Analyses of financial time series:
Moving Averages and Correlations

Gayatri TILAK
Final Year Internship Report

Promotion: 2011
Contents

Acknowledgements  2

1 Presentation  3

2 Introduction  4

3 Analyses using moving averages  5
   3.1 Introduction .................................................. 5
   3.2 Analytical treatment ........................................... 5
      3.2.1 Price and return series for a stock ..................... 5
      3.2.2 Realised returns and strategy .......................... 6
      3.2.3 Characteristic functions and expectation of the realized returns .......................... 7
   3.3 Numerical simulations and results ............................ 8
      3.3.1 No memory \( (S_t \text{ independent of } S_{t-1}) \) ................. 8
      3.3.2 Memory \( (S_t \text{ depends on } S_{t-1}) \) ..................... 9

4 Analyses of correlations  12
   4.1 Introduction .................................................. 12
   4.2 Cross-correlations and correlation matrix ....................... 13
   4.3 Minimum spanning tree (MST) approach ........................ 13
      4.3.1 Methodology .............................................. 13
      4.3.2 Results ................................................ 14
   4.4 Multidimensional scaling (MDS) approach ....................... 15
      4.4.1 Methodology .............................................. 15
      4.4.2 Results ................................................ 17

5 Brief outlook and future work  20
   5.1 Analyses using Moving averages ............................... 20
   5.2 Analyses using Correlations ................................... 20
Acknowledgements

A huge thank you to,

Anirban, for being my teacher and my friend, for teaching me how to think, how to write reports, how to look for information, besides everything related to the project...special mention of the Physics ↔ Philosophy classes

Frédéric, for being everybody’s dream boss besides giving me this one of a kind opportunity to work at the BNP Paribas Chaire de Finance Quantitative

Olaf, for the wonderful ride in his BMW besides his numerous invaluable project related suggestions

Aymen, for introducing me to Mathematica; the saviour of all those who do not like to calculate integrals by hand; besides the hours he put in to help me with the equations

Nicolas, for saving me from lunch at CROUS time and again besides the numerous career related conversations

Tamas, Philippe and Ravee for sharing their MATLAB code

Mehdi, Hui, Ran, and Saumya, for making lunch-time so much more interesting

Annie, Sylvie, and Fanny for making life easier and colourful!!!

Place: Châtenay-Malabry                                  Date: 15 July 2011
Chapter 1

Presentation

The Laboratoire de Mathématiques Appliquées aux Systèmes at the École Centrale Paris was created specially for research in Applied Mathematics and Computer Science. It is focussed on modelling, simulation, analysis and optimisation of complex systems in various fields: industrial design, life and health, financial markets, information and networks. The research is carried out in close collaboration with several French and international research teams, as well as through long-term industrial partnerships and multi-partner collaborative projects.

The BNP Paribas Chaire de Finance Quantitative was formally established at École Centrale Paris in October 2007. Its main motive is to carry out research dedicated to the empirical and theoretical study of market microstructure in order-driven markets. The research spans along three main axes: empirical market microstructure; mathematical modelling in continuous time and numerical simulation of order books and trading strategies.
Chapter 2

Introduction

Financial time series is a record of the sequence of prices of commodities or assets. Analysis of financial time series has been of much interest not only to the practitioners and financial engineers in banks and financial institutions, as an empirical discipline, but also to the scientists and academicians, as a theoretical discipline, for making inferences and predictions. The inherent uncertainty in the financial time series and its theory, and other properties and characteristics of a “complex adaptive system” has made the studies and analyses of financial time series specially interesting to economists and statisticians [1], and attracted even the physicists [2, 3].

Though different kinds of financial time series were being recorded and studied for decades, the scale changed about twenty years ago. The computerisation and automatisation of the stock exchanges (and financial markets) that took place all over the world in the mid 1980’s and early 1990’s has lead to the explosion of the amount of data recorded. Nowadays, all transactions on a financial market are recorded tick-by-tick, i.e. every event on a stock is recorded with a timestamp defined up to the millisecond, leading to huge amounts of data. For example, the empirical database which we have been using in our studies, the Reuters Datascope Tick History (RDTH) database, today records roughly 25 gigabytes of data per trading day. Prior to this tremendous increase in recording market activity, statistics would be computed mostly with daily data. Presently, scientists can compute intraday statistics in high-frequency. This allows one to check known properties at new time scales, but it also demands special care in the treatment and analyses of the high-frequency data.

