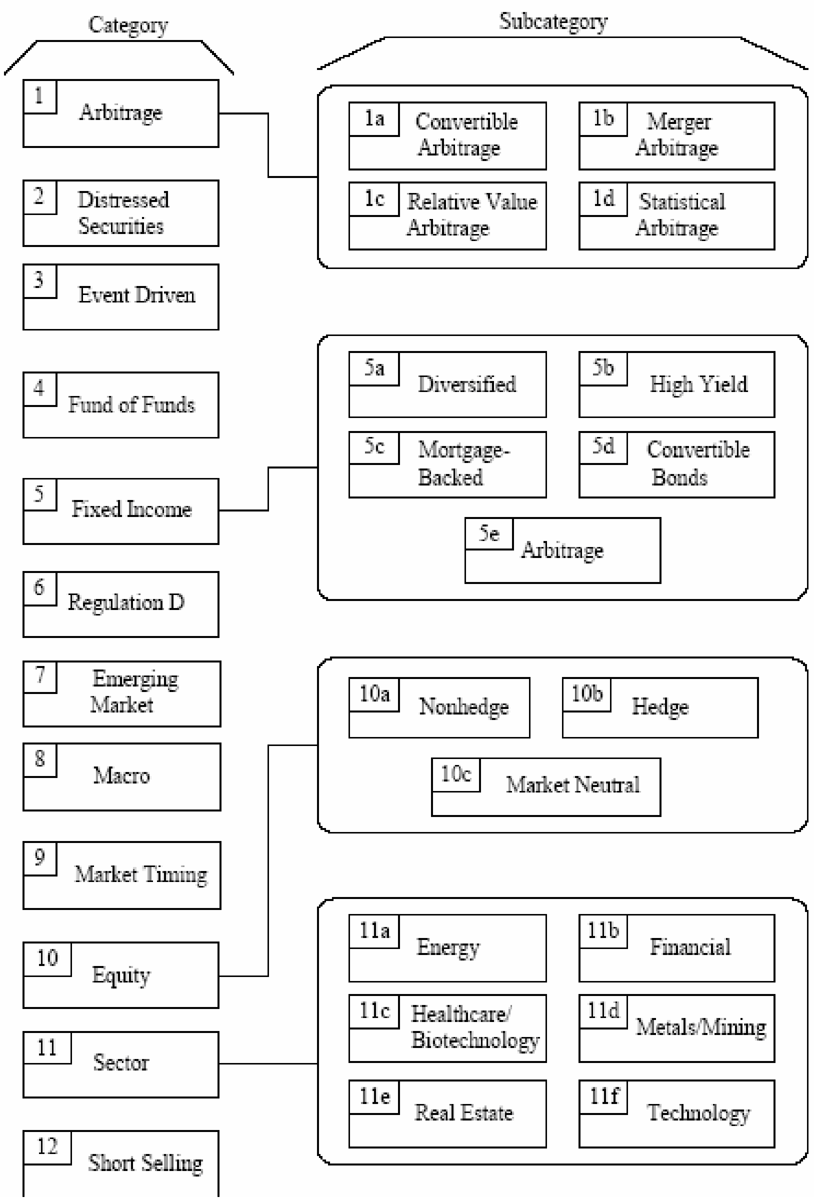


Fig.3. HFR classification of hedge funds

2.4.3. TASS Classification

TASS is the information and research subsidiary of Credit Suisse First Boston Tremont Advisers. It has nine categories of hedge funds, classified based on the investment styles of hedge fund managers. Fig. 4 shows the classification of TASS database. For more information see Appendix.

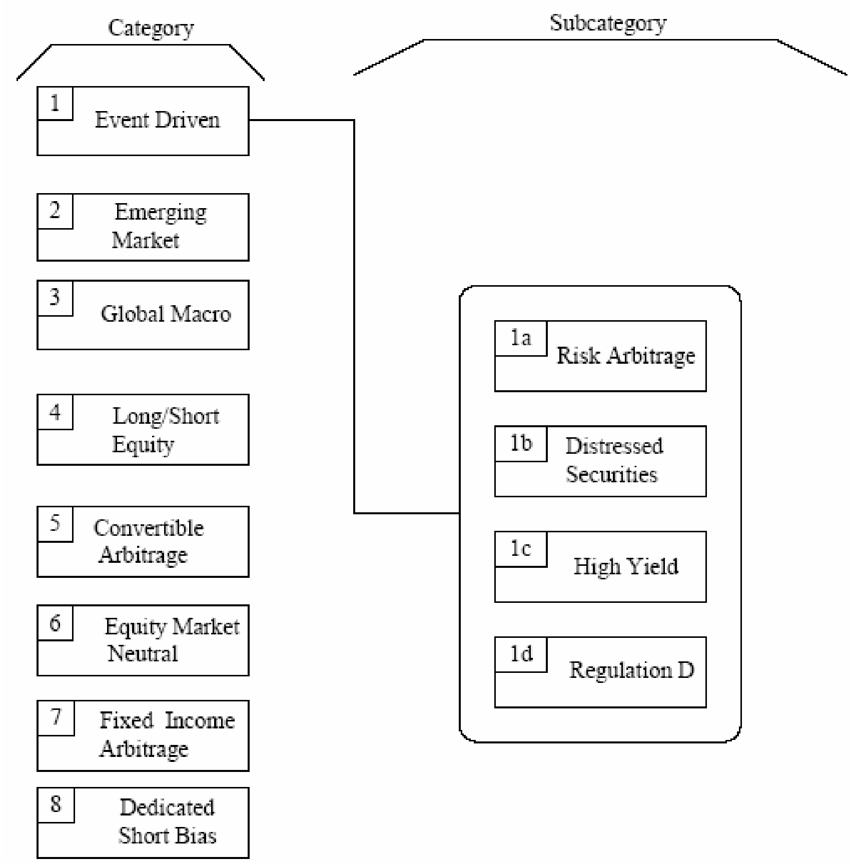


Fig.4. TASS classification of alternative investments

2.4.4. VanHedge Classification

VanHedge maintains an extensive database of hedge funds. It provides consultancy and detailed generic performance data on hedge fund styles. VanHedge database can be organized into thirteen categories and five subcategories, as shown in Fig. 5.

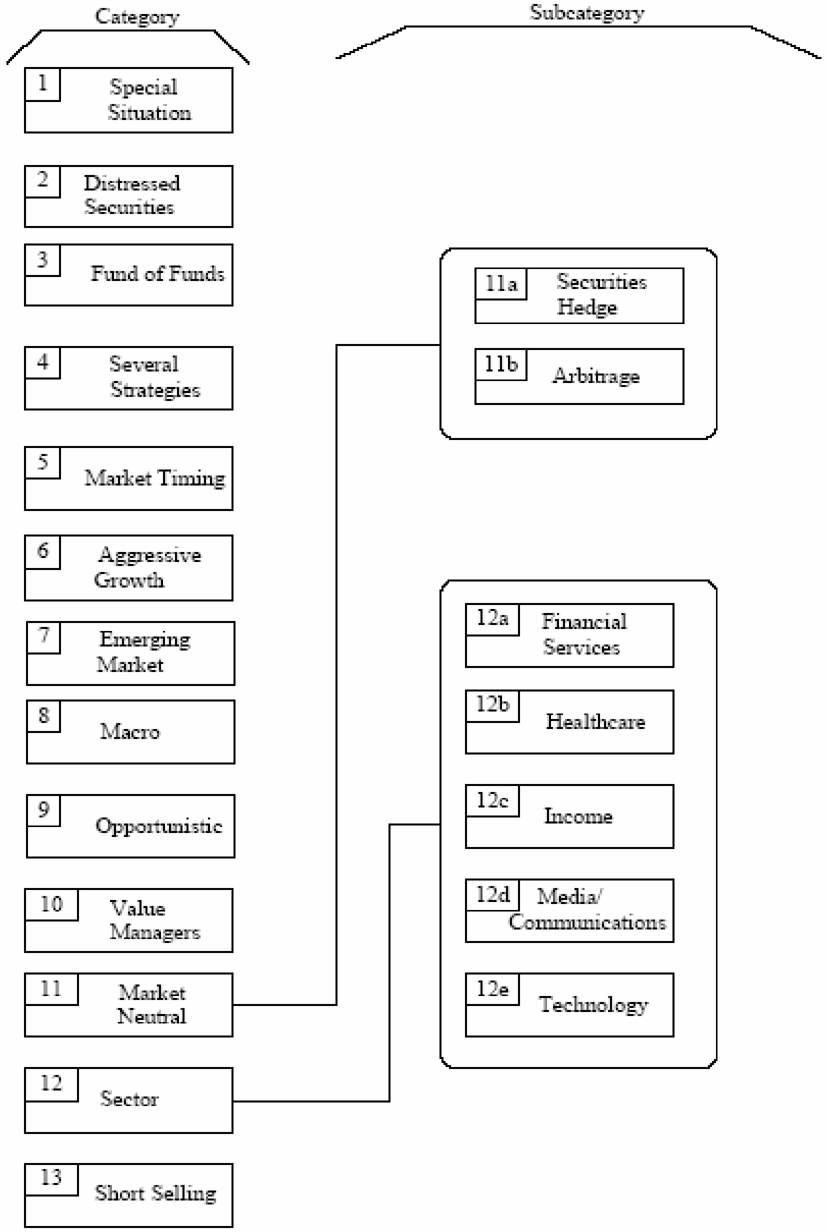


Fig.4. VanHedge classification of hedge funds

2.5. Alternative Classification Requirement

There exists a lot of variation in the definitions, calculation methodologies, assumptions, and data employed by the different managers and databases. It is necessary to benchmark hedge fund manager practices relative to their peers as hedge funds follow diverse strategies. The various classification schemes and multiple peer groups may vary depending on the strategies employed by the manager. It is important to clearly identify a peer for the various hedge fund strategies. This may not be an easy task since hedge fund managers refrain from disclosure.

Hedge funds are primarily distinguished by their use of short-selling, leverage, derivatives and portfolio concentration. Hedge fund manager refrains from disclosure for two reasons: They are not permitted by regulation to advertise to the public. Secondly, the proprietary nature of the traders may result in herding. Hedge fund managers profit by identifying arbitrage opportunities. These opportunities are based on very slim price differentials, but the manager hopes to profit by properly timing his trade and through portfolio concentration.

There is a need for an 'alternative approach' to hedge fund classification given the lack of 'pure' hedge fund types that exist in the industry. The hedge fund literature shows an almost complete reliance on the existing hedge fund classification. Performance comparison of various hedge funds with the existing hedge fund indices return data is not appropriate as a particular hedge fund could be classified in two or more classes depending on the database. Table 1 compares the classification of ZCM/Hedge, HFR, TASS and VanHedge databases.

Item	***ZCM/Hedge***	***HFR***	***TASS***	***VanHedge***
1a	Event Driven: Risk Arbitrage	Event Driven: Merger Arbitrage	Event Driven: Risk Arbitrage	Special Situation
1b	Event Driven: Distressed Securities	Distressed Securities	Event Driven: Distressed Securities	Distressed Securities
2	Fund of funds	Fund of funds	None	Fund of funds
3	Diversified	Fixed Income Diversified	None	Several Strategies
4	Niche	Fixed Income: High Yield Regulation D	Event Driven: Regulation D Event Driven: High Yield	None

5	Global	Emerging Markets	Emerging Markets	Emerging Markets
6	Macro Opportunistic	Macro Market Timing Relative Value Arbitrage Statistical Arbitrage	Global Macro	Opportunistic Value Managers
7	Long Only/ Leveraged	Equity Non-hedge	None	None
8a	Market Neutral: Long/Short	Equity Hedge	Long/Short Equity	Market Neutral: Securities Hedge
8b	Market Neutral: Arbitrage Convertible	Convertible Arbitrage	Convertible Arbitrage	Market Neutral: Arbitrage
8c	Market Neutral: Arbitrage Stock	Equity Market Neutral	Equity Market Neutral	Market Neutral: Arbitrage
8d	Market Neutral: Arbitrage Bond	Fixed Income Arbitrage	Fixed Income Arbitrage	Market Neutral: Arbitrage
9	Sector	Sector: Energy Sector: Financial Sector: Health Care/Biotechnology Sector: Metals/Mining Sector: Real Estate Sector: Technology	None	Financial Services Health Care Income Media/Communications Technology
10	Short Selling	Short Selling	Dedicated Short Bias	Short Selling

Table1. Comparison of ZCM/Hedge, HFR, TASS and VanHedge Classifications

It appears from Table 1 that research on hedge fund performance may produce different results, based on the database used. There seems to be no common comparison basis for the existing literature on hedge funds. The disparity that is observed in the number produced between different organisations measuring hedge funds performance could be attributed to the varied classification of hedge funds. Goldman Sachs & Co. & FRM describes various methods used by hedge fund managers. The description of various hedge fund styles certainly does not cover all the permutations, but provides an overall idea of the various strategies used by the managers. Table 2 compares the different segments of hedge fund in terms of investment strategy, use of leverage and risk control.

Segment	***Investment Strategy***	***Use of leverage***	***Risk Control***
Market Neutral or Relative Value	Seek out basic mispricings between securities.	Aggressively use leverage to capitalize on otherwise small pricing differences.	Broad market risk is eliminated completely to capitalize on relative mispricing.
Event Driven	Seek out valuation disparities produced by corporate events that are less dependent on overall stock market gains.	Use of leverage varies from situation to situation, but in general leverage is used conservatively.	Portfolio is diversified among a number of position s to reduce the impact of any single position that does not work out as anticipated. Hedge against market risk by purchasing index put options and short selling.
Long/Short	Seek out mispriced securities based on the business prospects of the firms, using both long and short positions.	Historically, they maintain leverage positions ranging from slightly short to 100% long.	It is often accomplished through market neutral positions. Some accomplish this within industry groups and employ greater amount of leverage.
Tactical Trading: Systematic Managers	Seek out mispriced securities using statistical analysis, which is applied to historical data.	A high degree of leverage is used to capitalize on small, but statistically significant, return opportunities.	Risk control is vital. Managers eliminate all risk except the risk that their models indicate as profitable.

Tactical Trading: Discretionary Managers	Seek out mispricing in global currency, stock and bound market using derivatives.	Use of leverage is kept to a minimum due to lack of risk control.	Risk control is difficult to achieve because of low correlation between currencies and indices within a market.
Fund of Funds	Seek out diversification by investing in a variety of hedge funds.	Not applicable.	Risk control is achieved through diversification of hedge funds.

Table2. Comparison of different core segments of hedge fund investments

2.4. Literature review

The study of hedge funds is a recent phenomenon. Most of the literature is less than a decade old and can be divided into three main categories: performance attribution (modelling returns), performance evaluation and characteristics and impacts on the financial markets.

2.4.1. Performance attribution (modelling returns)

This analysis attempts to find the factors affecting the hedge fund return. When *modelling hedge funds performance as a group*, no distinction is made between the different categories. Ackerman et al. (1999) isolated hedge funds characteristics that explain the performance and volatility of hedge funds and founds that incentive fees can be used to explain risk-adjusted performance.

Different managers and databases classify hedge funds differently. One particular hedge fund could be grouped under one category, based on a strategy, in one database, whereas the same hedge fund would be listed under a different category in some other database. The studies made *extract strategies from observed returns* and try to reclassify hedge funds based on observed return characteristics. Fung and Hsieh (1997) develop an integrated framework for analysing traditional managers with absolute return targets (mutual funds) as well as alternative managers with absolute return targets (hedge funds) and find that Sharpe's style regression is not appropriate for discovering performance attributes, but non-linear look-back straddles could be a good approach.

Another way of *modelling a particular hedge fund strategy* consists on taking the database classification as given and studying only one strategy at a a time. Fung and Heish (2001) modelled the nonlinear relationship between style factors and the markets where the hedge funds trade. They found that the trend-following strategies can be modelled using look-back straddles.

2.4.2. Performance evaluation

This analysis compares the return earned on a hedge fund with the return earned on some other standard investment asset. Research in this area can be divided into three main groups:

An investment *benchmark* is a passive representation of a manager's investment process. This represents the prominent financial characteristics that the investment would exhibit in the absence of active investment judgement. Key benchmarking research supports the fact that hedge funds outperform mutual funds, even on a risk adjusted basis. Ackermann et al. (1999) find that hedge funds are more volatile than both mutual and market indices; Agarwal and Naik (2000) analyse the degree of out-performance of hedge funds strategies over a portofolio of passive strategies and find that hedge fund managers exhibit superior market timing and security selection ability; Fung and Hsiesh (2001) show that hedge fund categories should be reclassified into key hedge-fund styles, that is pairs of strategy and location.

The second aspect of *performance evaluation, persistence* deals with the examination of whether the hedge fund managers demonstrate persistence in their performance and how the survival rate affects performance persistence. Support for performance persistence within individual hedge fund strategies using both parametric and non-parametric methods and also using a multi-period framework was found by Agarwal and Naik (2000); their results indicate that that the extent of persistence decreases as the return interval increases. Capocci, Corhay and Hübner (2004) showed that if persistence is present in hedge fund returns, excess return creation was present in most of the cases and there was a clear proof of persistency in hedge fund returns. Edwards and Caglayan (2001) proved that only three hedge fund strategies (Market Neutral, Event Driven and Macro) provide protection to investors when stock markets head south. Ennis' and Sebastian's (2003) research showed that hedge funds did not provide investor protection after the market downturn of March 2000

The third area of evaluation deals with *performance in a portfolio context* that is the diversification benefits of including hedge funds in a traditional portfolio of stocks and bonds. Some researchers (Agarwal and Naik (2000), Lamm and Ghaleg-Harter (2000)) support the diversification effects of hedge funds.

2.4.3. Characteristics and impact on financial market

This area starts with general characteristics and progresses to performance attributes. The researchers study the *characteristics of the hedge fund industry*, including the fee structure, data conditioning biases, and the risk/return characteristic of various hedge fund strategies. Returns are summarized in Fung and Hsiesh (1999), who studied the different types of biases presented in the hedge fund performance data, and suggested fund-of-funds as a better proxy for market portfolio based on the smaller impact of biases inherent to individual hedge fund returns. Brown et al. (2001) seeks whether hedge fund return variance depend upon the manager's performance, and finds that survival depends on volatility, age and both absolute and relative performance of the fund..

In the last area, researchers study the *role of hedge funds in the financial market crisis* and the implications for policy. Brown et al. (2000, 2001) tested the hypothesis that hedge funds were responsible for the 1997 crash in the Asian currencies, and found that hedge fund managers as a group did not cause the crash.

2.4.4. Other studies

Studies by Agarwal and Naik (2004) and Capocci (2002) confirm that hedge funds are significantly exposed to standard asset classes, although the exposure is small compared to the one found with mutual funds. Schneeweis and Spurgin and Amenc, Martellini and Vaissié (2002) show that hedge fund returns are not only exposed to the market risk, but that other risks like volatility risk, default risk or liquidity risk have to be considered.

Fung and Hsieh (1997) explain the limited explanatory power of standard asset classes with respect to hedge funds in another way: Contrary to mutual fund managers who have relative return targets, hedge fund managers have absolute return targets. While mutual fund managers follow generally buy-and-hold strategies with limited leverage and the only decision being where to invest, hedge fund managers can choose not only the location, but also the trading strategy of their investments. This leads to nonlinear option-like exposures of hedge funds to standard asset classes.

Starting from the findings of Fung and Hsieh (1997), Agarwal and Naik (2004) introduce a general asset class model containing buy-and-hold strategies (location factors) and passive option-based strategies (trading strategy factors). They find that the option-based strategies can explain a significant proportion of variation in hedge fund returns. Capocci (2002) compares several models, including the four-factor model by Carhart (1997) and an extension of the model by Agarwal and Naik (2004). The latter model proves to perform best in explaining the variation of hedge fund returns. The results of Agarwal and Naik (2004) and Capocci (2002) indicate that besides exposure to traditional asset classes, which is measured by traditional beta, hedge funds are exposed to additional alternative risk factors. The exposure to this alternative risk factors is measured by alternative beta. The results of Géhin and Vaissié (2005) confirm these findings. The authors observe that certain hedge fund styles have significant exposures to alternative risk factors.

McGuire, Remolona, and Tsatsaronis (2005) analyse the time-varying exposure of different hedge fund investment styles (directional, market-neutral and equity-focused) to various risk factors using moving window regression. They show that despite the homogeneity of hedge fund strategies, the exposure of the analysed strategies to some common risk factors, although similar between the strategies. Exposure to other market risk factors like fixed income is found to be homogeneous between strategies.

Time-varying exposure is the predominant characteristic in nowadays hedge fund's literature. The beta of a portfolio changes when either the underlying asset betas change and/or when the portfolio weights are changed. For hedge funds containing mainly stocks (long/short equity and equity market neutral), the presence of the first characteristic can be tested by looking at time-variation in stock betas. Research by Wells (1996) on Swedish stocks and Yao and Gao (2004) on Australian industry portfolios confirm that betas of stocks and stock portfolios are time-varying. Both authors use dynamic models and recursive filtering techniques. The presence of the second characteristic implies an active fund management. Fung and Hsieh (1997) argue that hedge fund managers reduce the correlation to asset class returns by adjusting the portfolio weights and hereby actively changing the exposure of their funds to asset classes. Mamaysky, Spiegel, and Zhang (2003) apply a model that assumes constant asset betas, but time-varying portfolio weights to analyse the dynamics of mutual funds alphas and betas. Contrary to the latter, Mamaysky, Spiegel, and Zhang (2003) assume that fund managers assign the weights of individual assets within a portfolio according to a lagged information variable which is latent and has no economic explanation. Based on this, they develop a model of the evolution of the portfolio's alpha and beta that requires no knowledge of the alphas and betas of the individual underlying stocks and their weights within the portfolio. The fund's dynamic alpha and beta are estimated using an Extended Kalman Filter in order to identify funds with substantially positive alphas. The authors show that using the model to select successful funds and to build a portfolio out of them leads to results that beat the market benchmark.

Considering the high fees of hedge funds, investors should not pay fees for static exposure which they can get cheaper by investing in mutual or index-tracking funds. It is therefore essential to know whether hedge funds still deliver absolute returns once the exposure to traditional and alternative risk factors has been neutralised. Fung and Hsieh (2004a) extract alternative alphas from diversified hedge fund portfolios after hedging away the exposure to S&P 500 and SMB. They show that these alternative alphas exhibit almost no sensitivity neither to the eight traditional asset-class indices5nor to the seven hedge fund risk factors, even under extreme market conditions, and are thus portable. The authors conclude that long/short equity hedge funds show significant absolute returns after taking into account alternative and traditional market risk and that these absolute returns are not only an effect of bull markets.

2.4.5. Traditional beta, alternative betas, and alpha

The specific exposures of hedge funds - Walter Géhin, Mathieu Vaissié (2005)

Two studies, by Watson Wyatt and UBS (2005), give a pessimistic view of the hedge fund industry's capacity to generate long-term returns, due to its increasing size. Unfortunately, these studies focus almost exclusively on alpha. Walter Géhin, Mathieu Vaissié (2005) showed the importance of considering not only the exposure to the market (the traditional beta), but also the other exposures (the alternative betas) to cover all the sources of hedge fund returns. They examined the real extent to which the variability and level of hedge fund returns are affected by (static) betas, dynamic betas (i.e. factor timing), and pure alpha (i.e. security selection).

Like a mutual fund, a hedge fund can be exposed to the traditional beta, in other words to the market risk consisting of unforeseeable variations in the prices of basic assets, stocks and bonds. However, a hedge fund is also exposed to **risk factors** which are different from those of long-only managers (so-called *alternative betas*):

- **volatility risk** - refers to unforeseeable variations in the variability of the prices
- **default risk -** is related to unforeseeable variations in the propensity of certain counterparties to no longer be able to respect their commitments
- **liquidity risk -** consists of unforeseeable variations in the capacity to move quantities of assets in a "reasonable" time scale at market prices

No directional strategies are generally considered to be non-exposed to market risk, but they are exposed to these risks. The hedge funds' exposure to risk factors by strategy, during the period January 1997 – December 2004 is presented in Table3.

	Implied volatility	Change in implied volatility	Value versus growth	Change in value versus growth	Small cap versus large cap	Change in small cap versus large cap	S&P 500	Term spread	Change in term spread	Credit spread	Change in credit spread	Lehman global bond index	Historical volatility in bond return	T-Bill 3 months
Equity market neutral		+			+	+	+	+		-	-	+		+
Fixed income arbitrage	+								+		-		-	
Convertible arbitrage		+					-			+	-	+		+
Merger arbitrage	+	+			+		+	-		-	-			+
Distressed securities	+	-			+		+				-			
Long/Short equity	+	+	-		+		+			-	-			
Global macro	+	+			+	-	+			-	-	+	-	
CTA macro	+	-					-				+	+	-	
Emerging markets	+	+			+		+				-			

Table3. Exposure to risk factors by strategy – from January 1997 to December 2004
[Source: Edhec Risk and Asset Management Research Centre]

Hedge fund returns are the addition of the:

- **traditional beta** - normal returns generated from exposure to rewarded market risk
- **alternative betas** - normal returns generated from exposure to other systematic risks – dynamic betas (factor timing)

- **alpha** - abnormal returns due to the manager's skill – security selection

$$\mathbf{R_{Hedge\ Funds} = \alpha + Traditional\ \beta + Alternative\ \beta s} \qquad (2.1)$$

Total alpha, i.e. the manager’s skill, is then the sum of pure alpha and dynamic betas. The authors then estimate the contribution of these components by applying a dynamic model with Kalman filtering (although they do not disclose the exact model they use). They find that while on average about half of the variability in returns comes from alpha (with 25% from pure alpha and 24% from dynamic beta), the contribution of static beta to performance is more than 99% on average, with a positive pure alpha of 4% and a negative dynamic beta of -3%. Despite these average values, in most cases the contribution of dynamic betas to total alpha is much more pronounced than the contribution of pure alpha, which indicates that total alpha is rather driven by factor timing than the selection skills and underlines the importance of correctly identifying time-varying betas. This argument however neglects the possibility that dynamic betas might not come from managers’ timing skills, but be a result of dynamic betas of the underlying assets. In most cases, the trend of value added through dynamic betas is much more pronounced than that of pure alpha, suggesting that the evolution of total alpha is mainly driven by the ongoing value added through dynamic betas. This result is particularly interesting as it suggests that hedge fund strategies' *alpha*, contrary to what has recently been said *is more limited by manager capacity than by market capacity*. This comes from the fact that the level of pure alpha primarily depends on the quantity of market opportunities that are available to hedge fund managers, while the level of value added through dynamic betas depends above all on the ability of hedge fund managers to time factors with success.

Taking the above decomposition of alpha and beta, Géhin and Vaissié (2005) conclude that hedge funds cannot simply be defined as absolute return vehicles that always deliver positive results without exposure to risks. They argue that exposures to traditional and alternative risk factors are undervalued compared to pure alpha although they are responsible for an overwhelming part of the hedge fund returns. Consequently, in the authors view hedge funds should be considered as an asset class which in combination with traditional investments offers beta diversification. They sustain that the beta-benefits of hedge fund investing are more convincing and attractive than the alpha-benefits and even though alpha does not appear to be seriously threatened by the increase in market participants, they are more prone to promoting the long-term diversification power presented by hedge funds than the difficult and random search for alpha.

This opinion is shared by Jaeger and Wagner (2005). They point out that as it is difficult to decompose hedge fund returns into alpha and beta and as there is no model for describing alpha directly, alpha is the remaining average part of the return when all beta contributions (traditional or alternative) are subtracted. From this point of view, any unaccounted beta will erroneously be attributed to alpha. Under- respectively overestimation of beta will lead to an under- respectively overestimation of alpha. The authors conclude that given the described decomposition difficulties it is hard to verify whether hedge funds really deliver absolute returns. Moreover they claim that estimations of beta are more accurate than those of alpha. In their view, investors should start to recognize the diversifying opportunities offered by alternative betas rather than just focus on absolute returns.

Chapter 3

Gaussian Approximate Bayesian Estimation – Kalman Filter Framework

Summary

3.1. Introduction

Probabilistic inference is the problem of estimating the hidden states of a system in an optimal and consistent fashion given a set of noisy or incomplete observations. The optimal solution to this problem is given by the recursive Bayesian estimation algorithm which recursively updates the posterior density of the system state as new observations arrive online. This posterior density constitutes the complete solution to the probabilistic inference problem, and allows us to calculate any "optimal" estimate of the state. Unfortunately, for most real-world problems, the optimal Bayesian recursion is intractable and approximate solutions must be used.

The Kalman filter was first presented in 1960 in a paper by R.E. Kalman. The filter provides a recursive solution to a discrete-data *linear* filtering problem. It estimates a hidden system state in a way that minimizes the mean squared error. Since its introduction and with the increase in computing power, the Kalman filter has become an important tool. Various books and publications cover the basics and applications of the filter. One of the most comprehensive of them was written by Harvey (1989). It covers the mathematical background of the filter as well as applications on financial time series. Welch and Bishop (2001) give a well understandable overview of the Kalman and Extended Kalman Filter.

The filter has been extended in order to cover also nonlinear filtering problems. Within the space of approximate solutions, the Extended Kalman filter (EKF) has become one of the most widely used algorithms with applications in state, parameter and dual estimation; this method uses the Gaussian approximations. Unfortunately, the EKF is based on a sub-optimal implementation of the recursive Bayesian estimation framework applied to Gaussian random variables. This can seriously affect the accuracy or even lead to divergence of any inference system that is based on the EKF or that uses the EKF as a component part. Rudolph van der Merwe & Eric Wan (2004) have generalized these algorithms, all based on derivative less *statistical linearization,* to a family of filters called *Sigma-Point Kalman Filters* (SPKF) – Unscented Kalman Filter (UKF), Central Difference Kalman Filter (CDKF), Square-Root Unscented Kalman Filter (SR-UKF), Square-Root Central Difference Kalman Filter (SR-UKF) - which will be presented further on in this chapter, and successfully expanded their use within the general field of probabilistic inference, both as stand-alone filters and as subcomponents of more powerful sequential Monte Carlo filters (particle filters). The particle filters will be presented in chapter 4.

3.2. Kalman Filter (KF)

The Kalman Filter is the oldest standard recursive solution for linear filtering problems. It does not require all the past data to be kept in memory and processed for each new state. It processes previous observations/ measurements to obtain the current state; each updated estimate of state is computed from previous estimate and new input data.

Kalman's original derivation of the Kalman filter did not require the underlying system equations to be linear or the probability densities to be Gaussian. The only assumptions made are:

- Consistent minimum variance estimates of the system random variables can be maintained by propagating only their *first and second order moments* (means and covariances) - the densities are not required to be Gaussian; only the Gaussian components (mean and covariance) of these densities in the estimator are maintained
- The estimator (measurement update) itself to be linear
- Accurate predictions of the state (using process model) and of the system observations (using observation model) can be calculated.

These expectations can in general only be calculated exactly for *linear Gaussian* random variables. This does not however disallow the application of the Kalman framework to nonlinear systems. It just requires further *approximations* to be made. One such approximation is the linearization of the dynamic state-space models through the use of a first order truncated Taylor series expansion around the current estimate of the system state. This algorithm is known as the *extended Kalman filter* (EKF), which will be presented next in this chapter.

The Kalman filter estimates a process by using a form of feedback control: the filter estimates the process state at some time and then obtains feedback in the form of (noisy) measurements. The algorithm resembles that of a *predictor-corrector* algorithm for solving numerical problems. As such, the equations for the Kalman filter fall into two groups:

- *time update* equations
    - responsible for projecting forward (in time) the current state and error covariance estimates to obtain the *a priori* estimates for the next time step
    - *predictor* equations

- *measurement update* equations
  - responsible for the feedback (for incorporating a new measurement into the *a priori* estimate to obtain an improved *a posteriori* estimate)
  - *corrector* equations

Kalman derived [***102] the following recursive form of the optimal Gaussian approximate linear Bayesian update (Kalman update for short) of the conditional mean of the state random variable, $\hat{x}_k = E\,[\mathbf{x}_k|\mathbf{y}_{1:k}]$ and its covariance, $\mathbf{P}_{\mathbf{x}k}$:

$$\hat{x}_k \;=\; (\text{prediction of } \mathbf{x}_k) \;+ \mathbf{K}_k\; (\mathbf{y}_k - (\text{prediction of } \mathbf{y}_k)) \qquad (3.1)$$

$$=\; \hat{x}_k^{-} + \mathbf{K}_k\,(\mathbf{y}_k - \hat{\mathbf{y}}_k^{-}) \qquad (3.2)$$

$$\mathbf{P}_{\mathbf{x}k} \;=\; \mathbf{P}_{\mathbf{x}k}^{-} - \mathbf{K}_k\,\mathbf{P}_{\tilde{\mathbf{y}}k}\mathbf{K}_k^{T} \qquad (3.3)$$

While this is a *linear* recursion, we have not assumed linearity of the model. The optimal terms in this recursion are given by:

$$\hat{x}_k^{-} \;=\; E\,[\mathbf{f}\,(\mathbf{x}_{k-1},\, \mathbf{u}_{k-1},\, \mathbf{w}_k)] \qquad (3.4)$$

This is the optimal prediction (prior mean at time k) of $\mathbf{x}_k$ that corresponds to the *expectation* (taken over the posterior distribution of the state at time $k - 1$) *of a nonlinear function of the random variables* $\mathbf{x}_{k-1}$ and $\mathbf{u}_{k-1}$. The random variables w_k and v_k represent the process and measurement noise in equation (3.4) and equation (3.5). The *non-linear* function f relates the state at the previous time step $k - 1$ to the state at the current time step. It includes as parameters any driving function u_k and the zero-mean process noise w_k. The *non-linear* function h in the measurement equation (3.5) relates the state x_k to the measurement y_k.

$$\hat{\mathbf{y}}_k^{-} \;=\; E\,[\mathbf{h}\,(\mathbf{x}_k^{-}\,,\, \mathbf{v}_k)] \qquad (3.5)$$

Similar interpretation for the optimal prediction $\hat{\mathbf{y}}_k^{-}$, except the expectation is taken over the prior distribution of the state at time k).

$$\mathbf{K}_k \;=\; E\,[(\mathbf{x}_k - \hat{x}_k^{-})\,(\mathbf{y}_k - \hat{\mathbf{y}}_k^{-})^{T}]\; E\,[(\mathbf{y}_k - \hat{\mathbf{y}}_k^{-})\,(\mathbf{y}_k - \hat{\mathbf{y}}_k^{-})^{T}] - 1 \qquad (3.6)$$

$$=\; \mathbf{P}_{\mathbf{x}k\,\tilde{\mathbf{y}}k}\mathbf{P}_{\tilde{\mathbf{y}}k}^{-1} \qquad (3.7)$$

The optimal gain term $\mathbf{K}_k$ is expressed as a function of the expected cross correlation matrix (covariance matrix) of the state prediction error and the observation prediction error, and the expected auto-correlation matrix of the observation prediction error. The error between the true observation and the predicted observation, $\tilde{\mathbf{y}}_k = \mathbf{y}_k - \hat{\mathbf{y}}_k^-$ is called the *innovation*. The evaluation of the covariance terms also requires taking *expectations of a nonlinear function of the prior state variable*.

These expectations can in general only be calculated exactly (in an analytical sense) for a *linear* dynamic state-space model and *Gaussian* random variables. Under these (linear, Gaussian) conditions, the Kalman filter framework is in fact an exact solution of the optimal Bayesian recursion. This is probably why the Kalman filter framework is often misconstrued as only applicable to such linear, Gaussian systems. This does not, however, disallow the application of the Kalman framework to nonlinear systems. As mentioned above, even for nonlinear, non-Gaussian systems, the Kalman filter framework is still the (minimum variance) optimal Gaussian approximate linear estimator, if the rest of the assumptions hold. This does however require further *approximations* to be made in order to practically apply this framework to nonlinear systems. Specifically, these approximations directly address how the optimal terms in the Kalman update (Equations 3.4 through 3.7) are calculated.

If we assume the following linear dynamic state space model, then we have the following equations:

Process equation:

$$x_{k+1} = F_{k+1;k} x_k + w_k \qquad (3.8)$$

Measurement equation:

$$y_k = H_k x_k + v_k \qquad (3.9)$$

- w_k and v_k are independent, zero mean white, Gaussian noise processes with covariance matrix Q_k and R_k
- The matrix F in the difference equation (3.8) relates the state at the previous time step k-1 to the state at the current step k, in the absence of either a driving function or process noise
- The matrix H in the measurement equation (3.9) relates the state to the measurement y_k.

Initial values for $k = 0$
Initial estimate of state:

$$\hat{x}_0 = E\,[x_0] \qquad (3.10)$$

Initial estimate of a posteriori covariance:

$$P_0 = E\,[(x_0 - E[x_0])\,(x_0 - E[x_0])^T] \qquad (3.11)$$

Time update equations

The time update equations are responsible for projecting forward the current state and error covariance estimates to obtain the a priori estimates for the next time step.

State estimate propagation – project the state ahead

$$\hat{x}_k^- = F_{k;\,k-1}\hat{x}_{k-1} \quad (3.12)$$

Error covariance propagation – project the error covariance ahead

$$P_k^- = F_{k;\,k-1}P_{k-1}F^T_{k;\,k-1} + Q_{k-1} \quad (3.13)$$

Measurement update equations

The measurement update equations are responsible for incorporating a new measurement into the a priori estimate to obtain an improved a posteriori estimate. The first step during the measurement update is to calculate the Kalman gain, which is chosen such that it minimizes the a posteriori error covariance. After each time and measurement update pair, the process is repeated with the previous *a posteriori* estimates used to project or predict the new *a priori* estimates.

Kalman gain matrix

$$K_k = P_k^- H_k^T\, [\, H_k P_k H_k^T + R_k]^{-1} \quad (3.14)$$

State estimation update with measurement y_k

$$\hat{x}_k = \hat{x}_k^- + K_k\,(y_k - H_k\,\hat{x}_k^-) \quad (3.15)$$

Error covariance update

$$P_k = (I - K_k H_k)\,P_k^- \quad (3.16)$$

3.3. Extended Kalman Filter (EKF)

When the strict assumptions of the Kalman filter do not hold, approximate filters must be used. The most simple and widespread is the Extended Kalman filter (EKF). As the standard Kalman filter, EKF assumes that the posterior density p $(x_k|y_{1:k})$ is normally distributed, i.e. it is approximated by a Gaussian. The EKF linearizes the state space model at each time instant around the most recent time estimate by computing the Jacobian matrices. These are defined as the derivatives of the nonlinear functions with respect to the corresponding variables in the system and observation equations. After the linearization, the equations for the standard Kalman filter can be used.

Consider a nonlinear dynamical system described by the following state-space model:

$$x_{k+1} = f(k;\, x_k) + w_k \quad (3.17)$$

$$y_k = h(k;\, x_k) + v_k \quad (3.18)$$

where w_k and v_k are independent zero-mean white Gaussian noise processes with covariance matrices R_k and Q_k.

The approximation follows in the following two steps:

Stage 1:

$$F_{k+1,k} = \frac{\partial f(k,x)}{\partial x}_{x=xk} \tag{3.19}$$

$$H_k = \frac{\partial h(k,x_k)}{\partial x}_{x=xk^-} \tag{3.20}$$

The ij-th entry of $F_{k+1;k}$ is equal to the partial derivative of the i-th component of $F(k; x)$ with respect to the j-th component of x. The ij-th component of H_k is equal to the partial derivative of the i-th component of $H(k; x)$ with respect to the j-th component of x.

Stage 2:

Once the matrices $F_{k+1;k}$ and H_k are evaluated, they are employed in a first order Taylor approximation of the nonlinear functions F($k; x_k$) and H($k; x_k$) around x_k and x_k^- .

$$F(k; x_k) \approx F(x; x_k) + F_{k+1;k}(x; x_k) \tag{3.21}$$

$$H(k; x_k) \approx H(x; x_k) + H_{k+1;k}(x; x_k^-) \tag{3.22}$$

Hence the non linear state equations are given as:

$$x_{k+1} \approx F_{k+1;k}x_k + w_k + d_k \tag{3.23}$$

$$\overline{y_k} \approx H_k x_k + v_k \tag{3.24}$$

where

$$\overline{y_k} = y_k - h(x; xk-) - H_k\, xk- \tag{3.25}$$

$$d_k = f(x; x_k) - F_{k+1;k}x_k \tag{3.26}$$

The extended Kalman filter equations are given below:

Initial values for $k = 0$

Initial estimate of state:

$$\hat{x}_0 = E\,[x_0] \tag{3.27}$$

Initial estimate of a posteriori covariance:

$$P_0 = E\,[(x_0 - E[x_0])\,(x_0 - E[x_0])^T] \tag{3.28}$$

1. Prediction step

- Compute the process model Jacobians

- Time update equations - predicted state mean and covariance

State estimate propagation

$$\hat{x}_k^- = f(k; \hat{x}_{k-1}) \qquad (3.29)$$

Error covariance propagation

$$P_k^- = F_{k;\,k-1} P_{k-1} F^T_{k;\,k-1} + Q_{k-1} \qquad (3.30)$$

2. Correction step

- Compute the process model Jacobians
- Measurement update equations – update estimates with latest observation

Kalman gain matrix

$$K_k = P_k^- H_k^T \left[H_k P_k H_k^T + R_k \right]^{-1} \qquad (3.31)$$

State estimation update with measurement y_k

$$\hat{x}_k = \hat{x}_k^- + K_k\, y_k - h(k, \hat{x}_k^-) \qquad (3.32)$$

Error covariance update

$$P_k = (I - K_k H_k) P_k^- \qquad (3.33)$$

The Extended Kalman Filter rests one of the most widely used estimation algorithm for non linear systems. This filter approximates the non linear model as time varying linear model, where the state distribution is propagated through the first - order linearization of the non linear system. The linearization method employed by the EKF does not take into account the fact that **x** is a random variable with inherent uncertainty. The linearization around a *single point* (the current state estimate) has large implications for the accuracy and consistency of the resulting EKF algorithm. These approximations often introduce large errors in the EKF calculated posterior mean and covariance of the transformed (Gaussian) random variable, which may lead to suboptimal performance and sometimes divergence of the filter.

3.4. Unscented Kalman Filter (UKF)

The Unscented Kalman Filter is proposed as an alternative to Extended Kalman Filter. Because the EKF only uses the first order terms of the Taylor series expansion of the nonlinear functions, it often introduces large errors in the estimated statistics of the posterior distributions of the states. This is especially evident when the models are highly nonlinear and the local linearity assumption breaks down, i.e., the effects of the higher order terms of the Taylor series expansion becomes significant. The UKF is provably superior to the EKF. It does not need to explicitly calculate the Jacobians or Hessians. In the UKF the state distribution is still represented by a Gaussian random variable, but it is specified using a minimal set of deterministically chosen sample points. These sample points completely capture the true mean and covariance of the Gaussian random variable, and when propagated through the true nonlinear system, captures the posterior mean and covariance accurately to the 2nd order for any nonlinearity, with errors only introduced in the 3rd and higher orders. This small number of carefully chosen sample points –called *sigma-points* – when propagated in each estimation step, provides a compact parameterization of the underlying distribution ([1] see Fig.2). An explicit description of UKF can be found in van der Merwe, Doucet, de Freitas, and Wan [42, 44].

Unscented Transformation

The Unscented Transformation is a method for calculating the statistics of a random variable which undergoes a nonlinear transformation when propagating a random variable x through a nonlinear function, $y = f(x)$. It builds on the principle that it is easier to approximate a probability distribution than an arbitrary nonlinear function. X has mean x and covariance P_x. To calculate the statistics of y, we form a matrix x of $2L + 1$ sigma vectors X_i according to the following:

$$X_o = \overline{x} \tag{3.34}$$

$$X_i = \overline{x} + (\sqrt{(L + \lambda)Px})_i \tag{3.35}$$

$$X_i = \overline{x} - (\sqrt{(L + \lambda)Px})_{i-L} \tag{3.36}$$

where $\lambda = \alpha^2(L + \kappa) - L$ is a scaling parameter. The constant α determines the spread of the sigma points around x, and is usually a small positive value. The constant κ is a secondary scaling parameter, equal to either 0 or 3 - L. β is an extra degree of freedom scalar parameter used to incorporate any extra prior knowledge distribution of x (for Gaussian distributions its optimal value is 2).

These sigma points are propagated through the non linear function and the mean and covariance are approximated using a weighted sample mean and covariance of the posterior sigma points.

$$\overline{y} \approx \sum_{0}^{2L} W_i^{(m)} Y_i \tag{3.37}$$

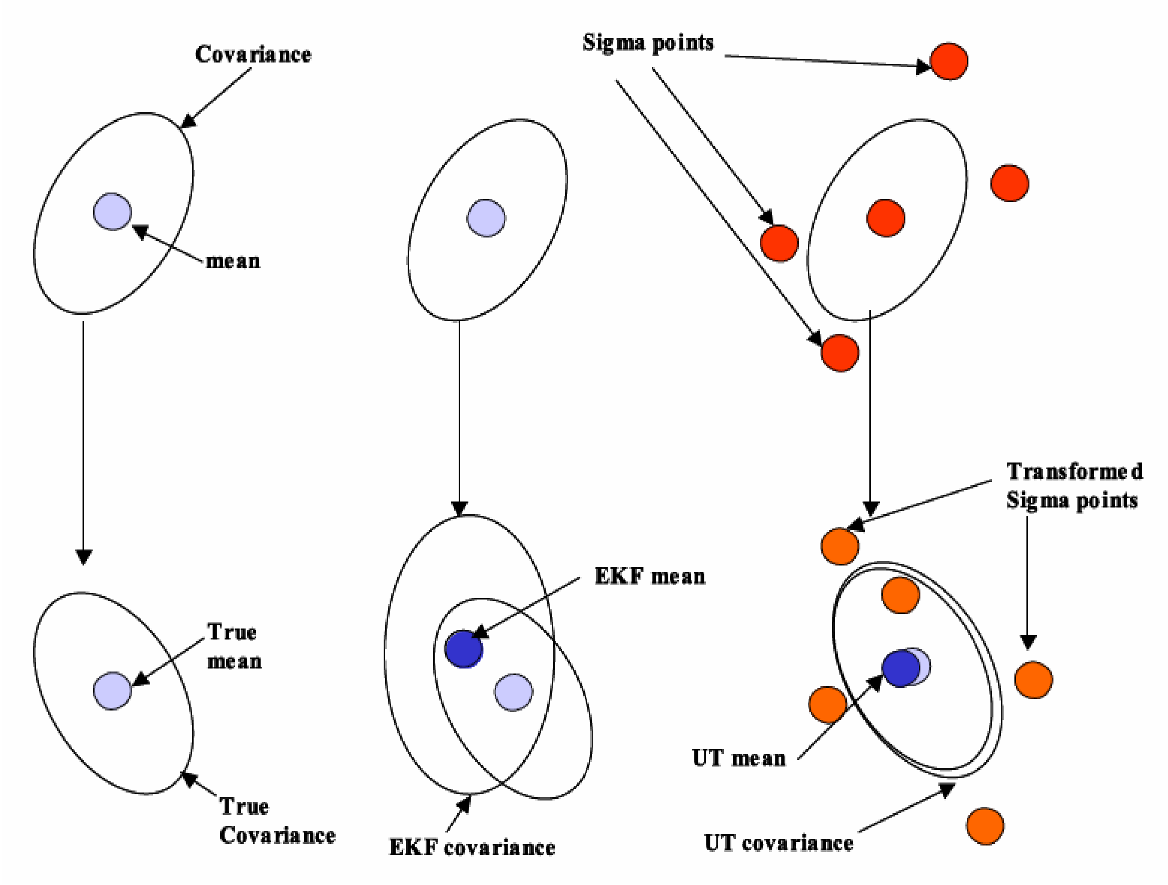
$$P_y \approx \sum_{0}^{2L} W_i^{(c)} (Yi - \overline{y}) (Yi - \overline{y})^T \tag{3.38}$$

The weights are calculated as follows:

$$W_o^{(m)} = \frac{\lambda}{\lambda + L}$$

$$W_o^{(c)} = \frac{\lambda}{\lambda + L} + 1 - \alpha^2 + \beta$$

$$W_i^{(m)} = W_i^{(c)} = \frac{1}{2\,(L + \lambda)} \qquad (3.39)$$



(a) (b) (c)

Fig.2: Graphical Depiction of the Superiority of the Unscented Transformation for mean and covariance propagation – Rudolph van der Merwe [10]
(a) Actual Propagation (b) First Order EKF Linearization (c) Unscented Transformation

The **Unscented Kalman Filter** is an extension of the unscented transformation. The random variable is redefined as the concatenation of the original state and noise variables:

$$\mathrm{x}^{a}_{k} = \begin{bmatrix} \mathbf{x}^{x}_{k} \\ \mathbf{x}^{v}_{k} \\ \mathbf{x}^{n}_{k} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{k} \\ \mathbf{v}_{k} \\ \mathbf{n}_{k} \end{bmatrix} \qquad (3.40)$$

The effective dimension of this augmented state random variable is now $L = L_x + L_v + L_{n,}$ where Lx is the original state dimension, L_v is the process noise dimension and L_n is the observation noise dimension.

The UKF equations are given below:

Initialization :

$$\hat{\mathrm{x}}_{0} = E\,[x_0] \qquad (3.41)$$

$$P_0 = E\,[(x_0 - E[x_0])\,(x_0 - E[x_0])^{\,T}] \qquad (3.42)$$

$$\hat{\mathrm{x}}_{0} = E[x^{a}] = [\hat{\mathrm{x}}^{T}\ 0\ 0]^{T} \qquad (3.43)$$

$$P_0 = E\,[(x_0^{\mathrm{a}} - \hat{\mathrm{x}}_0^{a})\,(x_0^{\mathrm{a}} - \hat{\mathrm{x}}_0^{a})^{\,T}] = \begin{bmatrix} \mathrm{P}^{0} & 0 & 0 \\ 0 & \mathrm{R}^{\mathrm{v}} & 0 \\ 0 & 0 & \mathrm{R}^{\mathrm{n}} \end{bmatrix} \qquad (3.44)$$

R^{v} and R^{n} are the process and observation noise covariances. By augmenting the state random variable with the noise random variables, we take the uncertainty in the noise random variables into account in the same manner as we do for the state during the sigma-point propagation. This allows for the effect of the noise on the system dynamics and observations to be captured with the same level of accuracy as with which we treat the state. In contrast, the EKF simply models the noise RV's using their expected values, which for zero-mean Gaussian noise is equal to **0**.

1. Calculate sigma points :

$$X^{a}_{k\text{-}1} = [\,\hat{x}_{k\text{-}1}^{\;a} \quad \hat{x}_{k\text{-}1}^{\;a} + \gamma\sqrt{P^{a}_{k\text{-}1}} \quad \hat{x}_{k\text{-}1}^{\;a} + \gamma\sqrt{P^{a}_{k\text{-}1}}\,] \qquad (3.45)$$

2. Time-update equations:

$$X^{x}_{k|k\text{-}1} = F\,(X^{x}_{k|k\text{-}1},\, u_{k-1,} X^{v}_{k\text{-}1}) \qquad (3.46)$$

$$\hat{x}_{k}^{\,-} = \sum_{0}^{2\mathrm{L}} \mathrm{W}_{i}^{(m)} X^{x}_{i,k|k\text{-}1} \qquad (3.47)$$

$$P_k^- = \sum_0^{2L} W_i^{(c)} \left(X_{i,k|k-1}^x - \hat{x}_k^- \right) \left(X_{i,k|k-1}^x - \hat{x}_k^- \right)^T \qquad (3.48)$$

3. Measurement update equations

$$Y_{k|k-1} = H(X_{i,k|k-1}^x, X_{k-1}^n) \qquad (3.49)$$

$$\hat{y}_k^- = \sum_0^{2L} W_i^{(m)} Y_{i,k|k-1} \qquad (3.50)$$

$$P_{\tilde{y}k\,\tilde{y}k} = \sum_0^{2L} W_i^{(c)} \left(Y_{i,k|k-1} - \hat{y}_k^- \right) \left(Y_{i,k|k-1} - \hat{y}_k^- \right)^T \qquad (3.51)$$

$$P_{xk\,yk} = \sum_0^{2L} W_i^{(c)} \left(X_{i,k|k-1} - \hat{x}_k^- \right) \left(Y_{i,k|k-1} - \hat{y}_k^- \right)^T \qquad (3.52)$$

$$K_k = P_{xkyk} P_{\tilde{y}k\,\tilde{y}k}^{-1} \qquad (3.53)$$

$$\hat{x}_k = \hat{x}_k^- + K_k (\hat{y}_k - \hat{y}_k^-) \qquad (3.54)$$

$$P_k = P_k^- K_k P_{\tilde{y}k\,\tilde{y}k} K_k^T \qquad (3.55)$$

Parameters: $\mathbf{x}_a = [\, \mathbf{x}^T \; \mathbf{v}^T \; \mathbf{n}^T \,]^T$,

$\boldsymbol{X}_a = [\, (\boldsymbol{X}_x)^T (\boldsymbol{X}_v)^T (\boldsymbol{X}_n)^T \,]^T$,

$\gamma = \sqrt{L + \lambda}$

where γ is a composite scaling parameter and $\lambda = \alpha^2(L + \kappa) - L$; L is the dimension of the augmented states, $\mathbf{R_v}$ is the process-noise covariance, $\mathbf{R_n}$ is the observation-noise covariance, and W_i are the weights as calculated in Eq. 3.39.

3.5. Central Difference Kalman Filter (CDKF)

Another way to approximate a nonlinear function over a certain interval, excepting the Taylor's series expansion, is to make use of an interpolation formula that uses a finite number of functional evaluations instead of analytical derivatives. One particular type of interpolation formula that uses central divided differences is ***Sterling's polynomial interpolation formula***, which for the scalar 2nd order case is given by:

$$g(x) = g(\overline{x}) + \tilde{D}_{\Delta x} g + \frac{1}{2!} \tilde{D}^2_{\Delta x} g \qquad (3.56)$$

where $\tilde{D}_{\Delta x}g$ and $\tilde{D}^{2}_{\Delta x}g$ are the first and second order central divided difference operators acting on $g(x)$. For the scalar case, these are given by:

$$\tilde{D}_{\Delta x}g = (x - \overline{x})\,\frac{g(\overline{x}+h) - g(\overline{x}-h)}{2h} \qquad (3.57)$$

$$\tilde{D}^{2}_{\Delta x}g = (x - \overline{x})^2\,\frac{g(\overline{x}+h) + g(\overline{x}-h) - 2g(\overline{x})}{h^2} \qquad (3.58)$$

where h is the interval length or central difference step size and $\overline{x}$ is the prior mean of x around which the expansion is done. So we can interpret the Sterling interpolation formula as a Taylor series where the analytical derivatives are replaced by central divided differences. For Gaussian random variables the optimal value is thus $h = \sqrt{3}$.