In this project, we tried to choose and briefly review two main areas: (i) Technical analysis using moving average rules, which works best with low frequency data, and (ii) Analysis and visualisation of correlations both in low and high frequency data. The report is thus organized in two chapters dedicated to these, along with an outlook and directions of future work in the last chapter.
Chapter 3

Analyses using moving averages

3.1 Introduction

The term “Technical Analysis” is a general heading for a myriad of trading techniques. Technical analysis is considered by many to be the original form of investment analysis. These techniques for discovering hidden relations in stock returns can range from extremely simple to quite elaborate [4]. The efficacy of technical analysis is disputed by the “efficient-market hypothesis” which states that stock prices are essentially unpredictable. For the purpose of this project though, we assume that technical analysis works, and it is possible to predict returns. We treat the problem analytically by doing the basic mathematics [5] and then verifying the results with numerical simulations.

The moving-average oscillator is one of the simplest and most widely used trading rule. According to this rule, buy and sell signals are generated by two moving averages (MA) of the level of the index - a long-period average and a short-period average. The idea behind computing moving averages is to smooth out an otherwise volatile series. In its simplest form this strategy is implemented as buying (or selling) when the short-period moving average rises above (or falls below) the long-period moving average. When the short-period moving average penetrates the long-period moving average, a trend is said to be initiated [4]. We consider the simplest rule of the MA() family, the MA(2), where the long-period moving average is of length two and the short-period moving average is of length unity.

3.2 Analytical treatment

3.2.1 Price and return series for a stock

We consider over $T$ periods, the price of an asset at any time period $t$ denoted by $P_t$. The “returns” over a time period with lag unity, are defined as:

$$X_t = P_t - P_{t-1}. \quad (3.1)$$

We thus have a set of $T-1$ returns to work with.

We then choose the simple “switching Markov model” to generate the returns [5], which can be mathematically stated as:

$$X_t = \alpha_0 (1 - S_t) + \alpha_1 S_t + (\sigma_0 (1 - S_t) + \sigma_1 S_t)\epsilon_t \quad (3.2)$$

where $S_t$ is a two state Markov variable such that $S_t \in \{0, 1\}$ , the parameter $\pi = P[S_t = 1]$ and $\epsilon_t$ is a random variable drawn from the standard Normal distribution ($\epsilon_t \sim N(0, 1)$). We can thus write the distribution of the returns $X_t$ as:

$$pdf(X) \sim (1 - \pi)N(\alpha_0, \sigma_0^2) + \pi N(\alpha_1, \sigma_1^2), \quad (3.3)$$

which can be interpreted as a mixture of the two normals with probability $\pi$ and $(1-\pi)$. 

5
3.2.2 Realised returns and strategy

We define the realised returns \( RR \) as:

\[
RR = \sum_{i=1}^{N} X_i,
\]

where \( N \) could be any natural number (a random variable), and depend on the trading rule or strategy we employ.

One of the simplest strategies could be, for example:

(i) buy at \( t = 0 \)

(ii) hold till price rises (while returns are positive)

(iii) sell when price falls for the very first time (the return is negative at \( t = N \)).

It is possible to construct more elaborate trading rules and strategies, but the above strategy fits well with the MA(2) rule, described earlier.
3.2.3 Characteristic functions and expectation of the realized returns

We next define the characteristic function: \( \phi_{X} (s) = E[e^{iXs}] \), where \( i = \sqrt{-1} \).

The characteristic function of any random variable completely defines its probability distribution. Thus it provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions:

\[
\varphi_X : \mathbb{R} \to \mathbb{C}; \quad \varphi_X(s) = \mathbb{E}[e^{isX}] = \int_{-\infty}^{\infty} e^{isx} dF_X(x) = \int_{-\infty}^{\infty} e^{isx} f_X(x) \, dx \tag{3.5}
\]

Characteristic functions can also be used to find moments of a random variable. Provided that the \( n \)-th moment exists for the series \( X_t \), the characteristic function can be differentiated \( n \) times to evaluate \( \mathbb{E}[X^n] = i^{-n} \phi^{(n)}_{X}(0) \).