The *weighted sigma-point set* used by Sterling's interpolation formula ($2L$+1 points, where L is the dimension of x) given by the prior mean plus/minus the columns (or rows) of the scaled matrix square-root of the prior covariance matrix is:

$$X_o = \overline{x} \qquad (3.59)$$

$$X_i = \overline{x} + (h\sqrt{Px})_i \qquad i=1,\ldots,L \qquad (3.60)$$

$$X_i = \overline{x} - (h\sqrt{Px})_i \qquad i=L+1,\ldots,2L \qquad (3.61)$$

The weights are:

$$W_o^{(m)} = \frac{h^2 - L}{h^2} \qquad (3.62)$$

$$W_{iI}^{(m)} = \frac{1}{2h^2} \qquad i=1,\ldots,L \qquad (3.63)$$

$$W_i^{(c1)} = \frac{1}{4h^2}) \qquad i=1,\ldots,L \qquad (3.63)$$

$$W_i^{(c2)} = \frac{h^2 - 1}{4h^4}) \qquad i=1,\ldots,L \qquad (3.64)$$

The central difference Kalman filter (CDKF) is a straightforward application of Sterling interpolation for posterior statistics approximation, to the recursive Kalman filter framework. The complete CDKF algorithm that updates the mean $\hat{x}_k$ and covariance $P_{\mathbf{x}k}$ of the Gaussian approximation to the posterior distribution of the states is described below:

- *Initialization:*

$$\hat{x}_0 = E\,[x_0] \qquad P_{x0} = E\,[(x_0 - E[x_0])\,(x_0 - E[x_0])^{T}] \qquad (3.65)$$

$$\overline{v} = E[v] \qquad R_v = E[(v - E[v])\,(v - E[v])^{T}\,] \qquad (3.66)$$

$$\overline{n} = E[n] \qquad R_n = E[(n - E[n])\,(n - E[n])^{T}\,] \qquad (3.67)$$

- *For k* = 1, . . . ,∞ :

1. Calculate sigma points for time-update:

$$\hat{x}_{k-1}^{av} = [\ \hat{x}_{k-1}^{av}\ \overline{v}\ ] \tag{3.68}$$

$$P^{a_v}_{k-1} = \begin{bmatrix} \mathrm{P}_{\mathrm{x\ k-1}} & \mathbf{0} \\ \mathbf{0} & \mathrm{R_v} \end{bmatrix} \tag{3.69}$$

$$X^{a_v}_{k-1} = [\ \hat{x}_{k-1}^{av} \quad \hat{x}_{k-1}^{av} + h\sqrt{P^{a_v}_{k-1}} \quad \hat{x}_{k-1}^{av} - h\sqrt{P^{a_v}_{k-1}}\ ] \tag{3.70}$$

2. Time-update equations:

$$X^{x}_{k|k-1} = f\,(X^{x}_{k-1}, X^{v}_{k-1}, u_{k-1}) \tag{3.71}$$

$$\hat{x}_k^{-} = \sum_{0}^{2L} \mathrm{W}_i^{(m)} X^{x}_{i,k|k-1} \tag{3.72}$$

$$P^{-}_{x_k} = \sum_{0}^{2L} [W_i^{(c1)}\ \ (X^{x}_{i,k|k-1} - X^{x}_{L+i,k|k-1})^2 + \\ W_{i\,i}^{(c2)}\ (X^{x}_{i,k|k-1}\ + X^{x}_{L+i,k|k-1} - 2\,X^{x}_{0,k|k-1})^2] \tag{3.73}$$

3. Calculate sigma points for measurement-update:

$$\hat{x}_{k-1}^{an} = [\ \hat{x}^{-}_{k-1}\ \ \overline{n}\ ] \tag{3.74}$$

$$P^{a_n}_{k|k-1} = \begin{bmatrix} \mathrm{P}^{-}_{\mathrm{x_k}} & \mathbf{0} \\ \mathbf{0} & \mathrm{R_n} \end{bmatrix} \tag{3.75}$$

$$X^{a_n}_{k|k-1} = [\ \hat{x}_{k|k-1}^{an} \quad \hat{x}_{k-1}^{an} + h\sqrt{P^{a_n}_{k|k-1}} \quad \hat{x}_{k-1}^{an} - h\sqrt{P^{a_n}_{k|k-1}}\ ] \tag{3.76}$$

4. Measurement-update equations:

$$Y_{k|k-1} = h\,(X^{x}_{k|k-1}, X^{n}_{k|k-1}) \tag{3.77}$$

$$\hat{y}_k^{-} = \sum_{0}^{2L} \mathrm{W}_i^{(m)}\ Y_{i,k|k-1} \tag{3.78}$$

$$P_{\tilde{y}k} = \sum_{1}^{2L} [\mathrm{W}_i^{(c1)}\ (Y_{i,k|k-1} - Y_{L+i,k|k-1})^2 +$$

$$W_i^{(c2)} (Y_{i,k|k-1} + Y_{L+i,k|k-1} - 2\, Y_{0,k|k-1})^2] \qquad (3.79)$$

$$P_{xk\,yk} = \sqrt{W_1^{(c1)} P_{xk}^-} \quad [Y_{1:L,k|k-1} - Y_{L+1:2L,k|k-1})^T \qquad (3.80)$$

$$K_k = P_{xkyk} P_{\tilde{y}k}^{-1} \qquad (3.81)$$

$$\hat{x}_k = \hat{x}_k^- + K_k (\hat{y}_k - \hat{y}_k^-) \qquad (3.82)$$

$$P_{xk} = P_{xk}^- - K_k P_{\tilde{y}k} K_k^T \qquad (3.83)$$

Parameters: $\mathbf{x}^{av} = [\, \mathbf{x}^T \ \mathbf{v}^T]^T$,

$\boldsymbol{X}^{av} = [\, (\boldsymbol{X}^x)^T (\boldsymbol{X}^v)^T]^T$,

$\mathbf{x}^{an} = [\, \mathbf{x}^T \mathbf{n}^T]^T$,

$\boldsymbol{X}^{an} = [\, (\boldsymbol{X}^x)^T (\boldsymbol{X}^n)^T]^T$,

$h \geq 1$ is the scalar central difference step size, L is the dimension of the augmented states, $R_\mathbf{v}$ is the process-noise covariance, R_n is the observation-noise covariance, and *Wi* are the weights

Rudolph van der Merwe [2] proved that both approaches (Scalar Unscented Transformation and Sterling approximation) can be summarized by three main steps - denoted as the **sigma-point approach** - for the approximating the statistics of a random variable that undergoes a nonlinear transformation:

1. A set of weighted sigma-points are deterministically calculated using the mean and square-root decomposition of the covariance matrix of the prior random variable. As a minimal requirement the sigma-point set must completely capture the first and second order moments of the prior random variable. Higher order moments can be captured, if so desired, at the cost of using more sigma-points.
2. The sigma-points are then propagated through the true nonlinear function using functional evaluations alone (no analytical derivatives are used) in order to generate a posterior sigma-point set.
3. The posterior statistics are calculated (approximated) using tractable functions of the propagated sigma-points and weights.

Both CDKF and UKF perform equally well with negligible difference in estimation accuracy. Both generate estimates however that are clearly superior to those calculated by an EKF. The performance similarity of the UKF and CDKF is clearly demonstrated on nonlinear time-series estimation problem [2]. However, there is one advantage the CDKF has over the UKF: The CDKF uses only a single scalar scaling parameter, the central difference interval size *h*, as opposed to the three (*α, κ, β*) that the UKF uses. Once again this parameter determines the *spread* of the sigma-points around the prior mean. The optimal setting for *h* is equal to the kurtosis of the prior random variable. For Gaussian random variables the optimal value is thus $h = \sqrt{3}$.

3.6. Square-Root Unscented Kalman Filter (SRUKF) and Square-Root Central Difference Kalman Filter (SRCDKF)

One of the most costly operations in the SPKF is the calculation of the matrix square-root of the state covariance at each time step in order to form the sigma-point set. Due to this and the need for more numerical stability (especially during the state covariance update), R. van der Merwe and A. Wan derived numerically efficient *square-root* forms of both the UKF and the CDKF[5, 6]. These forms propagate and update the square-root of the state covariance directly in Cholesky factored form, using the sigma-point approach the following three powerful linear algebra techniques: *QR decomposition*, *Cholesky factor updating* and *efficient pivot based least squares*. For implementation details, see [4, 5, 6]. The square-root unscented Kalman filter and the square-root central difference Kalman filter give a reduced computational cost for certain dynamic state space models and an increased numerical stability.

Chapter 4

Particle Filters

Summary

4.1. Generic Particle Filter

The *particle filter* is a sequential Monte Carlo (SMC) based method that allows for a complete representation of the state distribution using sequential importance sampling and re-sampling. It is a sophisticated model estimation technique based on simulation. The particle filters are usually used to estimate Bayesian models and are the sequential analogue of Markov chain Monte Carlo (MCMC) batch methods. Whereas the standard EKF and the sigma-point filters, presented in the previous chapter, make a Gaussian assumption to simplify the optimal recursive Bayesian estimation, particle filters make no assumptions on the form of the probability densities in question, that is full nonlinear, non-Gaussian estimation.

The working mechanism of particle filters is following: The state space is partitioned as many parts, in which the particles are filled according to some probability measure. The higher probability, the denser the particles are concentrated. The particle system evolves along the time according to the state equation, with evolving probability density function (pdf) determined by the FPK equation. Since the pdf can be approximated by the point-mass histogram, by random sampling of the state space, we get a number of particles representing the evolving pdf. However, since the posterior density model is unknown or hard to sample, we would rather choose another distribution for the sake of efficient sampling.

4.1.1. Monte Carlo approximation and sequential importance sampling

Particle filtering is based on *Monte Carlo simulation* with sequential importance sampling (SIS). The overall goal is to directly implement optimal Bayesian estimation by recursively approximating the complete posterior state density.

Importance sampling is a Monte-Carlo method that represents a distribution p(x) by an empirical approximation based on a set of weighted samples (particles)

$$\mathrm{p(x)} \approx \sum_{i=1}^{N} w^{(\mathrm{i})}\, \delta(\mathrm{x} - x^{(\mathrm{i})}) \tag{4.1}$$

where δ is the Dirac delta function, and the weighted sample set, $\{w^{(l)}, x^{(i)}; l = 1 \ldots N\}$ are drawn from some related, easy-to-sample-from *proposal* distribution π(x). The weights are given by:

$$w(i) = \frac{p(x^{(i)})/\, \pi\, (x^{(i)})}{\sum_{i=1}^{N} p(x^{(i)})/\, \pi\, (x^{(i)})} \tag{4.2}$$

Given this, any estimate of the system such as $E_p[g(x)] = \int g(x)p(x)dx$ can be approximated by $\hat{E}[g(x)] = \ w(i)g(X(i))$ [7]. Using the first order Markov nature of the dynamic state space model and the conditional independence of the observations given the state, a recursive update formula (implicitly a nonlinear measurement update) for the importance weights can be derived [7]. This is given by:

$$w_k^{(i)} = w_{k-1}^{(i)} p(\mathrm{y_k}|\mathrm{x_k}) p(\mathrm{x_k}|\mathrm{x_{k-1}})/\, \pi\, (\mathrm{x_k}|\mathrm{X_{k-1}}, \mathrm{Y_k}) \quad \text{for } \mathrm{x_k} = x^{(i)} \tag{4.3}$$

Equation (4.3) provides a mechanism to sequentially update the importance weights given an appropriate choice of proposal distribution, $\pi\, (\mathrm{x_k}|\mathrm{X_{k-1}}, \mathrm{Y_k})$. Since it is possible to sample from the proposal distribution and evaluate the likelihood $\mathrm{p(y_k|x_k)}$ and transition probabilities $\mathrm{p(x_k|x_{k-1})}$, all we need to do is generate a prior set of samples and iteratively compute the importance weights. This procedure then allows us to evaluate the expectations of interest by the following estimate:

$$E\, [g(\mathrm{x_k})] \approx \frac{1/N \sum_{i=1}^{N} w_k^{(i)} g(x_k^{(i)})}{1/N \sum_{i=1}^{N} w_k^{(i)}} = \sum_{i=1}^{N} \tilde{w}_k^{(i)}\, g(x_k^{(i)}) \tag{4.4}$$

where the normalized importance weights are given by:

$$\tilde{w}_k^{(i)} = w_k^{(i)} / \sum_{j=1}^{N} \tilde{w}_k^{(j)} \tag{4.5}$$

This estimate asymptotically converges if the expectation and variance of g $(\mathrm{x_k})$ and w_k exist and are bounded, and if the support of the proposal distribution includes the support of the posterior distribution. Thus, as N tends to infinity, the posterior filtering density function can be approximated arbitrarily well by the point-mass estimate:

$$\hat{p}\,(\mathrm{x_k}|\mathrm{Y_k}) = \sum_{i=1}^{N} \tilde{w}_k^{(i)}\ \delta(x_k - x_k^{(i)}) \qquad (4.6)$$

These point-mass estimates can approximate any general distribution arbitrarily well, limited only by the number of particles used and how well the above mentioned importance sampling conditions are met.

4.1.2. Re-sampling and Sample Depletion

The sequential importance sampling (SIS) algorithm presented above has a serious limitation: the variance of the importance weights increases stochastically over time. Typically, after a few iterations, one of the normalized importance weights tends to 1, while the remaining weights tend to zero. A large number of samples are thus effectively removed from the sample set because their importance weights become numerically insignificant. To avoid this degeneracy, a re-sampling or selection stage may be used to eliminate samples with low importance weights and multiply samples with high importance weights. Usually either *sampling-importance re-sampling* (SIR) or *residual re-sampling* is used. More theoretical and implementation detail on re-sampling are given in [7].

After the selection/re-sampling step at time k, we obtain N particles distributed marginally according approximately to the posterior distribution. Since the selection step favours the creation of multiple copies of the "fittest" particles, many particles may end up having no children (Ni = 0), whereas others might end up having a large number of children, the extreme case being Ni = N for a particular value i. In this case, there is a severe *depletion of samples*. Therefore, and additional procedure, such as a single Markov Chain Monte Carlo step, is often required to introduce sample variety after the selection step without affecting the validity of the approximation they infer. More details in 8].

4.1.3. The Particle Filter Algorithm

For the implementation, the choice of the proposal distribution $\pi\ (\mathrm{x_k}|\mathrm{X_{k-1}}, \mathrm{Y_k})$ is the most critical design issue. The optimal proposal distribution (which minimizes the variance on the importance weights) is given by:

$$\pi\ (\mathrm{x_k}|\mathrm{X_{k-1}}, \mathrm{Y_k}) = p\ (\mathrm{x_k}|\mathrm{X_{k-1}}, \mathrm{Y_k}) \qquad (4.7)$$

This represents the true conditional state density given the previous state history and all observations. Sampling from this is impractical for arbitrary densities; consequently the transition prior is the most popular choice of proposal distribution [7]:

$$\pi\ (\mathrm{x_k}|\mathrm{X_{k-1}}, \mathrm{Y_k}) \text{ chosen as} = p\ (\mathrm{x_k}|\mathrm{x_{k-1}}) \qquad (4.8)$$

The algorithm for the generic particle filter is described above:

1. *Initialization: k=0*
   - For i = 1. . . N, draw (sample) particle $\mathrm{x}_0^{(i)}$ from the prior p(x0).

2. For k = 1, 2 . . .

(a) *Importance sampling step*

- For i = 1. . . N, sample $x_k^{(i)} \sim \pi(x_k \mid x_{k-1}^{(i)}, Y_k)$
- For i = 1. . . N, evaluate the importance weights up to a normalizing constant:

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p(y_k \mid x_k^{(i)}) p(x_k^{(i)})}{\pi(x_k^{(i)} \mid x_{k-1}^{(i)}, Y_k) \mid} \tag{4.9}$$

- For i = 1. . . N, normalize the importance weights:

$$\tilde{w}_k^{(i)} = w_k^{(i)} / \sum_{j=1}^{N} \tilde{w}_k^{(j)} \tag{4.10}$$

(b) *Selection step (re-sampling)*

- Multiply/suppress samples $x_k^{(i)}$ with high/low importance weights $\tilde{w}_k^{(i)}$ to obtain N random samples approximately distributed according to $p(x_k|Y_k)$.
- For i = 1. . . N, set $\tilde{w}_k^{(i)} = w_k^{(i)} = \frac{1}{N}$
- (optional) Do a single MCMC (Markov chain Monte Carlo) move step to add further 'variety' to the particle set without changing their distribution.

(c) *Output:* The output of the algorithm is a set of samples that can be used to approximate the posterior distribution as follows:

$$\hat{p}(x_k|Y_k) = \frac{1}{N} \sum_{i=1}^{N} \delta(x_k - x_k^{(i)}) \tag{4.11}$$

From these samples, any estimate of the system state can be calculated:

$$\hat{x}_k = E[x_k|Y_k] = \frac{1}{N} \sum_{i=1}^{N} x_k^{(i)} \tag{4.12}$$

The effectiveness of this approximation depends on how close the proposal distribution is to the true posterior distribution. If there is not sufficient overlap, only a few particles will have significant importance weights when their likelihood is evaluated.

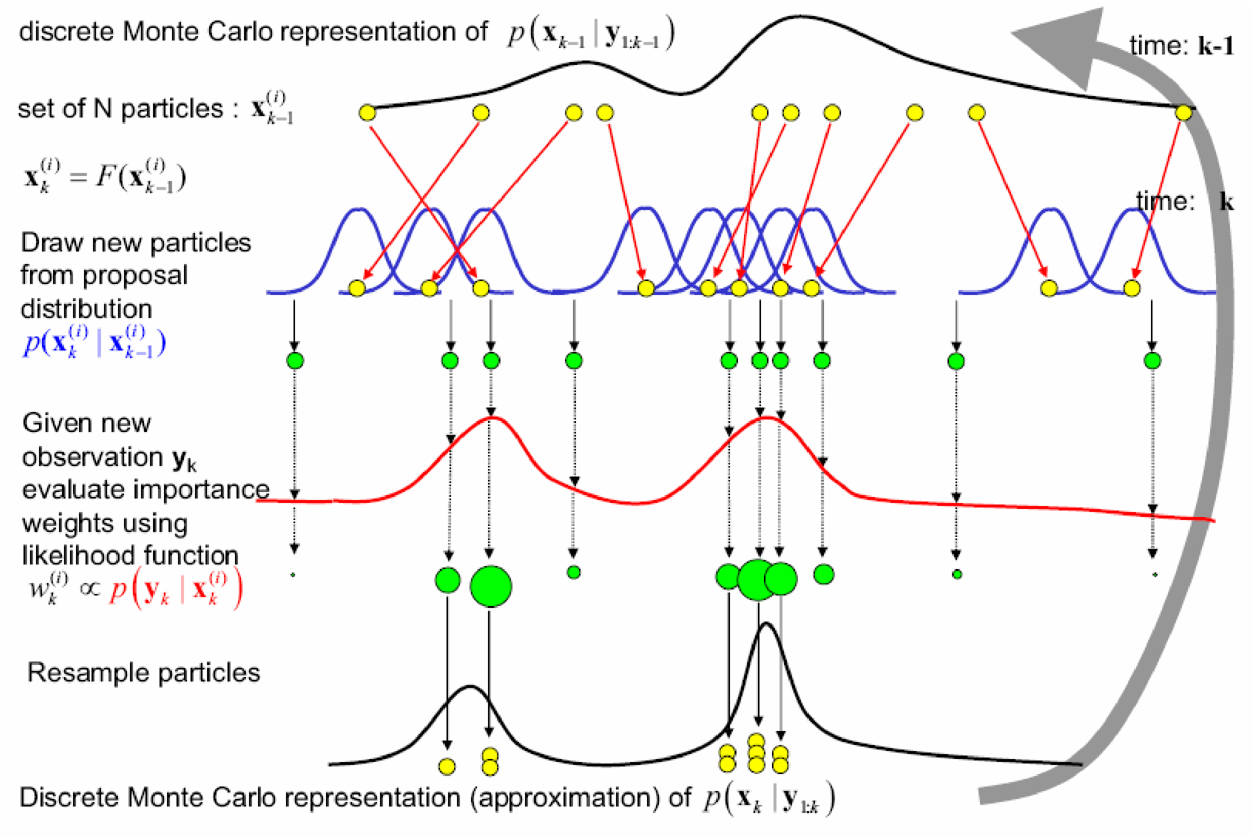

discrete Monte Carlo representation of $p(\mathbf{x}_{k-1} | \mathbf{y}_{1:k-1})$
time: k-1
set of N particles : $\mathbf{x}_{k-1}^{(i)}$
$\mathbf{x}_k^{(i)} = F(\mathbf{x}_{k-1}^{(i)})$
time: k
Draw new particles from proposal distribution
$p(\mathbf{x}_k^{(i)} | \mathbf{x}_{k-1}^{(i)})$
Given new observation $\mathbf{y}_k$ evaluate importance weights using likelihood function
$w_k^{(i)} \propto p(\mathbf{y}_k | \mathbf{x}_k^{(i)})$
Resample particles
Discrete Monte Carlo representation (approximation) of $p(\mathbf{x}_k | \mathbf{y}_{1:k})$

Fig.3: Schematic diagram of a generic particle filter (SIR-PF)

4.2. Sigma-point Particle Filter

An improvement in the choice of proposal distribution over the simple transition prior, which also address the problem of sample depletion, can be accomplished by moving the particles towards the regions of high likelihood, based on the most recent observation y_k.

An effective approach to accomplish this is to use an EKF generated Gaussian approximation to the optimal proposal:

$$\pi\ (x_k|x_{k-1},\ Y_k) \text{ chosen as} = q_{\mathcal{N}}(x_k|Y_k) \qquad (4.13)$$

which is accomplished by using a separate EKF to generate and propagate a Gaussian proposal distribution for each particle:

$$q_{\mathcal{N}}(x_k|Y_k) = \mathcal{N}(\ x_k;\ x_k,\ P_k^{(i)}\ ) \qquad i = 1, 2, \ldots N \qquad (4.14)$$

At time k one uses the EKF equations, with the new data, to compute the mean and covariance of the importance distribution for each particle from the previous time step k − 1. Next, the i-th particle is redrawn (at time k) from this new updated distribution. While still making a Gaussian assumption, the approach provides a better approximation to the optimal conditional proposal distribution and has been shown to improve performance on a number of applications [15].

By replacing the EKF with a sigma-point Kalman filter (UKF, CDKF, SRUKF, SRCDKF), we can more accurately propagate the mean and covariance of the Gaussian approximation to the state distribution. Distributions generated by the SPKF will have a greater support overlap with the true posterior distribution than the overlap achieved by the EKF estimates. In addition, scaling parameters used for sigma point selection can be optimized to capture certain characteristic of the prior distribution. The new filter that results from using a SPKF for proposal distribution generation within a particle filter framework is called the *Sigma-Point Particle Filter* (SPPF). See for implementation details [38, 42].

4.3. Gaussian Mixture Sigma-Point Particle Filter

Particle filters need to use a large number of particles for accurate and robust operation, which often make their use computationally expensive. They suffer from an ailment called "sample depletion" that can cause the sample based posterior approximation to collapse over time to a few samples. This problem can be addressed by moving particles to areas of high likelihood through the use of a SPKF generated proposal distribution. Although the SPPF has large estimation performance benefits over the standard PF, it still remains heavy to compute since it has to run a SPKF for each particle in the posterior state distribution.

The ***Gaussian Mixture Sigma-Point Particle Filter*** (GMSPPF) [19] has equal or better estimation performance when compared to standard particle filters and the SPPF, at a largely reduced computational cost. The GMSPPF combines an *importance sampling* (IS) based measurement update step with a *SPKF based Gaussian sum filter* for the time-update and proposal density generation. The GMSPPF uses a finite Gaussian mixture model (GMM) representation of the posterior filtering density, which is recovered from the weighted posterior particle set of the IS based measurement update stage, by means of a *Expectation-Maximization (EM)* step. The EM step either follows directly after the re-sampling stage of the particle filter, or it can completely replace that stage if a *weighted EM* algorithm is used. The EM or WEM recovered GMM posterior further mitigates the "sample depletion" problem through its inherent "kernel smoothing" nature.

Chapter 5

Wavelet analysis

Summary

5.1. Introduction

Wavelet analysis is a relatively new tool in the field of applied mathematics. Daubechies (1992), Chui (1992) and Graps (1995) provide the fundamentals of the wavelet theory. Wavelet analysis provides the opportunity to make semi-parametric estimations of highly complex structures without knowing the underlying functional form. Wavelet analysis had its impact on the area of signal processing, data compression and image analysis. The impact on signal processing was reviewed by Donoho (1995). Walker (2000) provides a primer on wavelets and their use in these applications of signal processing, image analysis and data compression.

Wavelet analysis, in contrast to Fourier analysis, gives insight in local behaviour, whereas Fourier analysis gives insight in global behaviour. The Fourier transforms processes time-series by transforming the signal from the time domain into the frequency domain. The new processed signal provides insight in the amount of frequencies and the amount of energy in each frequency existing in this time-series. However, local effects are only visible in the time domain and not in the frequency domain. If the signal is stationary, we don't need the "location" information, but in the real world most of our data sets are non-stationary.

The Windowed Fourier Transform (WFT) can locate the window of the data that are transformed in time. The WFT only transforms part of a signal and that segment of signal is small enough that we can assume that portion of signal is stationary. By using a particular window function and shifting the window along the time dimension of the signal we can localize the frequency in the signal, and we obtain a time-frequency representation of the signal. The transformation coefficients are the amplitudes of different frequencies at different times. But WFT has a problem, known as the resolution problem. The Heisenberg uncertainty principle states that we cannot know the exact time-frequency representation of a signal. We can know however the bands of frequencies associated with the time intervals in the signal. Here we have to have a requirement on the width of the window function. The wavelet transformation is a solution to the problem.

Wavelet analysis makes use of a fully scalable window, which is shifted along the signal in order to capture local behaviour in the time domain. This process is repeated several times with different window-sizes, with a collection of time-frequency representations of the signal as a result. The transformation of the signal into the several resulting wavelet coefficients, which provide information at different scales, is more often referred to as time-scale decomposition. However, as there is no direct connection between the Fourier frequency parameter and the Wavelet parameter, the term scale is preserved for wavelet analysis, whereas the term frequency is preserved for Fourier analysis. The use of wavelet analysis enables the analysis of non-stationary data, localization in time and time-scale decomposition, which proved to be useful in the analysis of economic and financial data (Ramsey 1999).

The Continuous Wavelet Transform (CWT) uses a particular wavelet waveform, which has some required or desired properties, as does the window function (which applies the same logic as WFT). There are two main differences between WT and WFT. First, in CWT we use a wavelet to replace the cosine in WFT, which will give us many spikes in the decomposed signal. Second, the most significant characteristic of a particular CWT is the width of the window, which is changed for different frequencies. As we noted for CWT, the windowed signal multiplied with window function is then continuously integrable across time. That is not a discrete transformation and contains highly redundant information.

The Discrete Wavelet Transform (DWT) reduces the signal sample by a factor of two each time according to Nyquist's rule, and then decomposes (resolves) the signal at different frequency bands with different resolution for each frequency band. Each frequency is localized in a particular place in the time domain, according to that band's resolution.

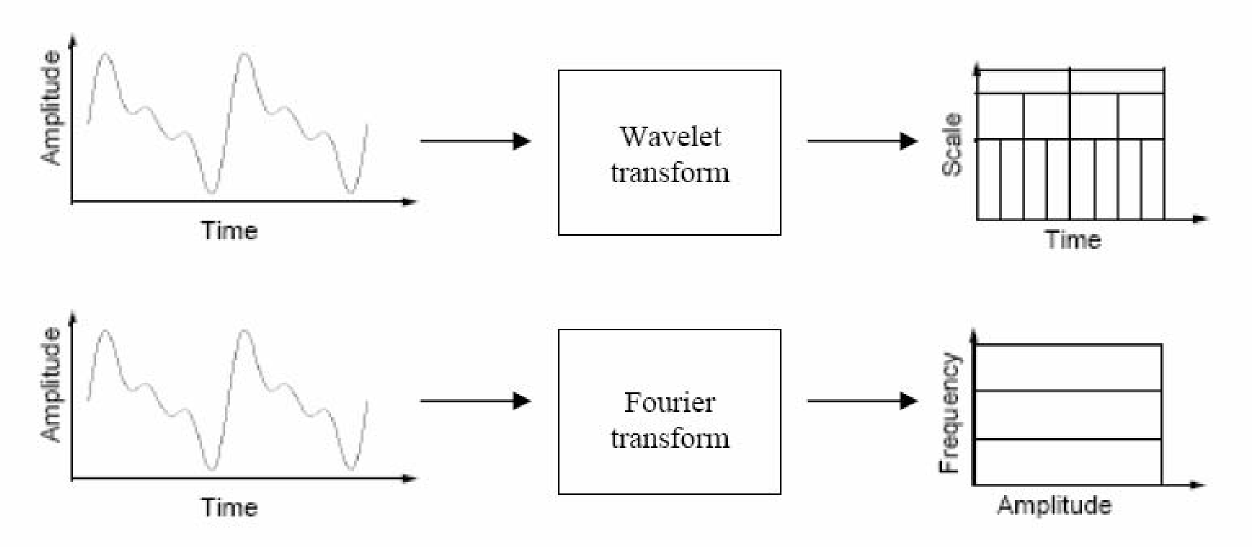


Fig. 4: Wavelet and Fourier transform

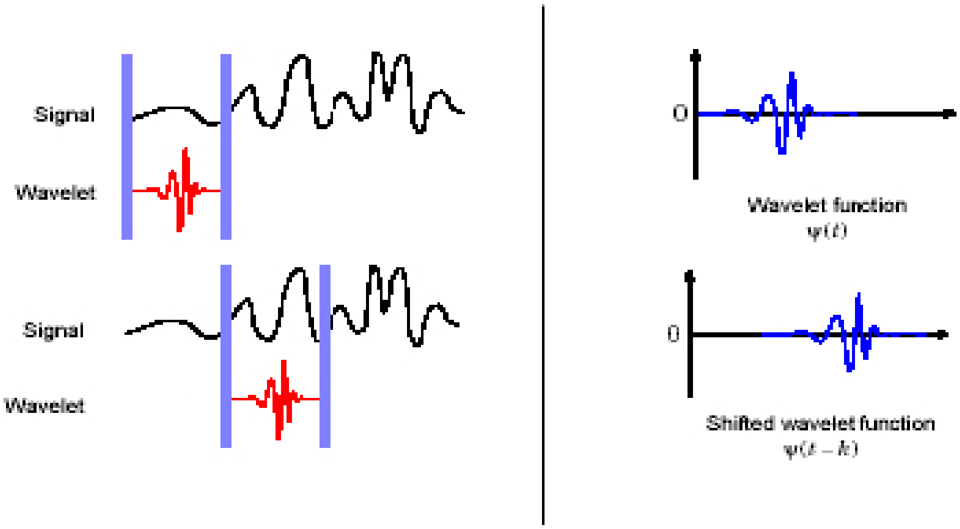


Fig. 5: Left: Shifing a wavelet along the signal. Right: Shifting a wavelet function in time

5.2. Overview of the wavelet transforms

A time signal, *f(t)* with finite energy in the space of all integrable functions, ***L²(R)*** can be approximated and represented using a wavelet transform by projecting the function onto the translated and dilated father and mother wavelets.

$$\phi_{J,k}(t) = 2^{-j/2}\phi(2^{-j}t - k) \qquad (5.1)$$

where $j,\ k \in Z = \{0, \pm1, \pm2, ...\}$

$$\Psi_{j,k}(t) = 2^{-j/2}\,\Psi(2^{-j}t - k) \qquad (5.2)$$

where $j,\ k \in Z = \{0, \pm1, \pm2, ...\}$

Father wavelets, denoted by ϕ, are used to represent the smooth and low frequency portions of the function while mother wavelets, denoted by Ψ, describe the details or high frequency components of the function.

The wavelet representation of the function, *f(t)* is defined as:

$$f(t) = \sum_k s_{J,k}\,\phi_{J,k}(t) + \sum_k d_{J,k}\,\Psi_{J,k}(t) + \sum_k s_{J-1,k}\,\phi_{J-1,k}(t) + \sum_k d_{J-1,k}\,\Psi_{J-1,k}(t) + \ldots + \sum_k d_{1,k}\,\Psi_{1,k}(t)$$

(5.3)

where *J* is the number of multi resolution components and *k* ranges from 1 to the number of coefficients in the specific component. The coefficients are the wavelet transform coefficients given by the projections:

$$s_{J,k} \approx \int \phi_{J,k}(t)\ f(t)\ dt \qquad (5.4)$$

$$d_{J,k} \approx \int \Psi_{J,k}(t)\ f(t)\ dt \qquad (5.5)$$

where $j = 1,2,3\ldots\ldots,J$ (3)

The magnitude of these coefficients actually reflects the significance of the corresponding wavelet function to the total original function.

The equations described above form the basis for the continuous wavelet transform. However, this form of wavelet analysis is known to be highly redundant (since it effectively expands a one dimensional time series into a two dimensional time-scale space, we would expect the resulting data to be rank deficient). A more efficient implementation should be able to represent all the information of a function using a minimum number of wavelet coefficients through critical sampling. This type of wavelet transform is known as the discrete wavelet transform, and is both practical and adequate in most cases especially for input data that is discretely sampled.

Consider $\boldsymbol{x}$ to be a vector of observations of dyadic length (n=2^J). Discrete wavelet transform maps a signal to a J+1 vectors of n wavelet coefficients $\boldsymbol{w} = (w_1, w_2, w_3, w_4, \ldots, w_n)^T$.

$$\boldsymbol{w} = W_{\boldsymbol{x}} \qquad (5.6)$$

The matrix W is composed of the wavelet and scaling filter coefficients. In practice, a pair of high pass and low pass filters are used to filter input signal, $\boldsymbol{x}$. Both the filter outputs are then sub-sampled to half their original lengths. The sub-sampled outputs from the high pass filter are kept as detailed coefficients while the filtering operation is repeated for the sub-sampled output from the low pass filter. The whole process is repeated until the Jth iteration where the Jth order detailed and smooth coefficients are extracted. The resulting vector $\boldsymbol{w}$ contains the detailed coefficients ($d_{J,k}, d_{J-1,k}, \ldots\ d_{1,k}$) and smooth coefficients ($s_{J,k}$). The smooth coefficients describe the underlying smooth behavior of the signal at coarse level 2^J while the detailed coefficients describe the deviations from the smooth behavior at different scales. By ensuring critical sampling and orthogonality, the minimum number of wavelet coefficients is retained to while preserving a perfect representation of the original signal or function. The number of coefficients at different scales is related to the width of the wavelet functions. For example, at scale 2, the translation steps are $2k$, and so $n/2$ terms are required in order for the function $\Psi_{1,k}$ to cover the interval of $1<t<$n. By similar reasoning, a summation involving $\Psi_{J,k}$ and $\phi_{J,k}$ will only require $n/2^j$ terms.

The wavelet transform thus decomposes the input into orthogonal components at different scales and translations. Consider a input signal $f(t)$, the multi resolution decomposition of the signal can be defined as following:

$$f(t) = S_J(t) + D_J(t) + \ldots\ldots. + D_2(t) + D_1(t) \qquad (5.7)$$

$$S_j(t) = \sum_k s_{J,k}\ \phi_{J,k}(t) \qquad (5.8)$$

$$D_j(t) = \sum_k d_{J,k}\ \Psi_{J,k}(t) \qquad (5.9)$$

for $j = 1,2, \ldots J$

$S_j(t)$ are called the smooth signal or approximation signal while *$D_j(t)$* is known as the detailed signals.

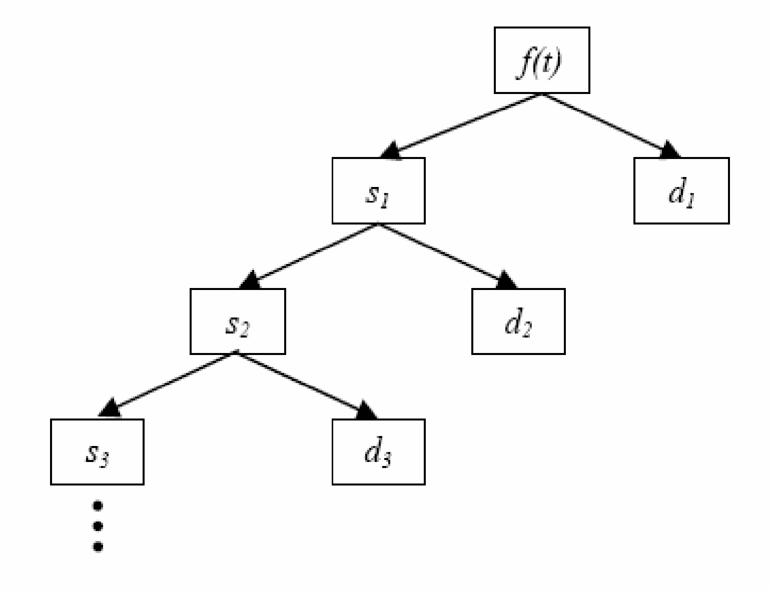


Fig. 6: Multi level decomposition of input

5.3. Wavelets in finance, economics and soft-computing

Wavelets theory has developed into a methodology for financial analysis, and wavelet analysis has been applied to a lot of situations, and most of them have had favourable results. Wavelet analysis has remarkable impact on many fields, mainly on mathematics, signal processing, image analysis, data compression, geophysics, numerical analysis, and statistics.

Wavelets applications in finance can be grouped into three categories. The first category is that wavelets methods are used to study the non-stationary property of financial time series; associated topics include structure change, local stationary and long-memory process. The second category is that wavelet methods provide alternatives for forecasting. Wavelets applications in statistics, discussed above, closely relates to the above two categories, which are technically oriented. The third category is more oriented by the financial theories; they concern how wavelets decompositions can be used to improve the hypothesis testing on exiting theories and also can provide insights of financial phenomena and enhance the development of theories. Next is a brief literature review of wavelets in finance.

Mark Jensen's article (Jensen 1997) is an introduction to application of wavelets to finance. In this article he discussed the advantages of using wavelets to deal with high frequency irregular spaced financial data. In a series of his papers ((Jensen 1999), (Jensen

2000), and (Jensen 2001)) on using wavelets of analysis long-memory process, he developed several methods to estimate the fractional differencing parameter in autoregressive, fractionally integrated, moving average model (ARFIMA). Based on his wavelet OLS estimator, Tkacz (Tkacz 2000) studied the order of integration of interest rates for U. S. and Canada, and find that most rates are mean-reverting in the very long run, with the fractional order of integration increasing with the term to maturity.

Ramsey (1999) gives an overview of the contribution of wavelets to the analysis of economic and financial data. The ability to represent highly complex structures without knowing the underlying functional form proved to be a great benefit for the analysis of these time-series. In addition, wavelets facilitate the precise location of discontinuities and the isolation of shocks. Furthermore, the process of smoothing found in the time-scale decomposition facilitates the reduction of noise in the original signal, by first decomposing the signal into the wavelet components, then eliminating all values with a magnitude below a certain threshold and finally reconstructing the original signal with the inverse wavelet transform (Walker 2000).

Ramsey and Lampart (1998) used wavelet analysis for time-scale decomposition. They researched both the relationships between consumption and income and money and GDP. The time-scale decomposition yielded a new transformed signal built up from the several wavelet coefficients representing the several scales. At each scale, a regression was made between the two variables. This research yielded three conclusions: First, the relationship between economic variables varies across different scales. Second, the decomposition resolved anomalies from the literature. Third, the research made clear that the slope relating consumption and income declines with scale. In this context, the role of real interest was strong in the consumption-income relation. Chew (2001) researched the relationship between money and income, using the same technique of wavelet-based time-scale decomposition as Ramsey and Lampart (1998) did. This research yielded a greater insight in the money versus income nexus in Germany. Arino (1996) used wavelet-based time-scale decomposition for forecasting applications. The approach used was to apply forecasting methods on each of the resulted coefficients from the time-scale decomposition. After applying forecast methods on each of these coefficients, the final forecast of the complete series was obtained by adding up the individual forecasts.

Aussem and Murtagh (1997) used neural networks to examine the individual coefficients. The trained neural network with its approximated variables in the target function was used for the final forecast. In the area of finance, multi-resolution analysis appears useful, as different traders view the market with different time resolutions, for example hourly, daily, weekly or monthly. The shorter the time-period, the higher the frequency. Different types of traders create the multi-scale dynamics of time-series.

Struzik (2001) applied the wavelet-based effective Holder exponent to examine the correlation level of the Standard & Poor's index locally at arbitrary positions and resolutions (time and scale).

Norsworty et al. (2000) applied wavelets to analyze the relationship between the return on an asset and the return on the market portfolio, or investment alternative. Similar to other researches in the field of finance and economics, they applied wavelet-based time-scale decomposition to investigate whether there are changes in behavior for different frequencies. The research indicated that the effect of the market return on an individual asset's return will be greater in the higher frequencies than in the lower.

Wavelets are also applied in the area of soft-computing, especially in the area of neural networks and fuzzy systems. Tan and Yu (1999) researched the complementarily and equivalence relationships between convex fuzzy systems and wavelets. After discussing the fundamentals of both fuzzy- and wavelet systems, the authors conclude that there is a complementarily relationship. As a result, fuzzy systems and wavelets can be combined together, to represent linguistic and numerical knowledge. In the case of function approximation, a fuzzy quantization and fuzzy rules can be constructed to effectively represent the function. In addition, wavelets can be used for further improvement of the approximation accuracy, by capturing the fine features. Furthermore, there is equivalence between multi-scale fuzzy systems and wavelets. This means that any result obtained from a multi-scale fuzzy approximation, can have its direct interpretation in an equivalent wavelet approximation. Fuzzy rules can be generated from wavelet coefficients.

Hoa, Zhang and Xu (2001) proposed a fuzzy wavelet network for function approximation. Such a FWN is built forth on wavelet neural networks. These networks are a combination of feed forward neural networks and wavelets. The main issue here is that wavelets are used as the transformation function in the hidden layer, whereas this was a sigmoid- or hyperbolic tangent function traditionally. The network is then trained by adapting the translation and dilation parameters in the wavelet function, in the variable wavelet variant. The FWN consists of four layers: input, fuzzification, inference and defuzzification layers. This is inspired by the traditional neuro-fuzzy systems, as described in Jang, Sun and Mizutani (1997). However, the difference between the FWN and the neuro-fuzzy systems is that the defuzzification-layer is built-up from a number of sub-wavelet neural networks, instead of using constants or linear equations. The FWN uses both globalized and localized approximation of the function, yielding better local accuracy and faster convergence. The input and fuzzification layer of the FWN construct the antecedents of the fuzzy rules from the input space. The sub-wavelet neural networks create the consequents of these rules, by making linear combinations of a finite set of wavelets, based on the same input space. Later on, in the defuzzification, the antecedents and consequents are combined to form the final output.

Some **advantages** using wavelet methods:

- Robustness of procedure (erroneous assumptions) (no parametric tests of procedures)
- Flexibility of regression fit (imprecise model formulations)
- Ability to handle complex relationships
- Efficiency of the estimators (few data points)
- Simplicity of implementation
- Dealing with non-stationarity of the stochastic innovations that inevitably are involved with economic and financial time series.

Chapter 6

Clustering methods

Summary

6.1. Introduction

The literature on cluster analysis is quite large and diverse. Significant work on cluster analysis has been done in various fields. Cluster analysis has frequently been employed as a classification tool. Classification is concerned with the identification of discrete categories, whereas structural representation is concerned with the development of a faithful representation of relationships. Cluster analysis is a statistical method of classification, yet it is different from classification. Classification in its purest form pertains to a known number of groups, and the operational objective is to assign new observations to one of these groups. In cluster analysis, no assumptions are made concerning the number of groups or the group structure. Grouping is based on similarities or distances (dissimilarities).

Cluster analysis is an exploratory technique in which the information provided by the analyst in the form of relevant attributes is used to come up with a natural grouping of data, if any. It is a tool of discovery that reveals structure and relations in data. The results of a cluster analysis can contribute directly to the development of classification schemes. Strictly speaking, a set of results applied only to the sample on which they are based; but through appropriate modification, technique employed can be extended to describe adequately the properties of other samples and ultimate the parent population.

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.

The main **requirements** that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- interpretability and usability.

There are a number of **problems** with clustering. Among them:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of "distance" (for distance-based clustering);
- if an *obvious* distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces;
- the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

6.2. Cluster analysis

6.2.1. The data

Clustering techniques can be applied to data that is quantitative (numerical), qualitative (categorical), or a mixture of both. The data are typically observations of some physical process. Each observation consists of n measured variables, grouped into an n-dimensional row vector $\mathrm{x}_k = [x_{k1}, x_{k2}, \ldots , x_{kn}]^T$; $\mathrm{x}_k \in \mathrm{R}^n$. A set of N observations is denoted by $\mathrm{X} = \{\mathrm{x}_k | k = 1, 2, \ldots , N\}$, and is represented as an N x n matrix:

$$\mathrm{X} = \begin{bmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1n} \\ \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{x}_{N1} & \mathrm{x}_{N2} & \cdots & \mathrm{x}_{Nn} \end{bmatrix} \qquad (6.1)$$

In pattern recognition terminology, the rows of X are called *patterns* or objects, the columns are called the *features* or attributes, and X is called the *pattern matrix*.

6.2.1.1. Choice of attributes

This very important step depends on the researcher's knowledge of the subject matter. The data for clustering should be described in terms of their characteristics, attributes, class membership, and other such properties. These descriptors collectively are the attributes of the problem. Attributes that are highly correlated add little in terms of distinguishing the data units. At the same time including attributes that have large variations among data units, but are not relevant to the problem will provide misleading results. The choice as to the number of attributes is different for different fields of study. Statisticians and social scientists emphasize parsimony and thus seek to minimize the number of measures attributes. Proper selection of attributes is a difficult, but important task.

6.2.1.2. Scaling and standardization of attributes

Once a decision is made as to the number of attributes to be included for clustering, the next step is to select the type(s) of attribute to be used. Like being said before, the attributes could be quantitative, qualitative or mixed type. The common problem in real data is the lack of homogeneity among attributes of interest. The philosophy of cluster analysis is based on measuring proximity between data points in a multi-dimensional framework. The type of attribute and the scale of measurement influence the measure of similarity calculated for the data points. Most analysis techniques assume homogeneity of data types, whereas real data sets often have mixed types. There are various ways of handling these three variations in calculating the similarity matrix.

Measurement scales could be sequentially ordered as nominal, ordinal, interval and ratio, with the progression reflecting increasing information demands for scale definition. Nominal and ordinal scales are referred to as qualitative attributes, and interval and ratio scales are referred to as quantitative attributes. If the problem at hand has mixed data type, one can reduce the quantitative attributes into qualitative attributes by dichotomizing the quantitative attributes. This strategy reduces the quantitative variable to the lowest common denominator. The process has the risk of losing information that the quantitative attributes may contain. This might be crucial in mathematical terms, but loss of information may not be crucial fir clustering purpose.

Even after the decision has been made as whether to use mixed data type or to convert the attributes into a homogeneous type, there remains the issue of standardisation of attributes. There are two main reasons for standardisation a data matrix. First, the units of measurement of the attributes can arbitrarily affect the similarities among data points. Standardization helps remove the arbitrarily affects. Second, standardization makes attributes contribute more equally to similarities among data points. If the original data matrix, the value of one particular attribute is much greater than the range of other attributes, the attributes with a larger value will carry more weight in determining the similarities among the data points. When this affects the clustering process adversely, the attributes should be standardized to remove the effect.

6.2.2. The clusters – similarities/dissimilarities

A cluster is a group of objects that are more similar to one another than to members of other clusters. The term "similarity" should be understood as mathematical similarity, measured in some well-defined sense. In metric spaces, similarity is often defined by means of a *distance norm*. Distance can be measured among the data vectors themselves, or as a distance form a data vector to some prototypical object of the cluster. The prototypes are usually not known beforehand, and are sought by the clustering algorithms simultaneously with the partitioning of the data.

The measure of similarity is defined for every pair-wise combination of entities to be clustered. The measure interacts with the cluster analysis criteria, so that some measures give identical results with some criterion and distinctly different with another. The combined choice of attributes, data transformation, and similarity measures leads to successful natural grouping. A basic assumption of all clustering methods is that these numerical measures of distance are all comparable to each other. If the similarity for a pair is 100 and for another pair it is 70, then the second pair is more similar than the first pair. There are various ways of handling quantitative, qualitative or mixed type data in calculating the similarity matrix. Romesburg (1988) discusses in detail the various resemblance coefficients that can be calculated for the three types of attributes.

6.2.3. Cluster partition

Since clusters can formally be seen as subsets of the data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy or crisp (hard). Hard clustering methods are based on classical set theory, and require that an object either does or does not belong to a cluster. Hard clustering in a data set X, means partitioning the data into a specified **number** of mutually exclusive subsets of X. The number of subsets (clusters) is denoted by *c*. Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership. The data set X is thus partitioned into *c* fuzzy subsets. In many real situations, fuzzy clustering is more natural than hard clustering, as objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial memberships. The discrete nature of hard partitioning also causes analytical and algorithmic intractability of algorithms based on analytic functional, since these functional are not differentiable.

6.3. Clustering algorithms

Clustering methods can be divided into two main groups: (1) hierarchical clustering methods and (2) non-hierarchical clustering methods.

6.3.1. Hierarchical clustering methods

The hierarchical clustering method can be further divided into two types: (a) agglomerative hierarchical methods and (b) divisive hierarchical methods. The first starts with a disjoint set of entities and merge them by certain rules into fewer and more inclusive clusters, until the formation of a conjoint set. The divisive techniques begin with the conjoint set and partition the sample into smaller and smaller subsets. There are several methods of hierarchical clustering, like the linkage methods and Ward's minimum variance method.

In hierarchical clustering method, there is no provision for reallocation of objects that may have been incorrectly grouped at an early stage. For a particular problem, it is important to try several clustering methods and, within a given method, a couple of different assigning distances (similarities). One should conclude a natural grouping only if the outcomes of several methods are consistent with one another.

6.3.2. Non-hierarchical clustering methods

For a data set of m entities, the hierarchical methods described above provide m nested classifications ranging from m clusters of one member each, to one cluster of m members. Non-hierarchical clustering method is designed to cluster data into a single classification of k clusters, where k is specified apriori or is determined as part of the clustering method. The main idea is to choose some initial partition of data units and then alter cluster memberships to obtain a better partition. The partitioning techniques differ from the hierarchical methods in several ways. First, partitioning leads to non-hierarchical single-rank solutions; second, it allows for correction of poor initial clustering, by iteratively reallocating assignment. In non-hierarchical methods, a set of cases is iteratively partitioned to maximize some predefined criterion function.

There are various methods of non-hierarchical clustering methods:

- Exclusive Clustering - data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster.

- Overlapping Clustering - uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.
- Probabilistic Clustering - use a completely probabilistic approach.

6.3.2.1. K-means clustering method

This is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as barycentres of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function is:

$$\sum_{i=1}^{c} \sum_{k \in A_i} \left\| x_k - v_i \right\|^2 \qquad (6.2)$$

where A_i is a set of objects (data points) in the i-th cluster and v_i is the mean for that points over cluster i. Equation (7.2) denotes actually a distance measure between a data point x_i and the cluster centre v_i. In K-means clustering, v_i is called the cluster prototypes, i.e. the cluster centres:

$$v_i = \frac{\sum_{k=1}^{N_i} x_k}{N_i}, \; x_k \in A_i \qquad (6.3)$$

where N_i is the number of objects in A_i.

Each cluster in the partition is defined by its member objects and by its centroid. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized.

The algorithm is composed of the following steps:

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

In a more mathematical form, the **K-means clustering algorithm** is:

Given the data set X, choose the number of clusters $1 < c < N$. Initialize with random cluster centres chosen from the data set.

Repeat for $l = 1, 2, \ldots$

Step **1** Compute the distances

$$D_{ik}^2 = (\mathrm{x_k} - \mathrm{v_i})^{\mathrm{T}} \ (\mathrm{x_k} - \mathrm{v_i}),\ 1<\mathrm{i}<\mathrm{c},\ 1<\mathrm{k}<\mathrm{N} \qquad (6.4)$$

Step **2** Select the points for a cluster with the minimal distances, they belong to that cluster.

Step **3** Calculate cluster centres

$$v_i^{(l)} = \frac{\sum_{k=1}^{N_i} x_k}{N_i} \qquad (6.5)$$

$$\text{until } v_i^{(l)} = \prod_{k=1}^{n} \max \left| v^{(l)} - v^{(l-1)} \right| \neq 0 \qquad (6.6)$$

Ending Calculate the partition matrix

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres because the calculation can run into wrong results, if the centres "have no data points". It is recommended to run K-means several times to achieve the correct results. To avoid the problem described above, the cluster centres are initialized with randomly chosen data points. If D_{ik} becomes zero for some $\mathrm{x_k}$, singularity occurs in the algorithm, so the initializing centres are not exactly the random data points, they are just near them. If the initialization problem still occurs for some reason, the "lonely" centres are redefined to data points. It is also necessary to take proper care in selecting initial number of clusters, avoiding local minima or misclassification.

6.3.2.2. Fuzzy C-means clustering method

In the considered *k*-means procedure, each data point is assumed to be in exactly one cluster. We can relax this condition and allow each instance to belong to some cluster with some degree, that is introduce a "fuzzy" membership in a cluster, so that a point may belong to several clusters with some degree in the range [0, 1]. This idea is used in the *fuzzy C-means* clustering algorithm. It is based on the *k*-means algorithm but in calculating a clauter's centre the coordinates of each instance are weighed by the value of the membership function.

The fuzzy c-means (FCM) algorithm, also known as fuzzy ISODATA, was originally developed by Dunn [21] and later generalized by Bezdek [22], [23]. The set of all points considered is data $X = \{x_1, x_2 \ldots, x_n\}$ and the vector of prototype centres $\mathbf{v} = \{v_1, v_2 \ldots, v_c\}$, which has to be determined. Therefore, fuzzy partition matrix for c clusters and n data points is:

$$M_{fc} = \{U|\mu_{ik} \in [0, 1]; \sum_{k=1}^{n} \mu_{ik} = 1; 0 < \sum_{k=1}^{n} \mu_{ik} < n\} \qquad (6.7)$$

where $i = 1, 2, \ldots, c$ and $k = 1, 2, \ldots, n$

The goal of fuzzy c-means algorithm is to select U and $\mathbf{v}$ in such a way to minimize the objective function, called *C-means functional*. It is defined by Dunn as:

$$J_m(X, U, \mathbf{v}) = \sum_{i=1}^{c} \sum_{k=1} (\mu_{ik})^{(m)} \|x_k - v_i\|^2{}_A \qquad (6.8)$$

where $m \in [1,\infty]$ is the weighting exponent determining the fuzziness of the clusters, μ_{ij} is the degree of membership of x_k in the cluster I and

$$D_{ikA}^2 = (\mathrm{x_k} - \mathrm{v_i})^{\mathrm{T}} \ (\mathrm{x_k} - \mathrm{v_i}), \ 1<\mathrm{i}<\mathrm{c}, \ 1<\mathrm{k}<\mathrm{N} \qquad (6.9)$$

is a Euclidean measure between the data sample x_k and cluster center v_i.

Statistically, (7.8) can be seen as a measure of the total variance of x_k from v_i. The minimization of the c-means functional (7.8) represents a nonlinear optimization problem that can be solved by using a variety of available methods. The most popular method is a simple Picard iteration through the first-order conditions for stationary points of (7.8), known as the *fuzzy c-means (FCM) algorithm*.

The FCM algorithm computes with the standard Euclidean distance norm, which induces hyper spherical clusters. Hence *it can only detect clusters with the same shape and orientation*.

Notes: If D_{ikA} becomes zero for some x_k, singularity occurs in the algorithm: the membership degree cannot be computed. Also, the correct choice of the weighting parameter (m) is important: as m approaches one from above, the partition becomes hard, if it approaches to infinity, the partition becomes maximally fuzzy, i.e. $\mu_{ik} = 1/c$.