Given the return time series \( X_t \), we introduce two parameters:

\[
\pi_1 = \mathbb{P}[X_1 > 0] \quad p_{11} = \mathbb{P}[X_t > 0 | X_{t-1} > 0] \tag{3.6}
\]

For the trading rule and strategy we have chosen earlier,

\[
\therefore \mathbb{P}[N = n] = \mathbb{P}[X_2 > 0, X_3 > 0, \ldots, X_{n-1} > 0, X_n < 0 | X_1 > 0] \cdot \mathbb{P}[X_1 > 0]
\]

\[
= p_{11}^{n-2}(1 - p_{11})\pi_1, \quad \text{if} \ n \geq 2 \tag{3.7}
\]

\[
= 1 - \pi_1, \quad \text{if} \ n = 2. \tag{3.8}
\]

We then have

\[
\phi(s) = (\phi(s) \mid (N = n))\mathbb{P}[N = n], \tag{3.9}
\]

We also define,

\[
\phi_+(s) = \mathbb{E}[e^{isX_t} \mid X_t > 0] \quad \phi_-(s) = \mathbb{E}[e^{isX_t} \mid X_t < 0] \]

\[
\phi_{++}(s) = \mathbb{E}[e^{isX_t} \mid X_t > 0, X_{t-1} > 0] \quad \phi_{+-}(s) = \mathbb{E}[e^{isX_t} \mid X_t < 0, X_{t-1} > 0], \tag{3.10}
\]

and using Eq. (3.4), we write

\[
\phi_R(s) = \mathbb{E}[e^{is\sum_{t=1}^{N} X_t}].
\]

Thus, we have

\[
\phi_R(s) = \sum_{1}^{N} \mathbb{E}[e^{is\sum_{t=1}^{N} X_t} \mid N = n] \cdot \mathbb{P}[N = n] \tag{3.11}
\]

\[
\sum_{1}^{N} \mathbb{E}[e^{is\sum_{t=1}^{N} X_t} \mid N = n] = \phi_{-}(s) + \phi_{+}(s)\phi_{++}^{n-2}(s)\phi_{+-}(s), \tag{3.12}
\]

or

\[
\phi_R(s) = \phi_{-}(s)(1 - \pi_1) + \sum_{1}^{N} \phi_{+}(s)\pi_1\phi_{++}^{n-2}(s)p_{11}^{n-2}\phi_{+-}(s)(1 - p_{11})
\]

\[
= \phi_{-}(s)(1 - \pi_1) + \phi_{+}(s)\pi_1\phi_{+-}(s)(1 - p_{11}) \sum_{1}^{N} \phi_{++}^{n-2}(s)p_{11}^{n-2}. \tag{3.13}
\]
We now use the identity, \( \sum_1^\infty \frac{z^k}{1-z} \), to derive
\[
\phi_R(s) = \phi_-(s)(1 - \pi_1) + \frac{\phi_+(s)\phi_-(s)\pi_1(1 - p_{11})}{1 - p_{11}\phi_++(s)}.
\] (3.14)

Once we have the above characteristic function, we derive the first moment \( \mathbb{E}(RR) = i^{-1}\phi'_R(0) \). We take the derivative and proceed as follows:
\[
\frac{\partial \phi_R(s)}{\partial s} = (1 - \pi_1)\frac{\partial \phi_-(s)}{\partial s} + \frac{\pi_1(1 - p_{11})\left\{\left(1 - p_{11}\phi_++(s)\right)\frac{\partial \phi_+(s)}{\partial s} + \frac{\partial \phi_+(s)}{\partial s} + p_{11}\frac{\partial \phi_+(s)}{\partial s}\right\}\phi_+(s)\phi_-(s)}{(1 - p_{11}\phi_++(s))^2}.
\] (3.15)

We use the notations,
\[
\phi'_+(0) = i\mathbb{E}[e^{iX_t} \mid X_t > 0] = i\mu_+,
\phi'_-(0) = i\mathbb{E}[e^{iX_t} \mid X_t < 0] = i\mu_-,
\phi_++(0) = i\mathbb{E}[e^{iX_t} \mid X_t > 0, X_{t-1} > 0] = i\mu_++,
\phi_+-(0) = i\mathbb{E}[e^{iX_t} \mid X_t < 0, X_{t-1} > 0] = i\mu_-+.
\] (3.16)