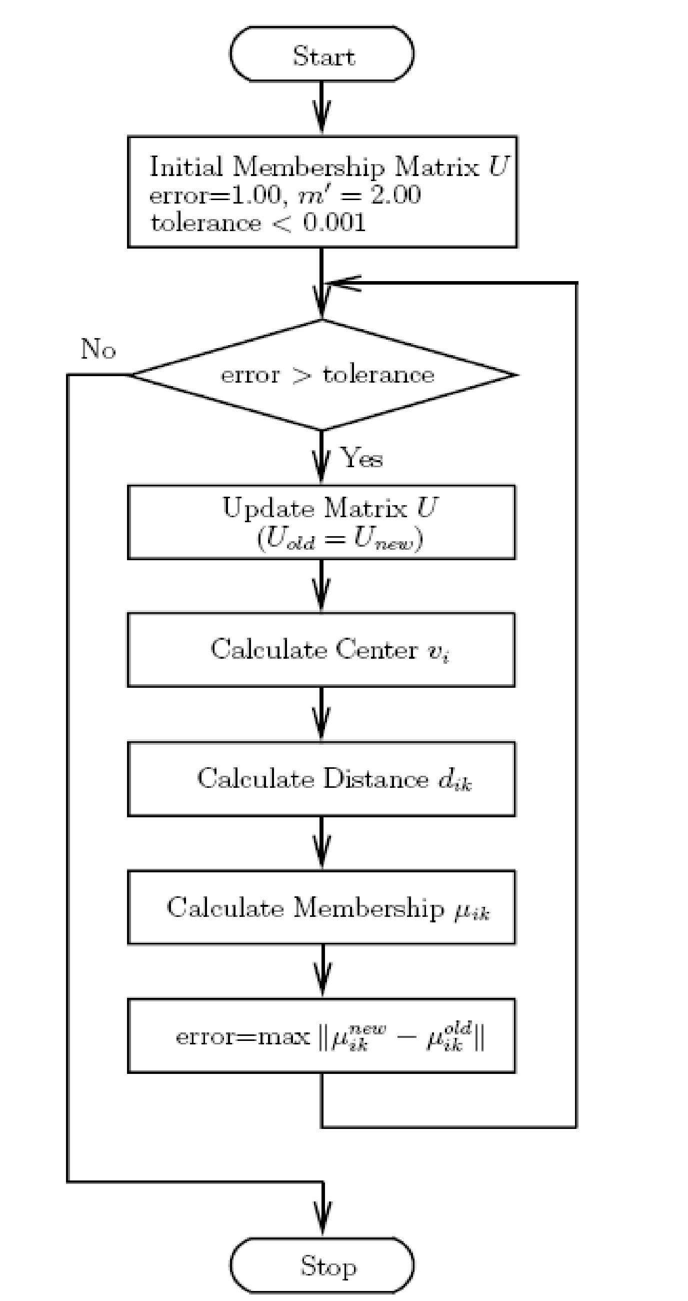


Fig. 7: A flow chart for the fuzzy c-means clustering algorithm.

The **Fuzzy C-means clustering algorithm** is:

Given the data set X, choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\varepsilon > 0$ and the norm-inducing matrix A. Initialize the partition matrix randomly, such that $U^{(0)} \in M_{fc}$.

Repeat for $l = 1, 2, \ldots$

Step **1** Compute the cluster prototypes (means)

$$v_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m x_k}{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m}, \ 1<i<c \qquad (6.10)$$

Step **2** Compute the distances:

$$D_{ik}^2 = (x_k - v_i)^T \ A \ (x_k - v_i), \ 1<i<c, \ 1<k<N \qquad (6.11)$$

Step **3** Update the partition matrix:

$$\mu_{i,k}^{(l)} = \frac{1}{\sum_{j=1}^{c} \left(\left(\frac{D_{ikA}}{D_{jkA}} \right)^{\frac{2}{m-1}} \right.} \qquad (6.12)$$

$$\text{until } \left\| U^{(l)} - U^{(l-1)} \right\| < \varepsilon$$

As already told, data are bound to each cluster by means of a Membership Function, which represents the fuzzy behaviour of this algorithm. To do that, we simply have to build an appropriate matrix named U whose factors are numbers between 0 and 1, and represent the degree of membership between data and centres of clusters.

6.3.2.3. The Gustafson - Kessel algorithm

Gustafson-Kessel clustering algorithm extends the Fuzzy C-means algorithm by employing an adaptive distance norm, in order to detect clusters with different geometrical shapes in the data set. Each cluster has its own norm-inducing matrix A_i, which yields the following inner-product norm:

$$D_{ikA}^2 = (x_k - v_i)^T \ A \ (x_k - v_i), \ 1<i<c, \ 1<k<N \qquad (6.13)$$

The matrices A_i are used as optimization variables in the c-means functional, thus allowing each cluster to adapt the distance norm to the local topological structure of the data. Let A denote a *c*-tuple of the norm-inducing matrices: A = (A_1, A_2 ..., A_c). The objective functional of the GK algorithm is defined by:

$$J_m(X, U, \mathbf{v}, \mathbf{A}) = \sum_{i=1}^{c} \sum_{k=1} (\mu_{ik})^{(m)} \mathrm{D}_{\mathrm{ikAi}}^2 \qquad (6.14)$$

This objective function cannot be directly minimized with respect to A_i, since it is linear in A_i. This means that *J* can be made as small as desired by simply making A_i less positive definite. To obtain a feasible solution, A_i must be constrained in some way. The usual way of accomplishing this is to constrain the determinant of A_i. Allowing the matrix A_i to vary with its determinant fixed corresponds to optimizing the cluster's shape while its volume remains constant:

$$\left\| A_i \right\| = \rho_i, \ \rho > 0 \qquad (6.15)$$

where ρ_i is fixed for each cluster.

So, this clustering method forces that each cluster has its own norm inducing matrix A$_i$, so they are allowed to adapt the distance norm to the local topological structure of the data points. The algorithm uses the Mahalanobis distance norm.

Notes:

- If there is no prior knowledge, ρ_i is 1 for each cluster, so the GK algorithm can find only clusters with approximately equal volumes.
- A numerical drawback of GK is: When an eigenvalue is zero or when the ratio between the maximal and the minimal eigenvalue, i.e. the condition number of the covariance matrix is very large, the matrix is nearly singular. Also the normalization to a fixed volume fails, as the determinant becomes zero. In this case it is useful to constrain the ratio between the maximal and minimal eigenvalue, this ratio should be smaller than some predefined threshold that is in the β parameter.

The **Gustafson - Kessel clustering algorithm** is:

Given the data set X, choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\varepsilon > 0$ and the norm-inducing matrix A. Initialize the partition matrix randomly, such that $U^{(0)} \in M_{fc}$.

Repeat for $l = 1, 2, \ldots$

Step **1** Calculate the cluster centers

$$v_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m x_k}{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m}, \ 1<i<c \tag{6.16}$$

Step **2** Compute the cluster covariance matrices

$$F_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m x_k (x_k - v_i^{(l)})(x_k - v_i^{(l)})^T}{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m}, \ 1<i<c \tag{6.17}$$

Add a scaled identity matrix:

$$F_i = (1-\gamma)F_i + \gamma (F_0)^{\frac{1}{n}} I \tag{6.18}$$

Extract eigenvalues, λ_{ij} and eigenvectors φ_{ij},
find $\lambda_{i;max} = max_j \, \lambda_{ij}$ and set:

$$\lambda_{i;max} = \lambda_{ij}/\beta; \ for\ any\ j\ for\ which \quad \lambda_{i;max}/\lambda_{ij} > \beta \tag{6.19}$$

Reconstruct F_i by:

$$Fi = [\varphi_{i,1} \ldots \varphi_{i,n}]\ diag(\lambda_{i;1} \ldots \lambda_{i;n})[\ \varphi_{i,1} \ldots \varphi_{i,n}]^{-1} \tag{6.20}$$

Step **3** Compute the distances:

$$D^2_{ikAi} = (\mathrm{x_k} - \mathrm{v_i}^{(l)})^{\mathrm{T}}\ [(\rho_{\mathrm{i}}\ \det(\mathrm{F_i}))^{1/\mathrm{n}}\mathrm{F_i}^{-1}]\ (\mathrm{x_k} - \mathrm{v_i}^{(l)}) \tag{6.21}$$

Step **3** Update the partition matrix:

$$\mu_{i,k}^{(l)} = \frac{1}{\sum_{j=1}^{c}\left(\left(\frac{D_{ikA}}{D_{jkA}}\right)^{\frac{2}{m-1}}\right)} \tag{6.12}$$

$$\text{until } \left\|U^{(l)} - U^{(l-1)}\right\| < \varepsilon$$

6.4. Validation

Cluster validity refers to the problem whether a given fuzzy partition fits to the data all. The clustering algorithm always tries to find the best fit for a fixed number of clusters and the parameterized cluster shapes. However this does not mean that even the best fit is meaningful at all. Either the number of clusters might be wrong or the cluster shapes might not correspond to the groups in the data, if the data can be grouped in a meaningful way at all. The determination of the **optimum number of the clusters** is the most important problem in the cluster analysis.

Two main approaches to determining the appropriate number of clusters in data can be distinguished:

- Starting with a sufficiently large number of clusters, and successively reducing this number by merging clusters that are similar (compatible) with respect to some predefined criteria. This approach is called *compatible cluster merging*.

- Clustering data for different values of *c*, and using *validity measures* to assess the goodness of the obtained partitions. This can be done in two ways:

    - The first approach is to define a validity function which evaluates a complete partition. An upper bound for the number of clusters must be estimated (c_{max}), and the algorithms have to be run with each $c \in \{2,$ 3 ..., $c_{max}\}$. for each partition, the validity function provides a value such that the results of the analysis can be compared indirectly. We define *S* as a fuzzy clustering validity function for selecting appropriate number of clusters. The number of clusters is determined so that the smaller *S* means a more compact and separate clustering. The goal should therefore be to minimize the value of *S*. A flow chart is shown in Fig. 8.

- The second approach consists of the definition of a validity function that evaluates individual clusters of a cluster partition. Again, c_{max} has to be estimated and the cluster analysis has to be carried out for c_{max}. The resulting clusters are compared to each other on the basis of the validity function. Similar clusters are collected in one cluster; very bad clusters are eliminated, so the number of clusters is reduced. The procedure can be repeated until there are *bad* clusters.

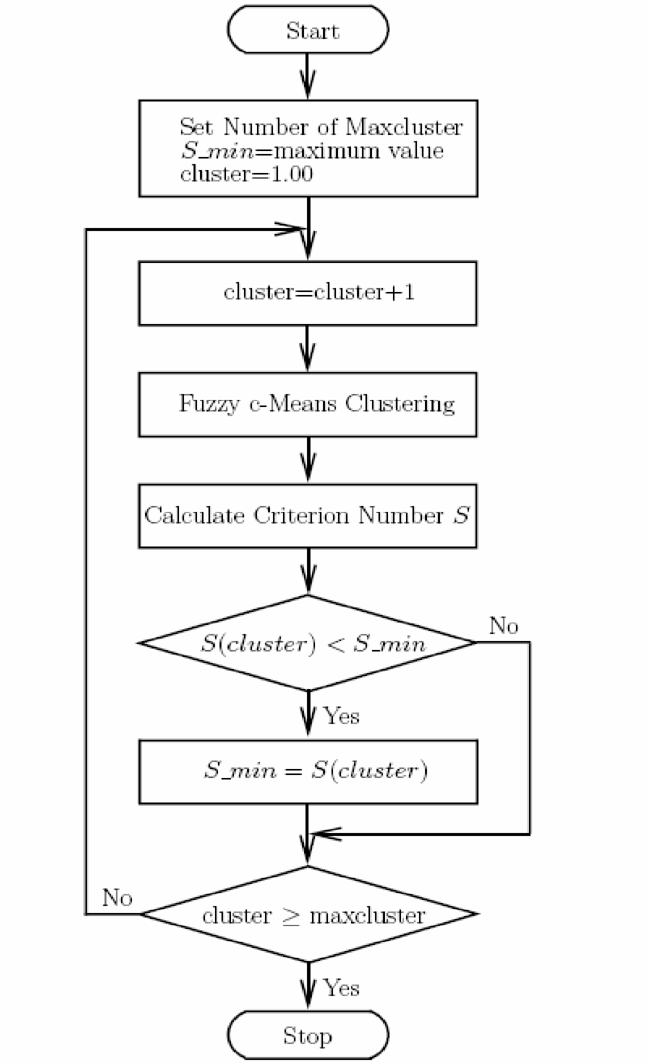


Fig. 8: A flow chart for the determination of appropriate number of clusters.

Different scalar validity measures have been proposed in the literature and some of them are presented below:

1. Partition Coefficient (PC): measures the amount of "overlapping" between clusters. It is defined by Bezdek as follows:

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2 \tag{6.13}$$

where μ_{ij} is the membership of data point j in cluster i. The disadvantage of PC is lack of direct connection to some property of the data themselves. The optimal number of cluster is at the maximum value.

2. Classification Entropy (CE): it measures the fuzziness of the cluster partition only, which is similar to the Partition Coefficient.

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log(\mu_{ij}) \tag{6.14}$$

3. Partition Index (SC): is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster.

$$SC(c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}(\mu_{ij})^{(m)}\left\|x_j - v_i\right\|^2}{N_i\sum_{k=1}^{c}\left\|v_k - v_i\right\|^2} \tag{6.15}$$

SC is useful when comparing different partitions having equal number of clusters. A lower value of *SC* indicates a better partition.

4. Separation Index (S): on the contrary of partition index (SC), the separation index uses a minimum-distance separation for partition validity.

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2\left\|x_j - v_i\right\|^2}{N\min_{i,k}\left\|v_k - v_i\right\|^2} \tag{6.16}$$

5. Xie and Beni's Index (XB): it aims to quantify the ratio of the total variation within clusters and the separation of clusters.

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2\left\|x_j - v_i\right\|^2}{N\min_{i,k}\left\|x_k - v_i\right\|^2} \tag{6.17}$$

The optimal number of clusters should minimize the value of the index.

6. Dunn's Index (DI): this index is originally proposed to use at the identification of "compact and well separated clusters". So the result of the clustering has to be recalculated as it was a hard partition algorithm.

$$DI(c) = \min_{i\in c}\left\{\min_{j\in c, i\neq j}\left\{\frac{\min_{x\in Ci, y\in Cj} d(x,y)}{\max_{k\in C}\left\{\max_{x,y\in C} d(x,y)\right\}}\right\}\right\} \qquad (6.18)$$

The main drawback of Dunn's index is computational since calculating becomes computationally very expansive as *c* and *N* increase.

7. Alternative Dunn Index (ADI): the aim of modifying the original Dunn's index was that the calculation becomes simpler, when the dissimilarity function between two clusters ($\min_{x\in Ci, y\in Cj} d(x,y)$) is rated in value from beneath by the triangle-non equality:

$$d(x, y) \geq \mid d(y, v_j) - d(x, v_j)\mid \qquad (6.19)$$

where v_j is the cluster centre of the *j*-th cluster.

$$ADI(c) = \min_{i\in c}\left\{\min_{j\in c, i\neq j}\left\{\frac{\min_{x_i\in Ci, x_j\in Cj}\left|d(y,v_j) - d(x_i,v_j)\right|}{\max_{k\in C}\left\{\max_{x,y\in C} d(x,y)\right\}}\right\}\right\} \qquad (6.20)$$

The only difference of *SC, S* and *XB* is the approach of the separation of clusters. In the case of overlapped clusters the values of *DI* and *ADI* are not really reliable because of re-partitioning the results with the hard partition method.

6.5. Visualization

The clustering-based data mining tools are important, since they are able to explore structures and classes in the data.

The Principal Component Analysis maps the data points into a lower dimensional space, which is useful in the analysis and visualization of the correlated high-dimensional data. The main drawback is that it is a linear transformation that is not adapted to non-linear data.

In this paper, the attention is focused more on the Sammon mapping method for the visualization of the clustering results, because it preserves inter-pattern distances and it is adapted to non-linear data. This kind of mapping of distances is much closer to the proposition of clustering than simply preserving the variances (like PCA). Anyway, there are two problems with the Sammon mapping application:

- The prototypes of clusters are usually not known apriori, and they are calculated along with the partitioning of the data. These prototypes can be vectors dimensionally equal to the examined data points, but they also can be defined as geometrical objects, i.e. linear or non-linear subspaces, functions. Sammon mapping is a projection method, which is based on the preservation of the

Euclidian inter-point distance norm, so it can be only used by clustering algorithms calculating with this type of distance norm[10],[11].

- The Sammon mapping algorithm forces to find in a high n-dimensional space N points in a lower q-dimensional subspace, such these inter-point distances correspond to the distances measured in the n-dimensional space. This affects a computationally expensive algorithm, since in every iteration step it requires computation of $N\,(N-1)/2$ distances.

To avoid these problems a modified Sammon mapping algorithm is used, described in detail in section 6.5.3.

6.5.1. Principal Component Analysis

The principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The main objectives of PCA are:

1. identify new meaningful underlying variables;
2. discover or to reduce the dimensionality of the data set.

The mathematical background lies in "eigen analysis": The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component.

6.5.2. Sammon mapping

The Sammon mapping method is used for finding N points in a q- dimensional data space, where the original data are from a higher n-dimensional space. The $d_{ij} = d(\mathrm{x}_i;\mathrm{x}_j)$ inter-point distances measured in the n-dimensional space approximate the corresponding $d^*_{ij} = d^*(\mathrm{x}_i;\mathrm{x}_j)$ inter-point distances in the q-dimensional space. This is achieved by minimizing an error criterion, E (called Sammon's stress) [10]:*

$$E = -\frac{1}{\lambda}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(d_{ij}-d^*_{ij})^2}{d_{ij}} \qquad (6.21)$$

Where $\lambda = \sum_{i<j} d_{ij} = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}$ but there is no need to maintain λ for a successful solution of the optimization problem, since as a constant, it does not changes the optimization result.

The minimization of E is an optimization problem in the $N * q$ variables y_{il}, $i = 1, 2,\ldots,N$ $l = = 1, 2,\ldots,q$, as $\mathrm{y}_i = [y_{i1},\ldots, y_{iq}]^T$. At the t-th iteration let to be the rating of y_{il} ,

$$y_{il}(t+1) = y_{il}(t) - \alpha \left[\frac{\dfrac{\partial E(t)}{\partial y_{il}(t)}}{\dfrac{\partial^2 E(t)}{\partial y^2{}_{il}(t)}} \right] \tag{6.22}$$

where α is a nonnegative scalar constant (recommended $\alpha \approx 0.3$ - 0.4), this is the step size for the gradient search. A drawback of this gradient-descent method is a possibility to reach a local minimum in the error surface, while searching for the minimum of E, so experiments with different random initializations are necessary. The initialization can be estimated based on information which is obtained from the data.

To conclude, the Sammon mapping is non-linear projection method, which reveals the structure present in data, but it's drawbacks are that it lacks generalization (new points cannot be added to the obtained map without recalculating it) and that it operates on all inter-point distances, so the complexity of finding the mapping is very high.

6.5.3. Fuzzy Sammon mapping

Avoiding the drawbacks of Sammon's mapping, the modified mapping method uses the basic properties of fuzzy clustering algorithms where only the distance between the data points and the cluster centres are considered to be important [14]. The modified algorithm takes into account only N x c distances, where c represents the number of clusters, weighted by the membership values:

$$E_{fuzz} = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ki})^m (d(x_k, v_i) - d^*_{ki})^2 \tag{6.23}$$

where $d(\mathrm{x}_k; \mathrm{v}_i)$ represents the distance between the x_k datapoint and the v_i cluster centre measured in the original n-dimensional space, while $d^*_{ki} = d^*(y_k, z_i)$ represents the Euclidian distance between the projected cluster centre z_i and the projected data y_k.

This means, in the projected two-dimensional space every cluster is represented by a single point, independently to the form of the original cluster prototype, v_i. The resulted algorithm is similar to the original Sammon mapping, but in this case, in every iteration after the adaptation of the projected data points, the projected cluster centres are recalculated based on the weighted mean formula of the fuzzy clustering algorithms.

The distances between the projected data and the projected cluster centres are based on the normal Euclidian distance measures. The membership values of the projected data can be also plotted based on the classical formula of the calculation of the membership values:

$$\mu^*_{ki} = \frac{1}{\sum_{j=1}^{c} \left(\dfrac{d^*(x_k, \eta_i)}{d^*(x_k, v_j)} \right)^{\frac{2}{m-1}}} \tag{6.24}$$

and $U^* = \left[\mu_{ki}^*\right]$ is the partition matrix containing the recalculated memberships. The resulted plot will only be an approximation of the original high dimensional clustering in two dimensions. The quality of this rating can be evaluated by determining the maximal value of the mean square error between the original and the re-calculated membership values:

$$P = \left\| U - U^* \right\| \tag{6.25}$$

The **Fuzzy Sammon mapping algorithm** is presented below.

1. Initialize the projected data points by y_k PCA based projection of x_k and compute the projected cluster centres by:

$$z_i = \frac{\sum_{k=1}^{N} (\mu_{ki})^m y_k}{\sum_{k=1}^{N} (\mu_{ki})^m} \tag{6.26}$$

   and compute the distances with the use of these projected points $D^* = \left[d_{ki}^* = d(y_k, z_i)\right]_{Nxc}$

2. As long as ($E_{fuzz} > \varepsilon$) and ($t \leq$ maxstep)

```
a.  { for ($i = 1 : i \cdot c : i++$
        { for ($j = 1 : j \cdot N : j++$
            {Compute $\frac{\partial E(t)}{\partial y_{il}(t)}, \frac{\partial^2 E(t)}{\partial y^2{}_{il}(t)}, \Delta y_{il} = \Delta y_{il} + \left[\frac{\frac{\partial E(t)}{\partial y_{il}(t)}}{\frac{\partial^2 E(t)}{\partial y^2{}_{il}(t)}}\right]$}
        }
        $y_{il} = y_{il} + \Delta y_{il} \forall i = 1,...,N, l = 1,...,q$
        Compute $z_i = \frac{\sum_{k=1}^{N} (\mu_{ki})^m y_k}{\sum_{k=1}^{N} (\mu_{ki})^m}$
        $D^* = \left[d_{ki}^* = d(y_k, z_i)\right]_{Nxc}$
        }
    Compute $E_{fuzz} = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ki})^m (d(x_k, v_i) - d_{ki}^*)^2$
```

Chapter 7

Tests and results

Summary

7.1. Introduction

Chapter 2 explained the characteristics and strategies of the hedge funds and the various studies existent in the literature. In chapter 3 the approximate Bayesian estimation theory is systematically investigated. Following the simplest case, the celebrated Kalman filter is briefly derived, followed by the discussion of optimal nonlinear filtering. Chapter 4 discusses a popular numerical approximation technique - Monte Carlo approximation and sequential sampling method - which results in various forms of particle filters. Chapter 5 explained the fundamentals of wavelet analysis and the basis for time-scale decomposition. In chapter 6 the cluster analysis techniques are presented; here were given different methods of clustering, which will be used in this chapter during the experiments, in order to compare, group and find structures in the various models for the hedge funds returns.

This chapter begins with an overview of the hard- and software used in the experiments. In addition, a detailed description of the dataset is given. Then, methodology used for the experiments and their results are explained.

The simulations and their results are divided into two main groups:

- Data + Clustering → Result 1
    - Artificial data + Clustering → Result 1.1.x
    - Real life data + Clustering → Result 1.2.x
- Data + Filtering → Result 2

- Artificial data + Filtering → Result 2.1.x

7.2. Hardware and software

The hardware used for the experiments is a personal computer with the following specifications:

- Toshiba Satellite M40 – Technologie Mobile Intel Centrino – Processeur Intel Pentium 1.73 GHz,
- 1 GB memory,
- 80GB Hard-Disk,

The software used for the experiments is the following:

- Microsoft Windows XP operating system,
- Microsoft Excel 2003,
- Matlab 6.5 Release 13,
- Fuzzy clustering Toolbox,
- REBEL Toolkit,
- Optimization Toolbox 2.2,
- Statistics toolbox 4.0,
- Wavelet Toolbox 2.2.

For more details, see Appendix.

7.3. Description of data sets

Two datasets were used for the empirical analysis: an *artificial dataset* and a *real-life dataset*.

Before explaining the method used for generating of the *artificial dataset*, the Sharpe's approach and some economical background must be presented. This will be helpful for understanding the process of obtaining the artificial returns.

Factor models for hedge fund strategies: Revisiting Sharpe's approach

In 1992 W. Sharpe introduced a unifying framework for such style models in an effort to describe active management strategies in equity mutual funds. In his model, he describes a certain active investment style as a linear combination of a set of asset class indices. In other words, an active investment strategy is a linear combination of passive, i.e. long-only, buy-and-hold, strategies. Fung and Hsieh were the first to extend Sharpe's model to hedge funds in 1997. They employed techniques similar to those Sharpe had applied to mutual funds five years earlier, but introduced short selling, leverage and derivatives – three important techniques employed by hedge funds - into their model. The resulting factor equation would account for all hedge fund return variation that derives from risk exposure to the risk factors of various asset classes. Adding alpha to the equation, it allows us to decompose hedge fund return as:

Hedge fund excess return = Manager's alpha + Σ (β_i * Factor$_i$) + random fluctuations (7.1)

Fung and Hsieh identified five risk factors, which they defined as modelling Global Macro, Systematic Trend-Following, Systematic Opportunistic, Value, and Distressed Securities. They further argued that hedge fund strategies are highly dynamic and create option like non-linear, contingent return profiles. These non-linear profiles, they argued, cannot be modelled in simple asset class factor models. As the formula above describes, we infer the hedge funds' alphas by measuring and subtracting out the betas times the beta factors. The obtained value of alpha therefore depends on the chosen risk factors. If we leave out a relevant factor in the model, the alpha will come out as fictively high. As a consequence, some of the returns not accounted for by these models are unaccounted beta rather than alpha. Surely, an incomplete model of systematic risk factors doesn't mean those additional risk factors do not exist; only that we do not yet know how to model them. Therefore the formula above on hedge fund returns should actually read as follows:

Hedge fund return = Manager's alpha + Σ (β_i * Factor$_i$ (modelled)) + Σ (β_i * Factor$_i$ (unmodelled) + random fluctuations (7.2)

As explained before, hedge fund returns can be expressed as the sum of the exposure to the market, which is measured by beta β (which is the sum between traditional betas - normal returns generated from exposure to rewarded market risk - and alternative betas - normal returns generated from exposure to other systematic risks), and the abnormal returns, which are defined by alpha. See chapter 2 for more details.

Sharpe's coefficient α is linked and correlated to β. The statistical study of instability is useful to better select funds; statistical studies and tests do not give a clear-cut result for hedge funds. For hedge funds, the β process is usually random walk

In order to analyse the dynamic evolution of hedge funds returns, the exposure and hereby the behaviour of hedge fund managers, the Random Coefficient Model with time-varying alphas and betas was chosen to generate more sets of **artificial data** which will be used further on during the experiments.

The Random Coefficient model (RCM) assumes that the state values alpha and beta vary randomly around a steady state mean: mean (α) and mean (β). Economically, this means that the exposure to the market deviates randomly and for a short time, from a long term mean. v_t and n_t are normally distributed random numbers with a mean of zero: $v_t \sim N(0, \sigma_n^2)$, $n_t \sim N(0, \sigma_n^2)$. For economical approach, R_i represent the total returns (the rentability), alpha the abnormal returns due to the manager's skill, beta the hedge funds' exposure to different risk factors, and F_i the different risk factors, which are explained in detail in chapter 2.

$$R_i = \text{mean}(\alpha) + \sum F_i \beta_i + \sqrt{\text{variance } \alpha_i} * v_t \qquad (7.3)$$

$$\beta_i = \text{mean}(\beta) + \sqrt{\text{variance}\beta_i} * n_t \qquad (7.4)$$

The process of obtaining the artificial returns consists of several steps that will be explained further, in sections 7.5.1.1.1 and 7.5.1.2.1. (Artificially generated data used for the clustering and filtering experiments).

The **real-life dataset** is the dataset called *returns*, and contains real-life continuously compounded returns obtained from the TASS and HFR databases (see Chapter 2, Appendix 4 and Appendix 5 for more details).

The TASS database of hedge funds used consists of monthly returns and accompanying information for 229 hedge funds. The index starts at 31st December 1992 and ends the 31st December 2003. The HFR database of hedge funds used consists of monthly returns and accompanying information for 268 hedge funds. The index starts at 31st December 1993 and ends the 31st December 2003.

7.4. Methodology

For the clustering analysis I chose an appropriate software package in order to acquire intuition for the theory and to conduct experiments. The software package used is the Fuzzy Clustering and Data Analysis Toolbox, which is a collection of Matlab functions. Its purpose is to divide a given data set into subsets, under different initial assumptions. A certain number of simulations and experiments with the proposed models on artificially generated data and real-life data from the TASS and HFR hedge fund databases were effectuated. I studied the results and checked whether the models gave an accurate estimation of hedge funds returns time-series. The simulations were divided in two categories: simple clustering (testing different methods and obtaining the optimal number of clusters) and filtering (testing the different algorithms presented in the theory). The filtering + clustering approach (in order to analyse how the filtering affects the data, if this process affects the initial clusters and if it diminishes the error) will be treated in a future work. Each simulation and validity test is repeated a significant number of times in order to get a reliable notion of the performance.

For the Gaussian Approximate Bayesian Estimation – Kalman Filter Framework, I chose an appropriate software package in order to acquire intuition for the theory and to conduct experiments. The software package used is the ReBEL toolkit, which is a Matlab toolbox designed to facilitate the sequential estimation in general state space models. ReBEL is developed and maintained by Rudolph van der Merwe.

7.5. Simulations

7.5.1. Clustering

The aim of these simulations is to present the differences, the usefulness and effectiveness of the partitioning clustering algorithms by partitioning different data sets (artificially generated and real data sets). Section 7.5.1.1 presents the clustering algorithms that are compared based on numerical results (validity measures). Section

7.5.1.2 deals with the problem of finding the optimal number of clusters; this information is rarely known *apriori*. At the beginning some artificially generated data sets are used and analysed, in order to validate the clustering methods, and afterwards two real data sets (TASS and HFR hedge fund databases) are treated.

The attempt is to classify hedge funds with a unified approach, to find similarities in the hedge funds returns evolution during time, and to try to group them together, in certain clusters. The investment process a manager uses to produce returns and manage risk can be a dominant factor in a fund's risk return profile. The clustering framework is used to identify groupings or classifications where common factors exist. Detailed classifications are valuable in comparing the risk and return characteristics of similar funds that face common factors.

7.5.1.1. Comparing the clustering methods

First of all it must be mentioned, that all these algorithms use random initialization, so different running issues in different partition results, i.e. values of the validation measures. On the other hand the results hardly depend from the structure of the data, and no validity index is perfect by itself for a clustering problem. Several experiment and evaluation are needed.

7.5.1.1.1. Artificial generated data

As explained before, in section 7.3, the model used for the experiments is the Random Coefficient model (RCM). It assumes that the state values alpha (the abnormal returns due to the manager's skill) and beta (the hedge funds' exposure to different risk factors) vary randomly around a steady state mean: mean (α) and mean (β). R_i represent the total returns (the rentability) and F_i the different risk factors. Economically, this model supposes that the exposure to the market deviates randomly and for a short time, from a long term mean.

The process of obtaining the artificial returns consists of several steps:

- Four curves weighted by a random coefficient are created, in order to simulate 4 different risk factors: F1, F2, F3, F4
- Four different labels are created, in order to distinguish 4 different hedge funds exposures to the different risk factors. To each label corresponds a certain return; each return is drawn from a normal distribution with mean μ and a standard deviation equal to the estimated local volatility - which is the square-root of the variance rate. The standard deviation is a statistic that tells you how tightly all the various examples are clustered around the mean in a set of data.

- Each initially generated curve, corresponding to the different risk factors, is multiplied with the corresponding randomly generated beta (the hedge funds' exposure to different risk factors) which varies randomly around a steady state mean: mean (β). n_t are normally distributed random numbers with a mean of zero $n_t \sim N(0, \sigma_n^2)$

$$\beta_j = \text{mean}(\beta) + \sqrt{\text{variance}\beta_i} * n_t \qquad (7.5)$$

- The sum of these products is computed

$$\sum F_i \beta_{ij} \qquad (7.6)$$

- In order to obtain the total return, we compute:

$$R_{ij} = \text{mean}(\alpha) + \sum F_i \beta_{ij} + \sqrt{\text{variance}\ \alpha_i} * v_t \qquad (7.7)$$

where v_t are normally distributed random numbers with a mean of zero $v_t \sim N(0, \sigma_n^2)$.

The experiments are made supposing standard, constant variances for alpha and beta, and some predefined initial curves for simulating the risk factors. Only the standard deviations are varying, from lower values, to higher values. The idea is to present the differences, the usefulness and effectiveness of the partitioning clustering algorithms by partitioning our artificially generated data sets.

For the experiments we have chosen the following numerical values:

- 4 initial curves weighted by a random coefficient are created, in order to simulate 4 different risk factors;
- Each curve, corresponding to a certain label, is repeated 500 times;
- Each curve consists of 120 points, in order to properly simulate the number of returns (the databases of hedge funds used consists of monthly returns and accompanying information for 229 hedge funds. The index starts at 31st December 1992 and ends the 31st December 2003, so during 10 years, i.e. 120 values);
- The constant variances for alpha and beta are established at:
  - Variance Alpha= [0.5 0 0.2 -0.8] ;
  - Variance Beta (label 1)=[0.05 0.05 0.05 0.0];
  - Variance Beta (label 2)=[0.15 0.0 0.0 0.0];
  - Variance Beta (label 3)=[-0.5 0.0 0.0 0.5];
  - Variance Beta (label 4)=[0.5 0.0 0.0 0.6];
- The standard deviation is varied for each experiment, from lower to higher values, in order to check the robustness of our algorithms.

So, to summarize, our artificially generated data set contains 4 classes (which we call labels) of 120 instances each; each class is repeated 500 times, where a class refers to a certain different type of hedge fund return. The predicted attribute for our clustering analysis is the class (or the label).

Case1. For the first experiment, the following numerical values have been chosen:

Standard Deviation Alpha = [0.05 0.05 0.05 0.05];
Standard Deviation Beta (Label 1… Label 4) = 0.05;

In Fig. 9 the four initial curves weighted by a random coefficient are described. Their randomly evolution during time can be observed. In our model they represent the risk factors that influence and affect our returns. We have used this randomly coefficient approach, in order to propose a very robust methodology that is not influenced by some particular values for the risks.

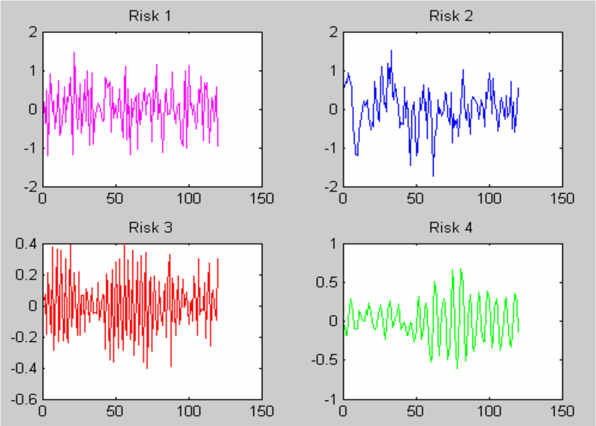


Fig.9. Random generated risk factors (variance of alpha and beta = 0.05)

The artificially generated returns are shown in Fig.10.

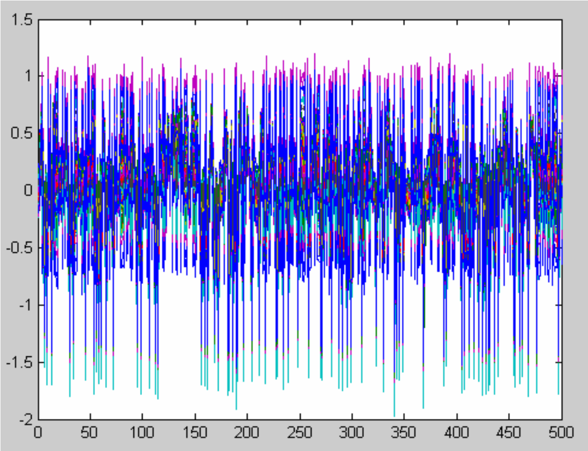


Fig.10. Total generated return (variance of alpha and beta = 0.05)

In Fig. 11 the result of the Principal Component Projection (PCA) of the K-means clustering algorithm is presented. The four clusters are very well determined, as it can be seen in the figure.

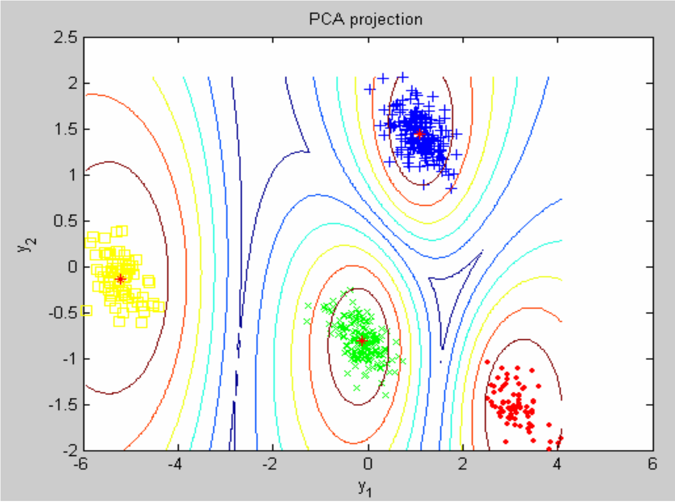


Fig.11. Result of PCA projection by the generated data set (variance of alpha and beta = 0.05)

In Fig. 12, respectively Fig. 13, are presented the results of the Fuzzy Sammon and Sammon mapping for our artificially generated data.

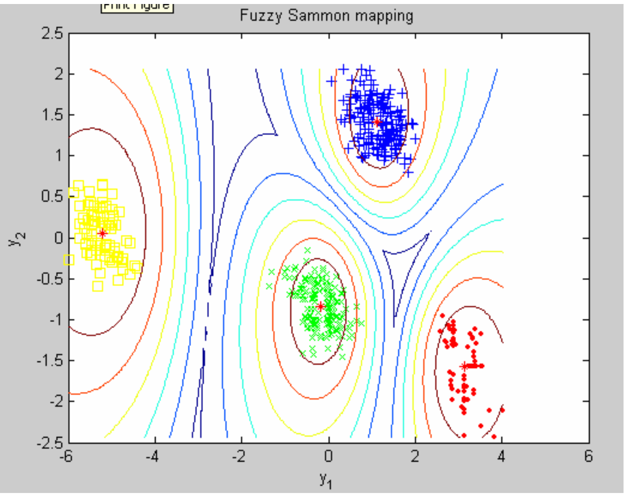


Fig.12. Result of Fuzzy Sammon projection by the generated data set (variance of alpha and beta = 0.05)

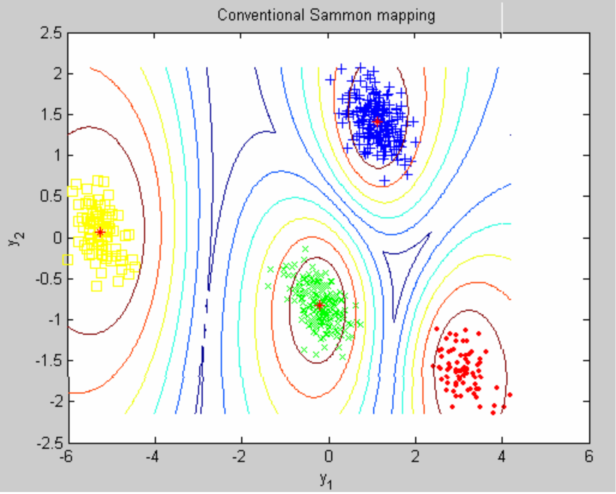


Fig.13. Result of Sammon projection by the generated data set (variance of alpha and beta = 0.05)

A projection evaluation function uses the results and the parameters of the clustering and the visualization functions. It evaluates the projected data. The distances between projected data and projected cluster centres are based on the Euclidean norm, so the function calculates only with a 2-by-2 identity matrix, generates the pair-coordinate points, calculates then new partition matrix, and draws a contour-map by selecting the points with the same partitioning values. It calculates the relation-indexes defined on the ground of (6.25):

$$P = \left\|\overline{U - U^*}\right\|, \sum_{k=1}^{N}\overline{\mu_k^2}, \sum_{k=1}^{N}\overline{\mu_k^{2*}}, E \qquad (7.8)$$

where

- P is the maximal value of the mean square error between the original and the re-calculated membership values (see Fuzzy Sammon mapping):

  $$P = \left\|U - U^*\right\|$$

- μ_{ij} are the degree of membership of x_k in the cluster I
- μ_{ki}^* are the membership values of the projected data.
- E is the error criterion, (called Sammon's stress)

Considering that projected figures are only approximations of the real partitioning results, the difference between the original and the projected partition matrix is also represented, and on the other hand one can observe the difference between the PCA, Sammon's mapping and the Modified Sammon Mapping (see Table 3. the relation-indexes)

	P	$\sum_{k=1}^{N} \overline{\mu_k^2}$	$\sum_{k=1}^{N} \overline{\mu_k^{2*}}$	E
PCA	0.0161	1.0000	0.9394	0.0039
Sammon	0.0202	1.0000	0.9243	0.0024
Fuzzy Sammon	0.0113	1.0000	0.9200	0.0016

Table3. Relation-indexes on generated data set (variance of alpha and beta = 0.05)

As Table3 shows, Fuzzy Sammon Mapping has better projection results by the value of P than Principal component Analysis, and it is computationally cheaper than the original Sammon Mapping. The original Sammon's stress for all the three techniques is calculated, in order to be able to compare them.

Case2. The numerical values chosen are:

Standard Deviation Alpha = [0.5 0.5 0.5 0.5];
Standard Deviation Beta (Label 1… Label 4) = 0.5;

The simulation is effectuated as in Case1. The numerical difference consists in the greater value for the standard variations of alpha and beta. The figures show the k-means clustering results.

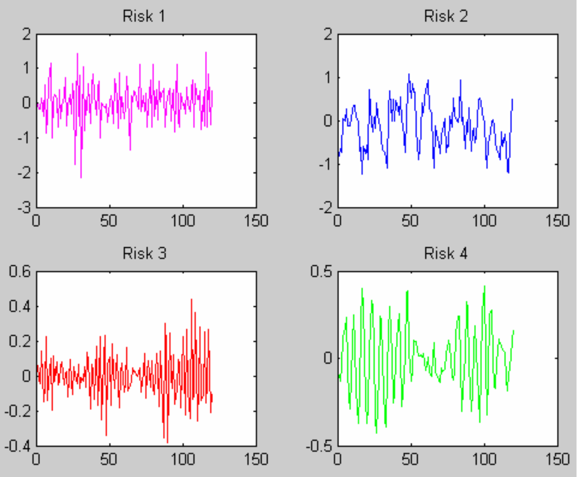


Fig.14. Random generated risk factors (variance of alpha and beta = 0.5)

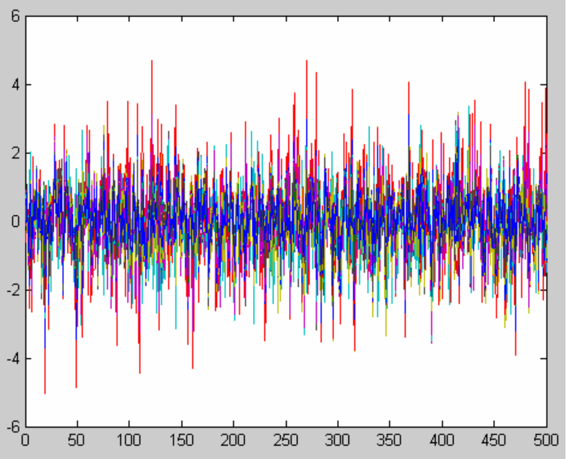


Fig.15. Total generated return (variance of alpha and beta = 0.5)

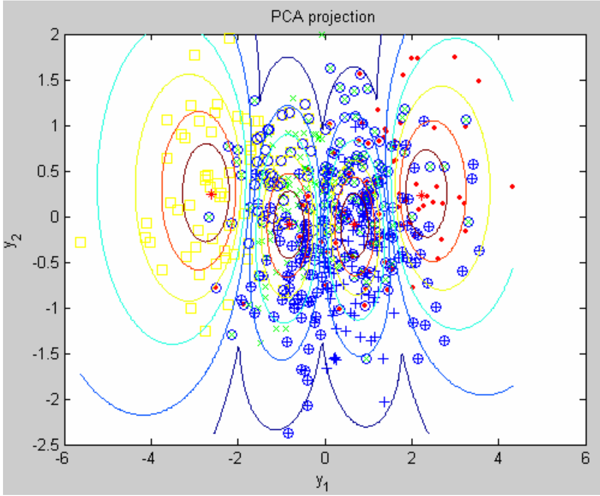


Fig.16. Result of PCA projection by the generated data set (variance of alpha and beta = 0.5)

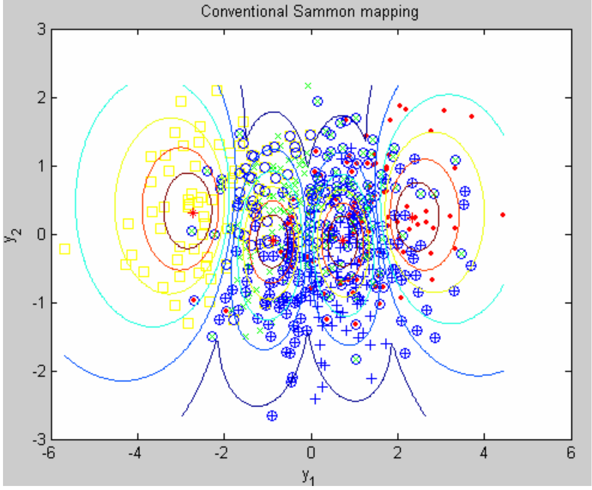


Fig.17. Result of Sammon projection by the generated data set (variance of alpha and beta = 0.5)

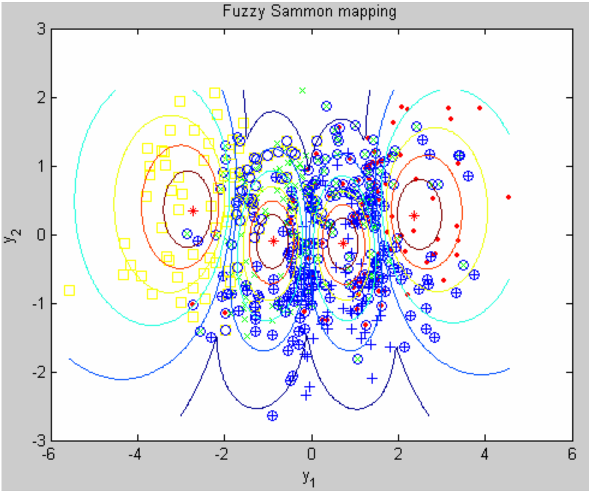


Fig.18. Result of Fuzzy Sammon projection by the generated data set (variance of alpha and beta = 0.5)

	P	$\sum_{k=1}^{N}\overline{\mu_k^2}$	$\sum_{k=1}^{N}\overline{\mu_k^{2*}}$	E
PCA	0.1538	1.0000	0.5708	0.0311
Sammon	0.1658	1.0000	0.5459	0.0202
Fuzzy Sammon	0.1466	1.0000	0.4911	0.0205

Table4. Relation-indexes on generated data set (variance of alpha and beta = 0.5)

As Table4 shows, Fuzzy Sammon Mapping has better projection results by the value of P than Principal component Analysis and conventional Sammon mapping. The original Sammon's stress for all the three techniques is calculated, in order to be able to compare them; it can be observed that Sammon and Fuzzy Sammon mapping in this particular case give similar results.

Case3. The numerical values chosen are:

Standard Deviation Alpha = [1 1 1 1];
Standard Deviation Beta (Label 1… Label 4) = 1;

The simulation is effectuated as in the previous examples. The numerical difference consists in the greater value for the standard variations of alpha and beta. The figures show the Fuzzy C-means clustering results.

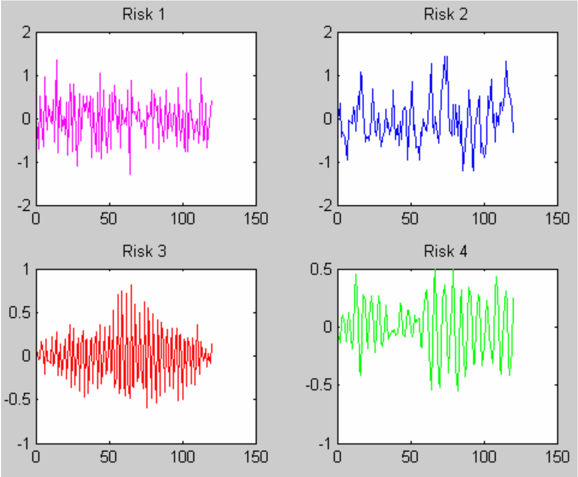


Fig.19. Random generated risk factors (variance of alpha and beta = 1)

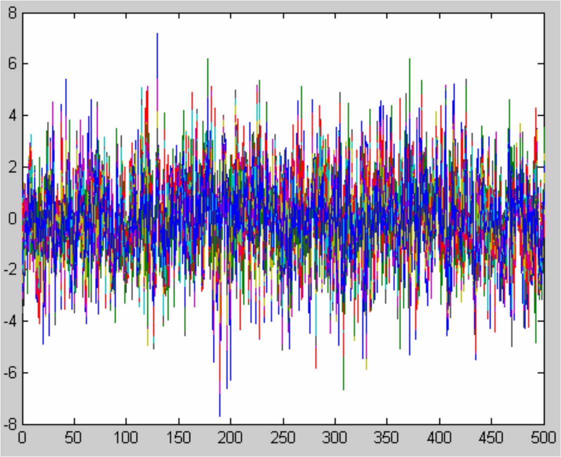


Fig.20. Total generated return (variance of alpha and beta = 1)

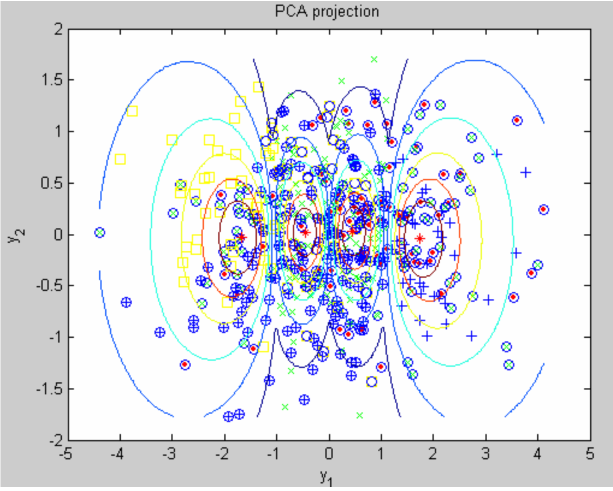


Fig.21. Result of PCA projection by the generated data set (variance of alpha and beta = 1)

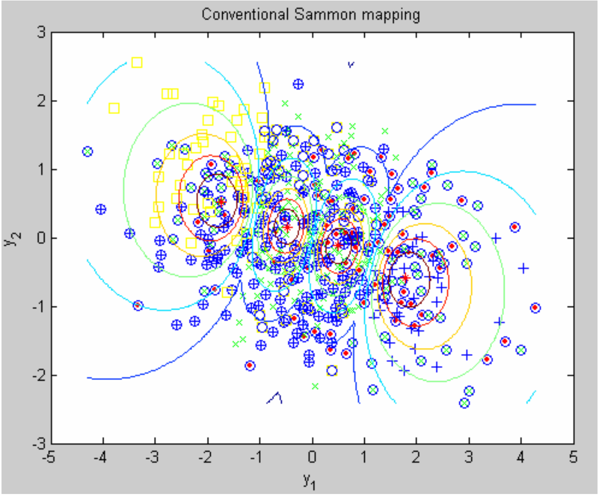


Fig.22. Result of Sammon projection by the generated data set (variance of alpha and beta =1)

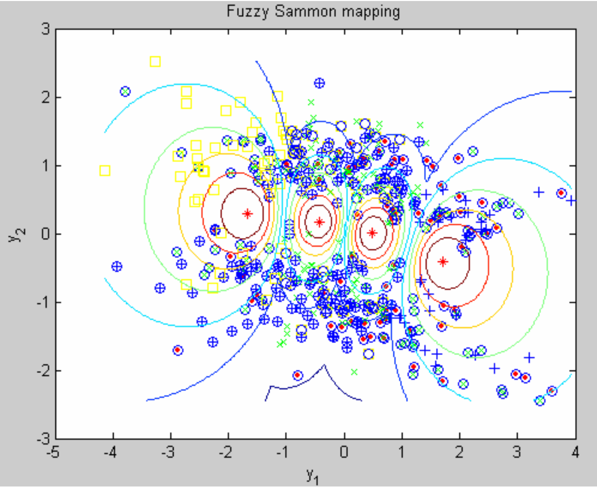


Fig.23. Result of Fuzzy Sammon projection by the generated data set (variance of alpha and beta = 1)

	P	$\sum_{k=1}^{N} \overline{\mu_k^2}$	$\sum_{k=1}^{N} \overline{\mu_k^{2*}}$	E
PCA	0.0666	0.7008	0.5352	0.0565
Sammon	0.0493	0.7008	0.5014	0.0330
Fuzzy Sammon	0.0073	0.7008	0.4051	0.0636

Table5. Relation-indexes on generated data set (variance of alpha and beta = 1)

As Table5 shows, Fuzzy Sammon mapping and conventional Sammon mapping have better projection results by the value of P than Principal component Analysis. The obtained values for the Sammon's stress show that the minimum value corresponds to conventional Sammon.

Conclusions

It must be underlined that the "advanced" algorithms do not have always the best results. It depends on the underlying structure of the data. During the simulations, for the comparison, many independent runs were estimated with each algorithm. Fuzzy C-means and Gustafson-Kessel returned always with the same minimum, while the results of K-means depend from the initialization. The main problem of K-means algorithm is that the random initialization of centres, because the calculation can run into wrong results, if the centers "have no data points". In order to avoid this problem I run K-means several times to achieve the correct results and the cluster centres were initialized with randomly chosen data points.

In this section only certain results are presented, by matter of space. The visualization of the clusters is also very important matter; the fuzzy Sammon mapping has proved very good projection results in comparison to the other techniques.

7.5.1.1.2. Real life data

- **HFR database**

The Hedge Fund Research (HFR) has twenty-six categories of hedge funds. Some of these categories are merely a type of financial instrument or a geographic area for investment. This classification can be reorganized into eleven categories as shown in Fig. 3, section 2.4.2. Some of the categories have further classification. The HFR database used consists of monthly returns and accompanying information for 258 hedge funds. The index starts at 31st December 1993 and ends the 31st December 2003.

1. **K-means clustering algorithm**

Each point in the figures is coloured according to the style membership of the fund.