Also, we note that \( \phi_+(0) = \phi_-(0) = \phi_++(0) = \phi_+-(0) = 1. \)

We therefore have
\[
\frac{\partial \phi_R(s)}{\partial s} = (1 - \pi_1)i\mu_- + \frac{\pi_1(1 - p_{11})\left\{(1 - p_{11})(i\mu_+-i\mu_+)+p_{11}i\mu_++\right\}}{(1 - p_{11})^2},
\] (3.17)
or
\[
\mathbb{E}(RR) = (1 - \pi_1)\mu_- + \pi_1(\mu_+-\mu_+) + \frac{p_{11}\pi_1\mu_++}{1 - p_{11}}.
\] (3.18)

### 3.3 Numerical simulations and results

We first carry out Monte Carlo simulations to verify the analytically obtained results in the earlier part. We basically generate several return time series with the same initial conditions and parameters \( \alpha_0, \sigma_0 \) and \( \alpha_1, \sigma_1 \), and evaluate the expected realized returns (equivalent to the expected profit and loss), \( \mathbb{E}(RR) \). We find that the numerical simulations agree well with the analytical results. For our numerical simulations we have always considered averages over 10,000 independent realizations.

The next logical step was to study the effect of the various parameters on the expected realized returns. For this we carried out simulations by varying a single parameter each time.

#### 3.3.1 No memory (\( S_t \) independent of \( S_{t-1} \))

The parameters which control the drift i.e. \( \alpha_0 \) and \( \alpha_1 \) mainly affect the \( \mathbb{E}(RR) \). The values of \( \sigma_0 \) and \( \sigma_1 \) do not seem to affect \( \mathbb{E}(RR) \). When the values of \( \alpha_0 \) and \( \alpha_1 \) are such that the returns are symmetric around 0, then \( \mathbb{E}(RR) = 0. \)
Figure 3.2: (a) Plot of the time series of the two state markov variable $S_t$ against time $t$, for 100 periods, when it does not have memory, and with parameter $\pi = 0.5$. (b) Histogram of $X_t$ with parameter $\pi = 0.5$, $\alpha_0 = 3.0, \alpha_1 = 1.0, \sigma_0 = 2.1, \sigma_1 = 1.0$. (c) The variation of the expected realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$ and $\pi = 0.5$. (d) Heat map: The variation of the Sharpe ratio of the realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$ and $\pi = 0.5$.

### 3.3.2 Memory ($S_t$ depends on $S_{t-1}$)

In order to generate a process where $S_t$ depends on $S_{t-1}$ we introduce two new parameters, $p_{00} = \mathbb{P}[S_t = 0 \mid S_{t-1} = 0]$ and $q_{11} = \mathbb{P}[S_t = 1 \mid S_{t-1} = 1]$. The figures produced are for different values of these two transition probabilities. These probabilities control how either of the normals is chosen. The effect is pronounced only at extreme values i.e. either close to 0 or 1.

As before, $\alpha_0$ and $\alpha_1$ being the drifts of the 2 normals, they have the maximum effect on $\mathbb{E}(RR)$.

Both the $\sigma$’s do not seem to have much effect except at extreme values of $p_{00}$ and $q_{11}$.
Figure 3.3: (a) Plot of the time series of the two state markov variable $S_t$ against time $t$, for 100 periods, when it has memory and the parameters $p_{00} = 0.5$ and $q_{11} = 0.5$. (b) Histogram of $X_t$ with parameter $\alpha_0 = 3.0, \alpha_1 = 1.0, \sigma_0 = 2.1, \sigma_1 = 1.0$. (c) The variation of the expected realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$. (d) Heat map: The variation of the Sharpe ratio of the realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$.

Figure 3.4: (a) Plot of the time series of the two state markov variable $S_t$ against time $t$, for 100 periods, when it has memory and the parameters $p_{00} = 0.1$ and $q_{11} = 0.9$. (b) Histogram of $X_t$ with parameter $\alpha_0 = 3.0, \alpha_1 = 1.0, \sigma_0 = 2.1, \sigma_1 = 1.0$. (c) The variation of the expected realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$. (d) Heat map: The variation of the Sharpe ratio of the realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$. 