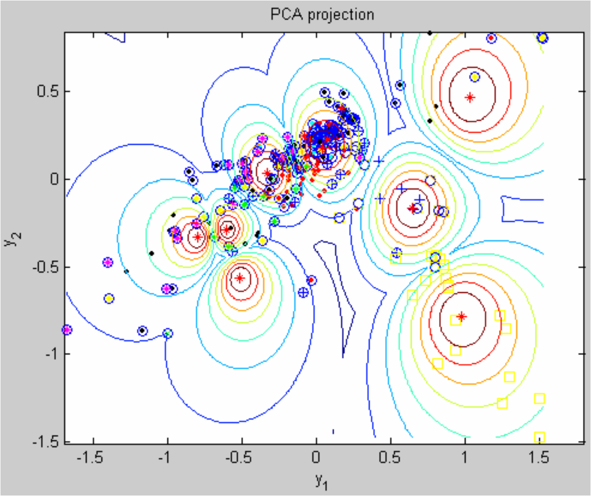


Fig.24. Result of PCA projection by the price returns of HFR database over the results obtained with the K- means clustering algorithm

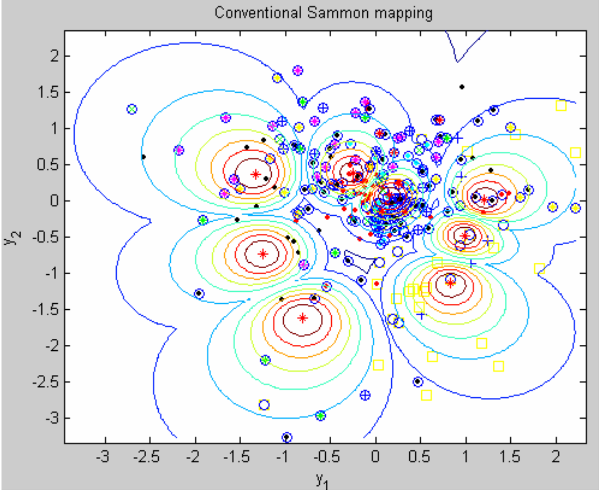


Fig.25. Result of conventional Sammon projection by the price returns of HFR database over the results obtained with the K- means clustering algorithm

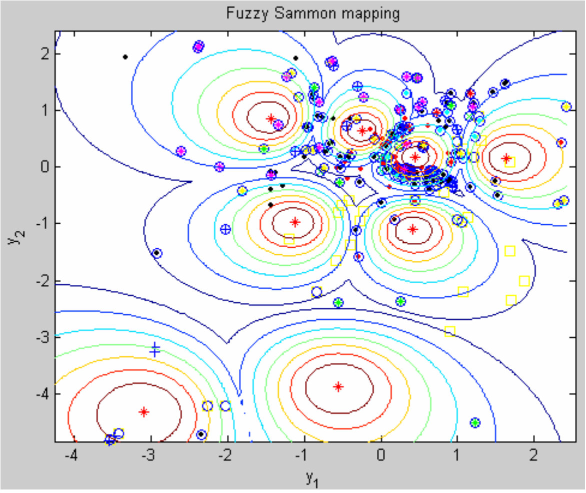


Fig.26. Result of Fuzzy Sammon projection by the price returns of HFR database over the results obtained with the (K- means clustering algorithm

	P	$\sum_{k=1}^{N} \overline{\mu_k^2}$	$\sum_{k=1}^{N} \overline{\mu_k^{2*}}$	E
PCA	0.1453	1.0000	0.5414	0.3192
Sammon	0.1477	1.0000	0.3879	0.1713
Fuzzy Sammon	0.1041	1.0000	0.4014	0.4007

Table6. Relation-indexes on the price returns of HFR database (K- means clustering algorithm)

As Table6 shows, the Fuzzy Sammon Mapping has better projection results by the value of P than the Principal component Analysis projection applied to the results of the K-means clustering method and conventional Sammon mapping over the same results. The original Sammon's stress for all the three techniques is calculated, in order to be able to compare them; it can be observed that conventional Sammon mapping in this particular case gives the best results.

As it can be noticed (the PCA projection gives better results than the conventional Sammon mapping projection), the "advanced" visualization algorithms do not have always the best results. It depends on the underlying structure of the data. The main problem of K-means algorithm is that the random initialization of centres, because the calculation can run into wrong results, if the centers "have no data points". In order to avoid this problem I run K-means several times to achieve the correct results and the cluster centres were initialized with randomly chosen data points.

2. Fuzzy C-means clustering algorithm

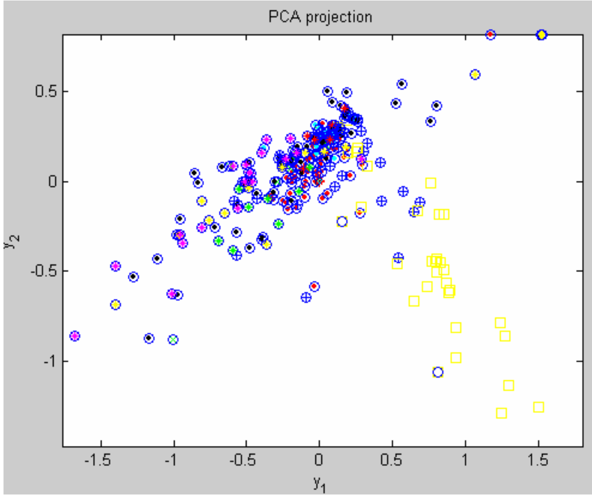


Fig.27. Result of PCA projection by the price returns of HFR database over the results obtained with the Fuzzy C - means clustering algorithm

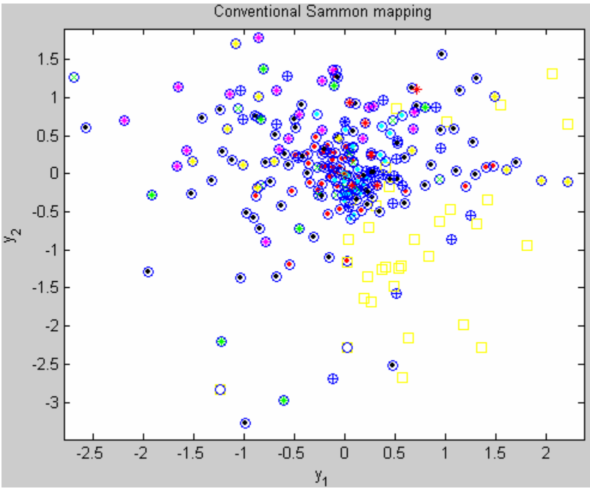


Fig.28. Result of Sammon projection by the price returns of HFR database over the results obtained with the Fuzzy C - means clustering algorithm

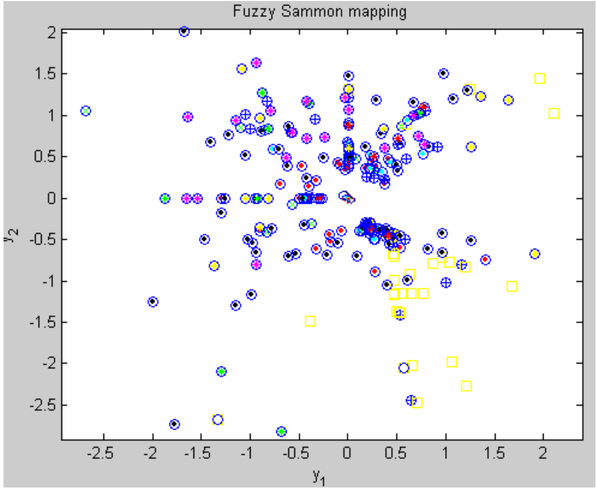


Fig.29. Result of Fuzzy Sammon projection by the price returns of HFR database over the results obtained with the Fuzzy C - means clustering algorithm

	P	$\sum_{k=1}^{N}\overline{\mu_k^2}$	$\sum_{k=1}^{N}\overline{\mu_k^{2*}}$	E
PCA	0.1250	0.8500	0.5414	0.3192
Sammon	0.0877	0.8500	0.3879	0.1713
Fuzzy Sammon	0.0041	0.8500	0.4014	0.1013

Table7. Relation-indexes on the price returns of HFR database over the results obtained with the Fuzzy C- means clustering algorithm

As Table7 shows, the Fuzzy Sammon Mapping has better projection results by the value of P than Principal component Analysis and conventional Sammon mapping. The original Sammon's stress for the Fuzzy Sammon Mapping gives the best results. The best stable results has the Fuzzy C-means clustering for this data set.

3. Gustafson - Kessel clustering algorithm

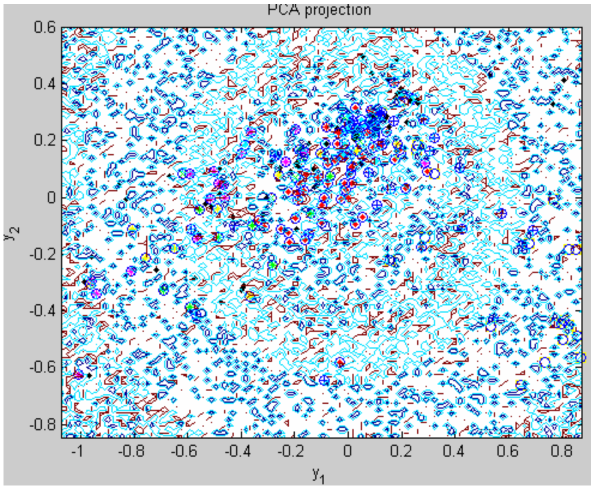


Fig.30. Result of PCA projection by the price returns of HFR database over the results obtained with the Gustafson - Kessel - clustering algorithm

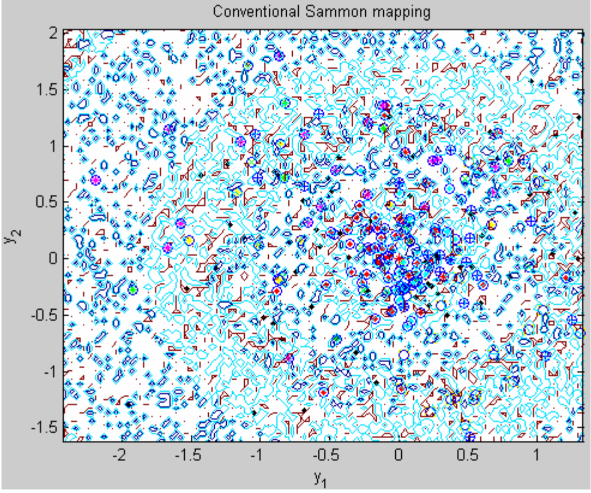


Fig.31. Result of Sammon projection by the price returns of HFR database over the results obtained with the Gustafson - Kessel clustering algorithm

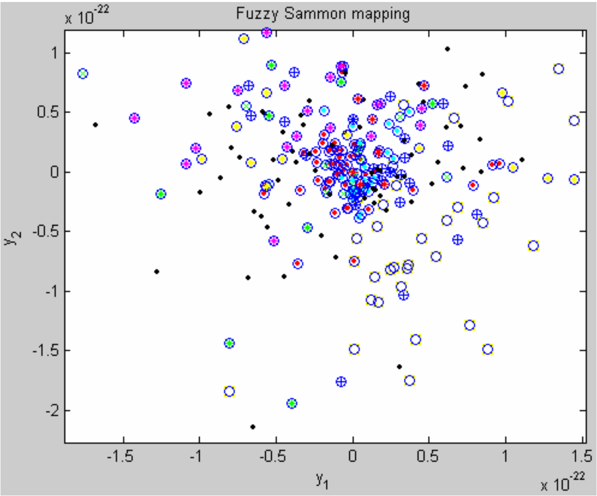


Fig.32. Result of Fuzzy Sammon projection by the price returns of HFR database over the results obtained with the Gustafson - Kessel clustering algorithm

	P	$\sum_{k=1}^{N} \overline{\mu_k^2}$	$\sum_{k=1}^{N} \overline{\mu_k^{2*}}$	E
PCA	0.1250	0.8500	0.5414	0.3192
Sammon	0.877	0.8500	0.3879	0.1713
Fuzzy Sammon	0.0041	0.8500	0.4014	0.1013

Table8. Relation-indexes on the price returns of HFR database (Gustafson - Kessel clustering algorithm)

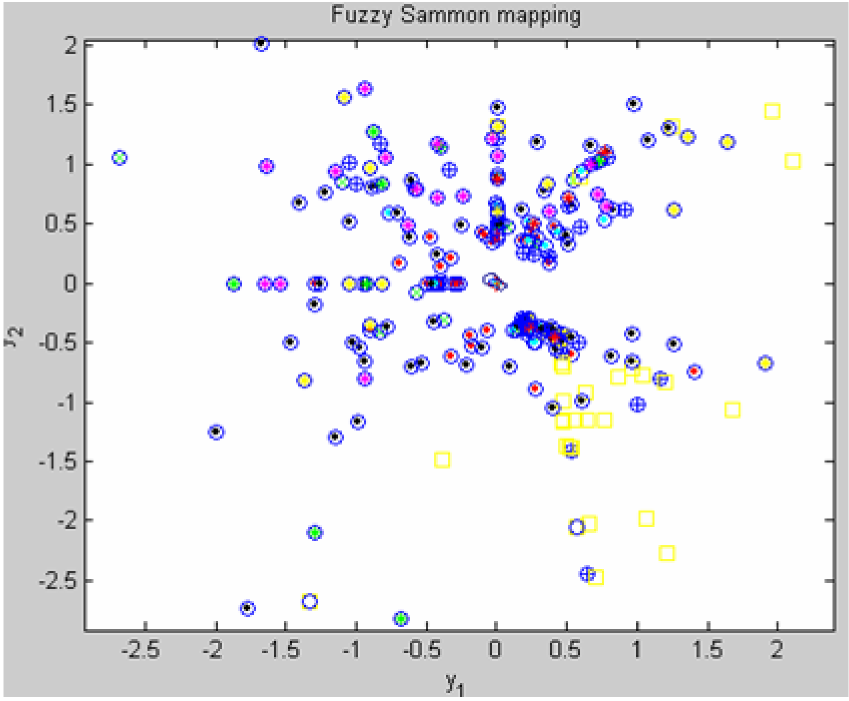
As Table8 shows, Fuzzy Sammon Mapping has better projection results by the value of P than Principal component Analysis and conventional Sammon mapping. The original Sammon's stress for all the three techniques is calculated, in order to be able to compare them; it can be observed that conventional Sammon mapping gives the best results.

As it can be noticed, the Principal component Analysis and conventional Sammon mapping visualization algorithms do not give good results. All depends on the underlying structure of the data. The Fuzzy Sammon Mapping instead gives better results.

- **TASS database**

TASS is the information and research subsidiary of Credit Suisse First Boston Tremont Advisers. It has nine categories of hedge funds, classified based on the investment styles of hedge fund managers. The TASS database of hedge funds used consists of monthly returns and accompanying information for 229 hedge funds. The index starts at 31st December 1992 and ends the 31st December 2003.

The price returns corresponding to each hedge fund are obtained by using the indexes given in the TASS database as following:



1. **K-means clustering algorithm**

Each point in the figures is coloured according to the style membership of the fund.

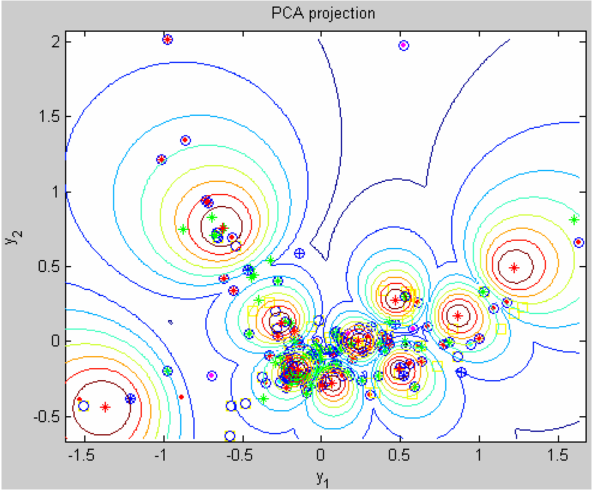


Fig.33. Result of PCA projection by the price returns of TASS database over the results obtained with the K- means clustering algorithm

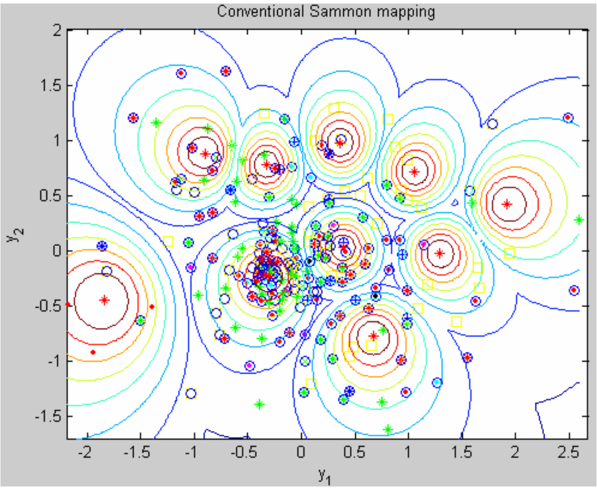


Fig.34. Result of conventional Sammon projection by the price returns of TASS database over the results obtained with the K- means clustering algorithm

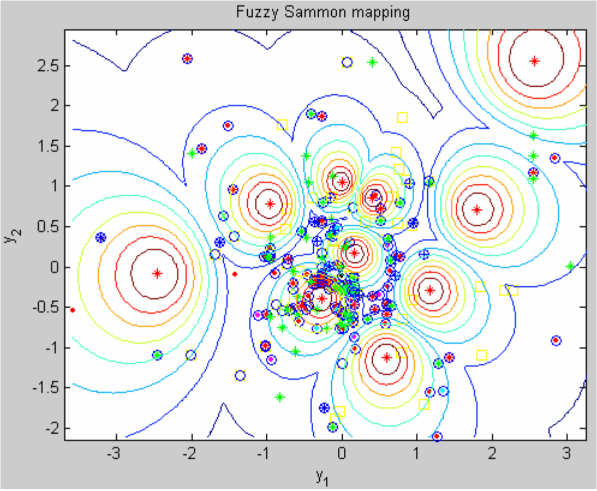


Fig.35. Result of Fuzzy Sammon projection by the price returns of TASS database over the results obtained with the K- means clustering algorithm

	P	$\sum_{k=1}^{N} \overline{\mu_k^2}$	$\sum_{k=1}^{N} \overline{\mu_k^{2*}}$	E
PCA	0.1983	1.0000	0.4892	0.2850
Sammon	0.1018	1.0000	0.4325	0.0628
Fuzzy Sammon	0.1246	1.0000	0.3309	0.2469

Table9. Relation-indexes on the price returns of TASS database (K- means clustering algorithm)

2. Fuzzy C-means clustering algorithm

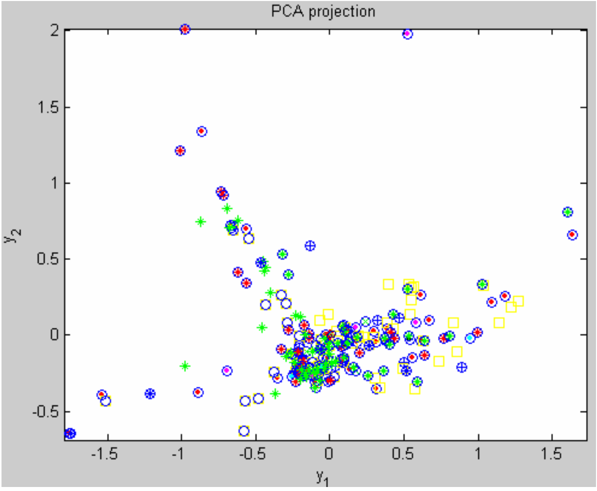


Fig.36. Result of PCA projection by the price returns of TASS database over the results obtained with the Fuzzy C - means clustering algorithm

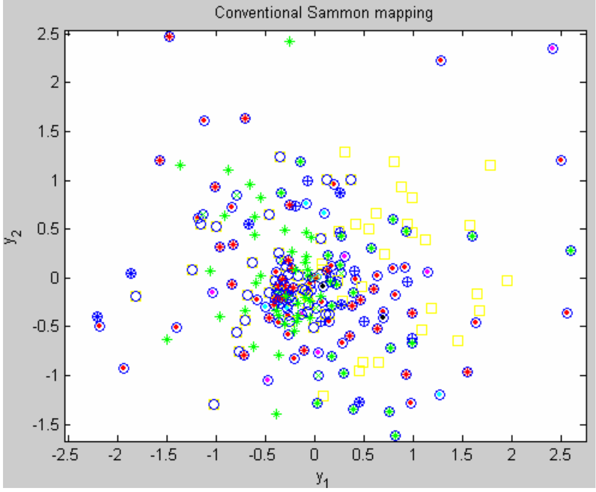


Fig.37. Result of Sammon projection by the price returns of TASS database over the results obtained with the Fuzzy C - means clustering algorithm

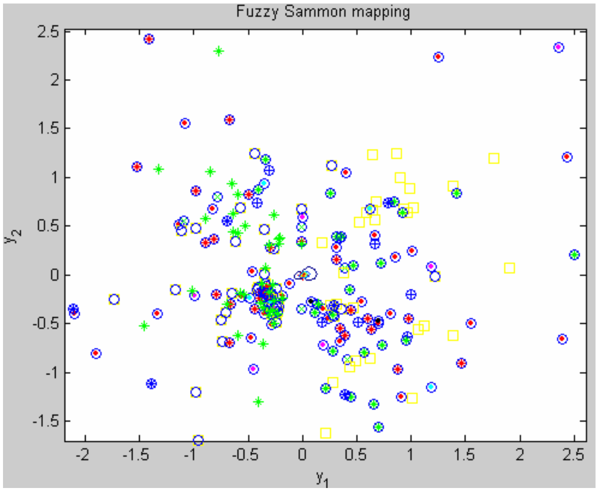


Fig.38. Result of Fuzzy Sammon projection by the price returns of TASS database over the results obtained with the Fuzzy C - means clustering algorithm

	P	$\sum_{k=1}^{N}\overline{\mu_k^2}$	$\sum_{k=1}^{N}\overline{\mu_k^{2*}}$	E
PCA	0.0002	0.6000	0.3000	0.2850
Sammon	0.0002	0.6000	0.3000	0.0628
Fuzzy Sammon	0.0001	0.6000	0.3000	0.0763

Table10. Relation-indexes on the price returns of TASS database (Fuzzy C- means clustering algorithm)

7.5.1.2. Optimal number of clusters

In the course of every partitioning problem the number of subsets (called the clusters) must be given by the user before the calculation, but it is rarely known *apriori*, in this case it must be searched also with using validity measures. The validity function provides cluster validity measures for each partition. The optimal partition can be determined by the point of the extreme of the validation indexes in dependence of the number of clusters. The indexes calculated are explained in detail in section 6.4.

- Partition Coefficient (PC),
- Classification Entropy (CE)
- Partition Index (SC)
- Separation Index (S)
- Xie and Beni's Index (XB)
- Dunn's Index (DI)
- Alternative Dunn Index (DII).

The number of clusters is determined so that the smaller *S* means a more compact and separate clustering. The resulting clusters are compared to each other on the basis of the validity function. Similar clusters are collected in one cluster; very bad clusters are eliminated, so the number of clusters is reduced. The goal should therefore be to minimize the value of *S*. See Figure 5, section 6.4.

7.5.1.2.1. Artificial generated data - results

As explained before, the validity measure indexes are calculated in order to help find the optimal number of clusters for the wanted data set. For more details see section 6.4 (Validation). No validation index is reliable only by itself; that is why all the programmed indexes are shown, and the optimum can be only detected with the comparison of all the results. The partitions with fewer clusters are better, when the differences between the values of a validation index are minor.

During the simulations described in section 7.5.1.1.1 (Artificial generated data), the following sets of validity measures were obtained.

Case1. This corresponds to Case1 from section 7.5.1.1.1.

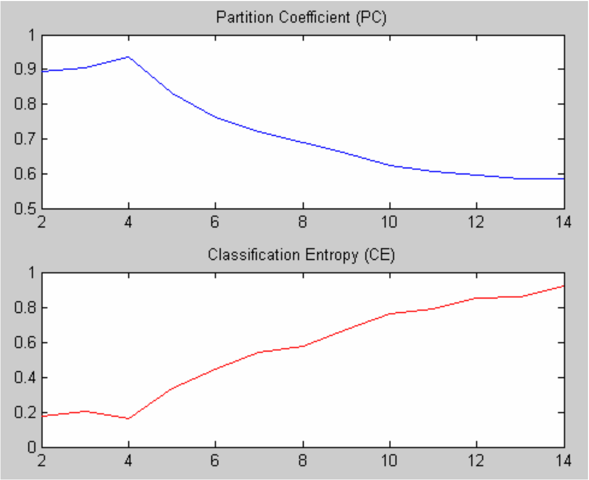


Fig.42. Values of Partition Coefficient and Classification Entropy (variance of alpha and beta = 0.05)

The main drawback of the partition coefficient (PC) is the monotonic decreasing with the number of clusters - *c* and the lack of direct connection to the data. The classification entropy (CE) has the same problems: monotonic increasing with the number of clusters - *c* and hardly detectable connection to the data structure. On the score of Fig. 42, the optimal number of clusters is 4. However, this result corresponds in totality with the expected one, knowing that we generated four distinct labels (see section 7.5.1.1.1 – Case 1 for more details).

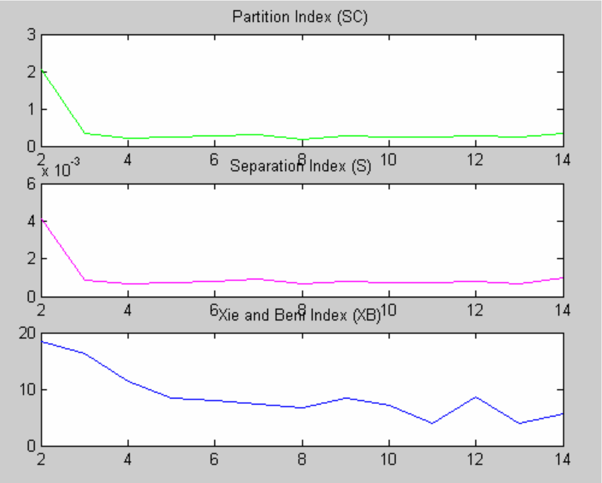


Fig.43.Values of Partition Index and Separation Index and Xie and Beni's Index (variance of alpha and beta = 0.05)

In Fig. 43 more informative diagrams are shown: partition index (SC) and separation index (S) decreases at the *c* = 3 point; this shows that the optimal number of clusters is 3. The Xie and Beni's Index (XB) index reaches this local minimum at $c \approx 5$, or more.

Considering that SC and S are more useful, when comparing different clustering methods with the same c, we chose the optimal number of clusters to 4, which is confirmed by the Dunn's index (DI) and the Alternative Dunn Index (ADI) too in Fig. 44.

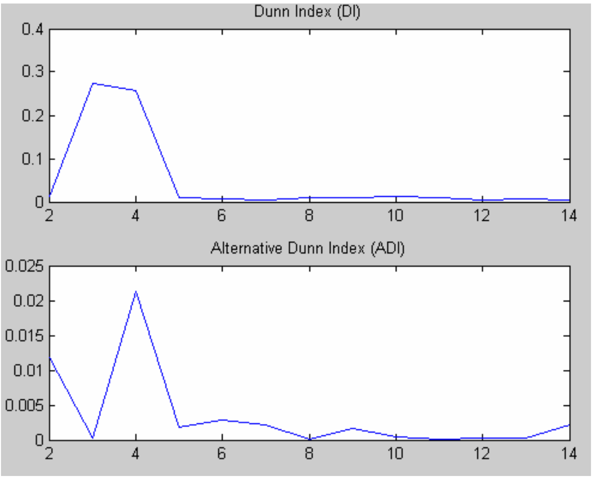


Fig.44. Values of Dunn's Index and Alternative Dunn Index (variance of alpha and beta = 0.05)

Case2. This corresponds to Case2 from section 7.5.1.1.1.

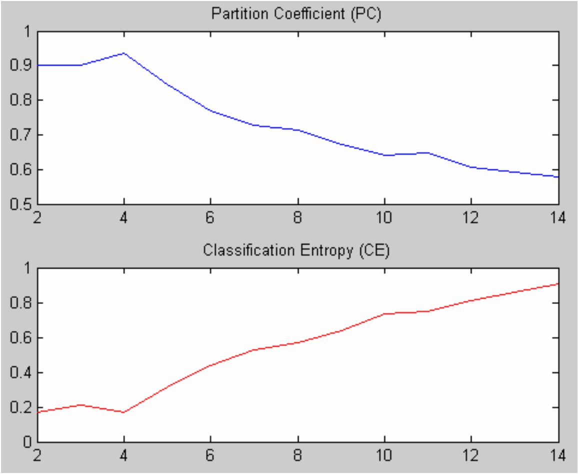


Fig.45. Values of Partition Coefficient and Classification Entropy (variance of alpha and beta = 0.5)

The values of the PC and CE from Fig.45 show the value 4 as optimal number of clusters for the artificially generated data; where the variance of alpha and beta is equal to 0.5. This was the value expected, so the algorithms work properly.

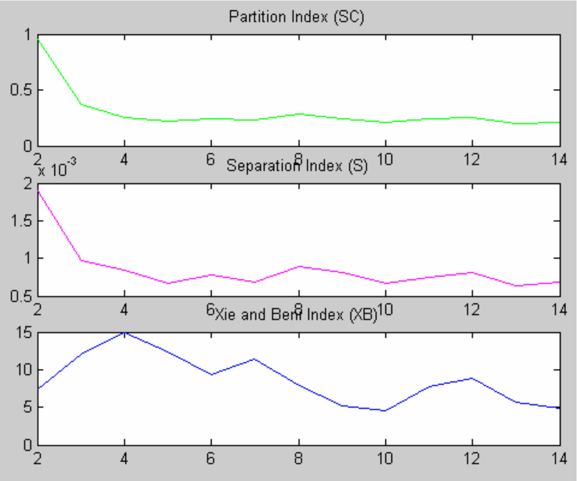


Fig.46.Values of Partition Index and Separation Index and Xie and Beni's Index (variance of alpha and beta = 0.5)

In Fig. 46 there are more informative diagrams: the partition index (SC) and the separation index (S) decreases at the *c* = 3 point; this shows that the optimal number of clusters is 3. The Xie and Beni's Index (XB) varies a lot and it cannot give a proper optimal number. In this case, considering that SC and S are more useful, when comparing different clustering methods with the same *c*, we chose the optimal number of clusters to 3, which is confirmed by the Dunn's index (DI). The Alternative Dunn Index (ADI) in Fig. 47 gives 4 the optimal number.

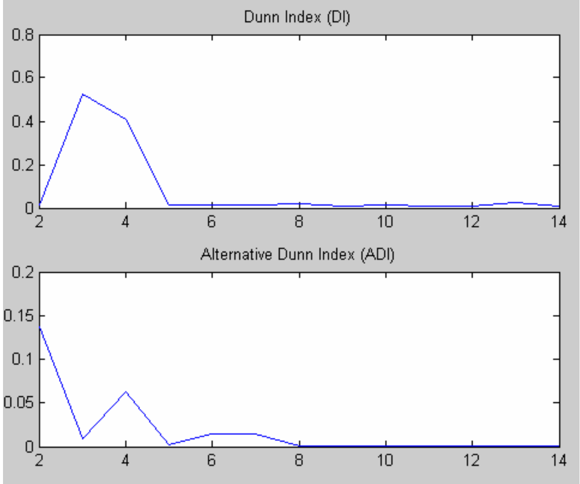


Fig.47. Values of Dunn's Index and Alternative Dunn Index (variance of alpha and beta = 0.5)

Case3. This corresponds to Case3 from section 7.5.1.1.1.

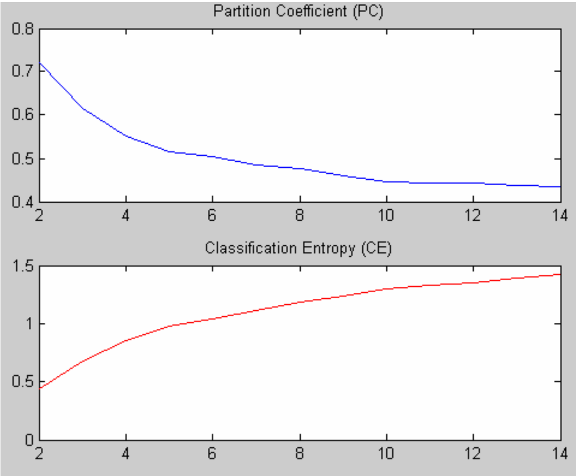


Fig.48. Values of Partition Coefficient and Classification Entropy (variance of alpha and beta = 1)

As seen in Fig.48, the main drawback of the partition coefficient (PC) is the monotonic decreasing with the number of clusters - *c* and the lack of direct connection to the data. The classification entropy (CE) has the same problems: monotonic increasing with the number of clusters - *c* and hardly detectable connection to the data structure. On the score of Fig. 49, we cannot give the exact optimal number of clusters, but we can choose 4 as the suitable one.

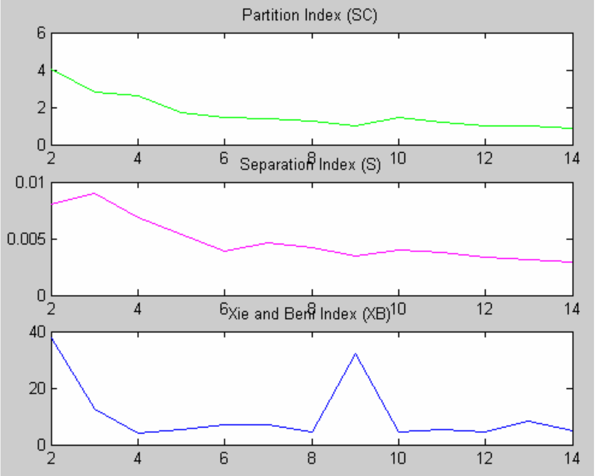


Fig.49.Values of Partition Index and Separation Index and Xie and Beni's Index (variance of alpha and beta = 1)

As shown in Fig.49, the partition index (SC) decreases continuously, and the separation index (S) decreases at the *c* = 3 point; this shows that the optimal number of clusters is 3. The Xie and Beni's Index (XB) varies a lot and it cannot give a proper optimal number; however, in c=4 it reaches the local minimum, so this value can be taken as the optimal one. When comparing different clustering methods, we chose the optimal number of clusters to be somewhere between 3 and 4, which is confirmed by the

Alternative Dunn Index (ADI). The Dunn's index (DI) in Fig. 50 cannot give a proper optimal number, so its value is not taken into account.

The more the variances of alpha and beta have higher values, the more difficult is to obtain a single optimal number of clusters.

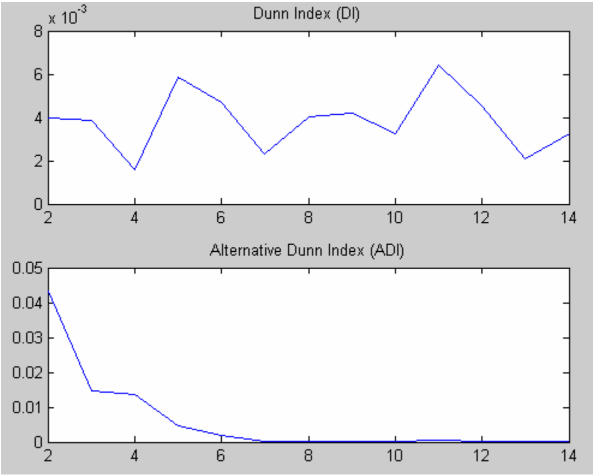


Fig.50. Values of Dunn's Index and Alternative Dunn Index (variance of alpha and beta = 1)

7.5.1.2.2. Real life data – results

- **HFR database**

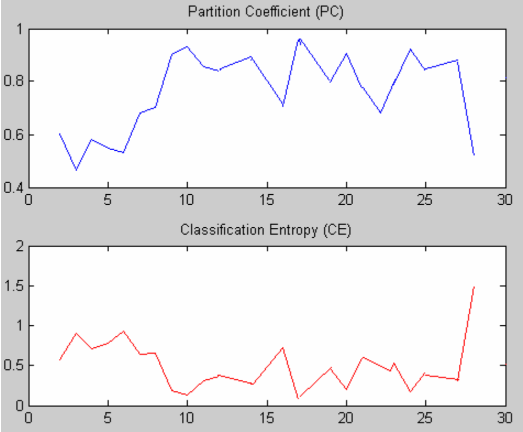


Fig.51. Values of Partition Coefficient and Classification Entropy for the price returns of HFR database

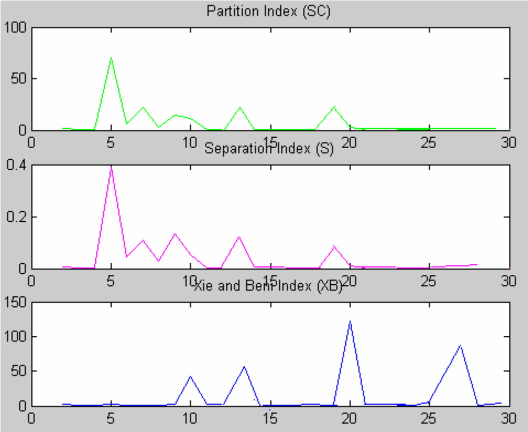


Fig.52.Values of Partition Index and Separation Index and Xie and Beni's Index for the price returns of HFR database

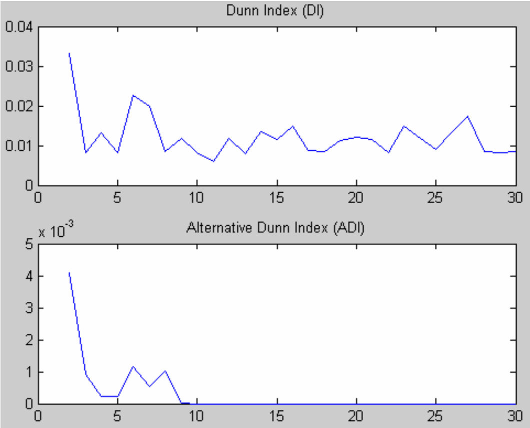


Fig.53. Values of Dunn’s Index and Alternative Dunn Index for the price returns of HFR database

- **TASS database**

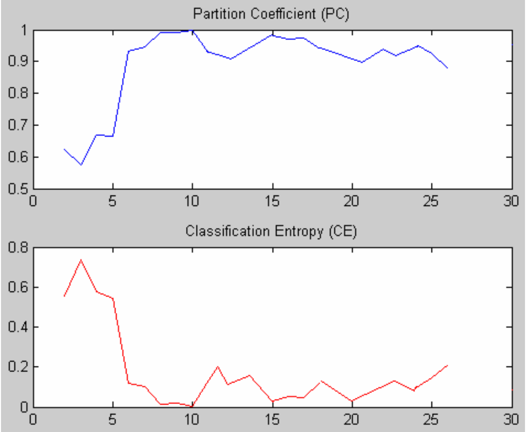


Fig.54. Values of Partition Coefficient and Classification Entropy for the price returns of TASS database

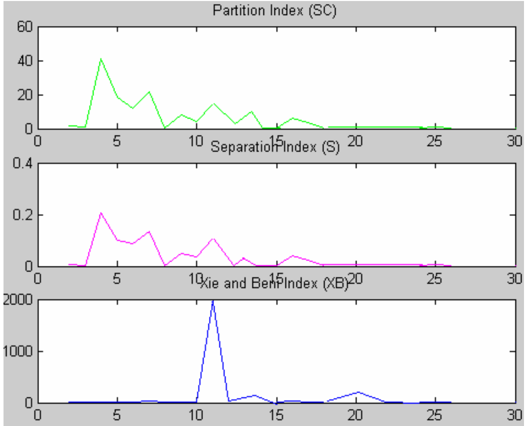


Fig.55.Values of Partition Index and Separation Index and Xie and Beni's Index for the price returns of TASS database

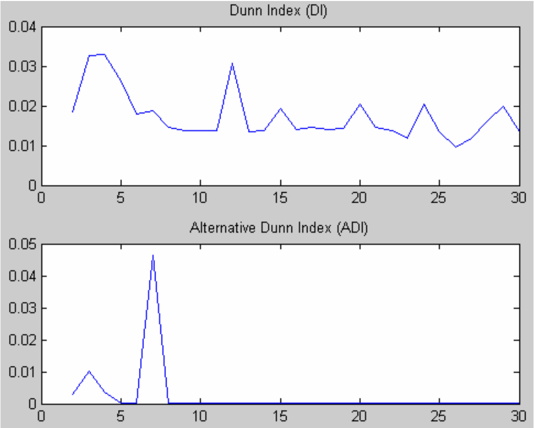


Fig.56. Values of Dunn’s Index and Alternative Dunn Index for the price returns of TASS database

7.5.2. Filtering

In this section the results of several simulations are presented, in order to show the importance and the accuracy of the proposed filtering methods; See section 3 (Gaussian Approximate Bayesian Estimation – Kalman Filter Framework) for the theory. Several filters are applied on some artificially simulated data, corrupted with some additive white noise. The performances of the following filters were evaluated:

- Kalman Filter (KF)
- Extended Kalman Filter (EKF)
- Unscented Kalman Filter (UKF)
- Central Difference Kalman Filter (CDKF)
- Square-Root Unscented Kalman Filter (SRUKF)
- Square-Root Central Difference Kalman Filter (SRCDKF)

A randomly coefficient time-series is generated and corrupted by additive white noise. The filtering results are presented in the next figures:

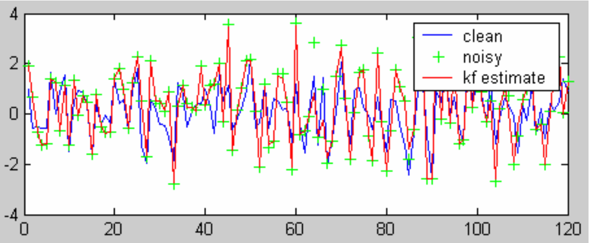


Fig.57. Kalman Filter estimation of the artificially generated noisy time-series

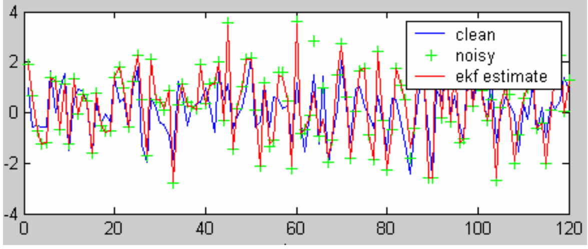


Fig.58. Extended Kalman Filter estimation of the artificially generated noisy time-series

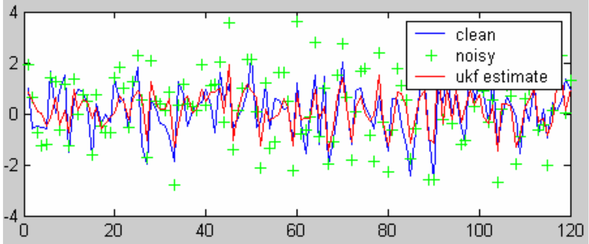


Fig.59. Unscented Kalman Filter estimation of the artificially generated noisy time-series

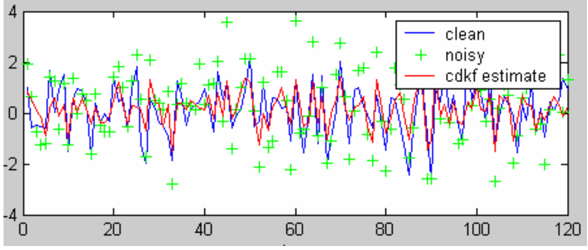


Fig.60.Central Difference Kalman Filter estimation of the artificially generated noisy time-series

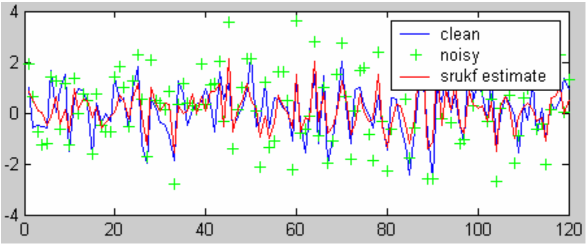


Fig.61. Square-Root Unscented Kalman Filter estimation of the artificially generated noisy time-series

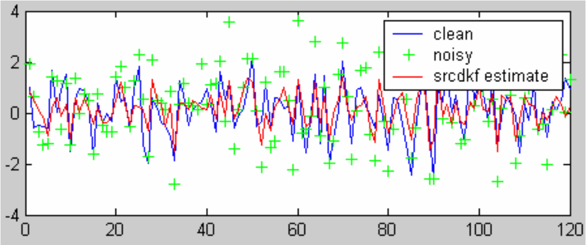


Fig.62. Square-Root Central Difference Kalman Filter estimation of the artificially generated noisy time-series

- kf : Mean square error (MSE) of estimate : 0.99635
- ekf : Mean square error (MSE) of estimate : 0.99635
- ukf : Mean square error (MSE) of estimate : 0.61812
- cdkf : Mean square error (MSE) of estimate : 0.55779
- srukf : Mean square error (MSE) of estimate : 0.57407
- srcdkf : Mean square error (MSE) of estimate : 0.55779

Chapter 8

Conclusions and future research

This paper tries to give some answers and to propose some solutions to the three questions articulated in the introduction.

Firstly, several types of filtering-analysis are proposed and tested on artificially generated data, in order to give an appropriate solution in estimating the hidden hedge funds time-series state in a way that minimizes the error. Here the extensions of the Kalman filter give much better results. The Kalman filter is a special case of a Bayesian filter, giving the optimal solution for linear state and observation equations and Gaussian noises. The Extended Kalman Filter (EKF) is a linearization of the Kalman filter for nonlinear state and/or observation equations which uses a first-order Taylor expansion. Still, it assumes Gaussian noises. In the case of linear equations, the EKF is equivalent to the standard Kalman filter. The Unscented Kalman Filter (UKF) is another approach to nonlinear systems with Gaussian noises. Unlike EKF, it does not linearize the equations and therefore needs no Jacobians, but it approximates the state variable by using an unscented transformation to it. Another category of filters are Particle filters and their extensions. Whereas the Kalman filter and its extensions make a Gaussian assumption to simplify the optimal recursive Bayesian estimation, particle filters make no assumptions on the form of the probability densities in question, that is a full nonlinear, non-Gaussian estimation; these filters, together with their extensions are very appropriate to "black box" systems. This filtering analysis could consist as a challenge in a future work.

Secondly, several clustering methods should were described and chosen, in order to better distinguish and classify the different hedge funds evolutions in time, based on the existent measurements. The K-means and the Fuzzy C-means clustering algorithms have been deeply studied, together with some better visualization methods: the Sammon and the Fuzzy Sammon mapping which give much better results than the classical method – the Principal Component Analysis.

Of course, the experiments came together with new research questions and the short time didn't allow us to give answers and find the solutions to all these problems. The main remaining questions are regarding the clustering efficiency after filtering (in the non-linear case), the visualization of the clusters (a financial analysis would be needed), the TASS and HFR hedge funds database interpretation (due to their differences in hedge funds style).

The answer to another question articulated in the introduction remains to be found in some future work and research – whether wavelet-analysis could decompose the hedge funds returns time-series into multiple levels, such that each level captures specific useful information? For the future research I would like to propose the study of a temporal cluster analysis framework consisting of three important stages: feature extraction from the hedge funds returns time series, dimension reduction of the high-dimensional feature sets and clustering of the already-processed feature sets.

Bibliography

[1] E. A. Wan, R. van der Merwe, and A. T. Nelson, \Dual estimation and the unscented transformation," in *Advances in Neural Information Processing Systems*. S. A. Solla, T. K. Leen, and K. R. Miller, Eds., Cambridge, MA: MIT Press, 2000

[2] Oregon Graduate Institute of Science and Technology, Rudolph van der Merwe, Rebel Toolkit, http://choosh.ece.ogi.edu/rebel/, August 2003.

[3] S. Haykin, *Kalman Filtering and Neural Networks*, New York, NY: John Wiley and Sons Inc., 2001.

[5] R. van der Merwe and E. Wan, "*Efficient Derivative-Free Kalman Filters for Online Learning*". In *Proc. of ESANN*, Bruges, April 2001.

[6] R. van der Merwe and E. Wan. The square-root unscented kalman filter for state and parameter-estimation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, May 2001.

[7] E. A.Wan and R. van der Merwe. *Kalman Filtering and Neural Networks, Ed. Simon Haykin.*, chapter 7 - The Unscented Kalman Filter. Wiley, 2001.

[8] Kalman, R. E. *A new approach to linear filtering and prediction problems. ASME Journal of Basic Engineering* (1960)

[9] Rudolph van der Merwe, Nando de Freitas, Arnaud Doucet, and Eric Wan. *The unscented particle filter.* Technical Report CUED/F-INFENG/TR 380, Cambridge University Engineering Department, August 2000.

[10] E. A.Wan and R. van der Merwe. *The Unscented Kalman Filter for Nonlinear Estimation*. In *Proc. of IEEE Symposium 2000 (AS-SPCC)*, Lake Louise, Alberta, Canada, October 2000.

[11] Capocci, D. (2002): "*An Analysis of Hedge Fund Performance 1984-2000,*" Working paper, University of Liège.

[12] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte-Carlo Methods in Practice*. Springer-Verlag, April 2001.

[13] D. Fox. KLD-Sampling: Adaptive Particle Filters. In *Advances in Neural Information Processing Systems 14*, 2001.

[14] D. Fox, S. Thrun, F. Dellaert, and W. Burgard. *Sequential Monte Carlo Methods in Practice.*, chapter Particle filters. Springer Verlag, 2000.

[15] J. F. G. de Freitas. *Bayesian Methods for Neural Networks*. PhD thesis, Cambridge University Engineering Department, 1999.

[16] R. van der Merwe, N. de Freitas, A. Doucet, and E. Wan. The Unscented Particle Filter. In *Advances in Neural Information Processing Systems 13*, Nov 2001.

[17] R. van der Merwe and E. Wan. Efficient Derivative-Free Kalman Filters for Online Learning. In *Proc. of ESANN*, Bruges, April 2001.

[18] R. van der Merwe and E. Wan. The square-root unscented kalman filter for state and parameter-estimation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, May 2001.

[19] R. van der Merwe and E. A. Wan. Gaussian Mixture Sigma-Point Particle Filters for Sequential Probabilistic Inference in Dynamic State-Space Models. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, April 2003. IEEE.

[20] R. van der Merwe. *Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models*. PhD thesis, OGI School of Science & Engineering at Oregon Health & Science University, 1999.

[21] Dunn, J. C., *A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters*, J. Cybernetics., 1974.

[22] Bezdek, J. C., *Cluster Validity with fuzzy sets*, J. Cybernetics., 1974.

[23] Bezdek, J. C., *A covariance theorem for the fuzzy ISODATA clustering algorithm*, IEEE Trans. Pattern Anal. Machine Intell., 1980.

[24] R. Babuska, P.J. van der Veen, and U. Kaymak. Improved covariance estimation for GustafsonKessel clustering. *In Proceedings of 2002 IEEE International Conference on Fuzzy Systems*, Honolulu, Hawaii, May 2002.

[25] Juha Vesanto, Neural Network Tool for Data Mining: SOM Tool- box,*Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000*, 2000.

[26] A. Kovacs - J. Abonyi, Vizualization of Fuzzy Clustering Results by Modified Sammon Mapping, *Proceedings of the 3rd International Symposium of Hungarian Researchers on Computational Intelligence*, 2002.

[27] Agarwal, V., and N. Y. Naik (2004): "*Risks and Portfolio Decisions involving Hedge Funds*," Review of Financial Studies.

[28] Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002): "*A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking*," IEEE Transactions on Signal Processing

[29] Brealey, R. A., and E. Kaplanis (2001): "*Changes in the Factor Exposure of Hedge Funds*," London Business School.

[30] Fung, W., and D. A. Hsieh (1997): "*Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds*," The Review of Financial Studies

[31] Géhin, W., and M. Vaissié (2005): "*The right place for alternative betas in hedge fund performance: an answer to the capacity effect fantasy*," Edhec Risk and Asset Management Research Center.

[32] Harvey, A. C. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

[33] Jaeger, L., and C. Wagner (2005): "*Factor Modelling and Benchmarking of Hedge Funds: Are passive investments in hedge funds possible*?," Working paper, Partners Group, Baar-Zug.

[34] Kat, H. M., and J. Miffre (2003): "*Performance Evaluation and Conditioning Information: The Case of Hedge Funds*," Working Paper # 0006, Cass Business School, City University London.

[35] Krail, R. J., C. S. Asness, and J. M. Liew (2001): "*Do Hedge Funds Hedge?*," The Journal of Portfolio Management.

[36] Mamaysky, H., M. Spiegel, and H. Zhang (2003): "*Estimating the Dynamics of Mutual Fund Alphas and Betas*," Yale ICF Working Paper.

[37] McGuire, P., E. Remolona, and K. Tsatsaronis (2005): "*Time-varying exposures and leverage in hedge funds*," BIS Quarterly Review.

[38] Sharpe, W. F. (1992): "*Asset Allocation: Management Style and Performance Measurement*," Journal of Portfolio Management.

[40] Welch, G., and G. Bishop (2001): "*An Introduction to the Kalman Filte*r," ACM.

[41] Wells, C. (1996): *The Kalman Filter in Finance*, Kluwer Academic Publishers.

Appendix

Appendix 1: REBEL – Recursive Bayesian Estimation Library – Toolkit

ReBEL [11]is a Matlab toolbox designed to facilitate the sequential estimation in general state space models. ReBEL is developed and maintained by Rudolph van der Merwe [22]. These scripts have been used in the estimation performed in this paper.

Appendix 2: TASS Fund Category Definitions

The following is a list of category descriptions, taken directly from TASS documentation, that define the criteria used by TASS in assigning funds in their database to one of 17 possible categories:

- **Equity Hedge** This directional strategy involves equity-oriented investing on both the long and short sides of the market. The objective is not to be market neutral. Managers have the ability to shift from value to growth, from small to medium to large capitalization stocks, and from a net long position to a net short position. Managers may use futures and options to hedge. The focus may be regional, such as long/short US or European equity, or sector specific, such as long and short technology or healthcare stocks. Long/short equity funds tend to build and hold portfolios that are substantially more concentrated than those of traditional stock funds. US equity Hedge, European equity Hedge, Asian equity Hedge and Global equity Hedge is the regional Focus.

- **Dedicated Short Seller** Short biased managers take short positions in mostly equities and derivatives. The short bias of a manager's portfolio must be constantly greater than zero to be classified in this category.

- **Fixed Income Directional** This directional strategy involves investing in Fixed Income markets only on a directional basis.

- **Convertible Arbitrage** This strategy is identified by hedge investing in the convertible securities of a company. A typical investment is to be long the convertible bond and short the common stock of the same company. Positions are designed to generate profits from the fixed income security as well as the short sale of stock, while protecting principal from market moves.

- **Event Driven** This strategy is defined as 'special situations' investing designed to capture price movement generated by a significant pending corporate event such as a merger, corporate restructuring, liquidation, bankruptcy or reorganization. There are three popular sub-categories in event-driven strategies:
  - risk (merger) arbitrage,
  - distressed/high yield securities,
  - regulation D.

- **Non Directional/Relative Value** This investment strategy is designed to exploit equity and/or fixed income market inefficiencies and usually involves being simultaneously long and short matched market portfolios of the same size within a country. Market neutral portfolios are designed to be either beta or currency neutral, or both.

- **Global Macro** Global macro managers carry long and short positions in any of the world's major capital or derivative markets. These positions reflect their views on overall market direction as influenced by major economic trends and or events. The portfolios of these funds can include stocks, bonds, currencies, and commodities in the form of cash or derivatives instruments. Most funds invest globally in both developed and emerging markets.

- **Global Opportunity** Global macro managers carry long and short positions in any of the world's major capital or derivative markets on an opportunistic basis. These positions reflect their views on overall market direction as influenced by major economic trends and or events. The portfolios of these funds can include stocks, bonds, currencies, and commodities in the form of cash or derivatives instruments. Most funds invest globally in both developed and emerging markets.

- **Natural Resources** This trading strategy has a focus for the natural resources around the world.

- **Leveraged Currency** This strategy invests in currency markets around the world.

- **Managed Futures** This strategy invests in listed financial and commodity futures markets and currency markets around the world. The managers are usually referred to as Commodity Trading Advisors, or CTAs. Trading disciplines are generally systematic or discretionary. *Systematic* traders tend to use price and market specific information (often technical) to make trading decisions, while *discretionary* managers use a judgmental approach.

- **Emerging Markets** This strategy involves equity or fixed income investing in emerging markets around the world.

- **Property** The main focus of the investments is property.

- **Fund of Funds** A ‘Multi Manager’ fund will employ the services of two or more trading advisors or Hedge Funds who will be allocated cash by the Trading Manager to trade on behalf of the fund.