10
Figure 3.5: (a) Plot of the time series of the two state markov variable $S_t$ against time $t$, for 100 periods, when it has memory and the parameters $p_{00} = 0.9$ and $q_{11} = 0.1$. (b) Histogram of $X_t$ with parameter $\alpha_0 = 3.0, \alpha_1 = 1.0, \sigma_0 = 2.1, \sigma_1 = 1.0$. (c) The variation of the expected realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$. (d) Heat map: The variation of the Sharpe ratio of the realized returns (PnL) against $(\alpha_0, \sigma_0)$, keeping $\alpha_1 = 1.0, \sigma_1 = 1.0$. 

Chapter 4

Analyses of correlations

4.1 Introduction

The English word “correlation” is defined as “a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone”\(^1\). When we discuss about correlations in stock prices, we are really interested in the relations between variables such as stock prices, transaction volumes, etc. and more importantly how these relations affect the nature of the statistical distributions which govern the price time series\(^6\).

As mentioned earlier, the computerization of financial exchanges has lead to the availability of huge amount of tick-by-tick data, and computing correlation using these intraday data raises lots of issues concerning usual estimators. Let us assume that we observe \(d\) time series of prices or log-prices \(p_i, i = 1, \ldots, d\), observed at times \(t_m, m = 0, \ldots, M\). The usual estimator of the covariance of prices \(i\) and \(j\) is the realized covariance estimator, which is computed as:

\[
\hat{\Sigma}_{ij}(t) = \sum_{m=1}^{M} (p_i(t_m) - p_i(t_{m-1}))(p_j(t_m) - p_j(t_{m-1})).
\]

The problem is that high-frequency tick-by-tick data record changes of prices when they happen, i.e. at random times. Tick-by-tick data is thus asynchronous, contrary to daily close prices for example, that are recorded at the same time for all the assets on a given exchange. Using standard estimators without caution, could be one cause for the “Epps effect”, first observed in [7], which stated that “[c]orrelations among price changes in common stocks of companies in one industry are found to decrease with the length of the interval for which the price changes are measured.” Here we do not discuss the various estimators and corrections used; we rather use data at frequencies where the Epps effect is not pronounced.

We then try to review methods that will help us to analyse and visualise the dynamic evolution of the cross-correlations in stock prices over time. This is because financial markets are often characterized as “evolving complex systems”\(^8\), and the evolution may be a reflection of the changing power structure in the market or the manifestations of the passing of different products and product generations, new technologies, management teams, alliances and partnerships, etc. We aim to understand intuitively measures which can be used to characterize the market taxonomy and its state, and the consequences of the market events.

4.2 Cross-correlations and correlation matrix

We consider $N$ assets. In order to study the correlations between stocks we first denote the closure price of stock $i$ at time $\tau$ by $P_i(\tau)$ (Note that $\tau$ refers to a date, not a time window). We focus our attention to the logarithmic return of stock $i$, given by $r_i(\tau) = \ln P_i(\tau) - \ln P_i(\tau - 1)$ which, for a sequence of consecutive trading days, i.e. those encompassing the given window $t$, form the return vector $r_i^t$. In order to characterize the synchronous time evolution of assets, we use the equal time correlation coefficients between assets $i$ and $j$ defined as

$$\rho_{ij}^t = \frac{\langle r_i^t r_j^t \rangle - \langle r_i^t \rangle \langle r_j^t \rangle}{\sqrt{\langle (r_i^t)^2 \rangle - \langle r_i^t \rangle^2} \sqrt{\langle (r_j^t)^2 \rangle - \langle r_j^t \rangle^2}}$$

(4.1)

where (...) indicates a time average over the consecutive trading days included in the return vectors. These correlation coefficients obviously fulfill the condition $-1 \leq \rho_{ij} \leq 1$ and form an $N \times N$ correlation matrix $C^t$, which serves as the basis of further analyses.

4.3 Minimum spanning tree (MST) approach

4.3.1 Methodology

This concept was introduced by Mantegna as a method for finding a hierarchical arrangement of stocks through study of the clustering of companies by using correlations of asset returns [9]. With an appropriate metric, based on the correlation matrix, a fully connected graph is defined in which the nodes are companies, or stocks, and the “distances” between them are obtained from the corresponding correlation coefficients. The minimum spanning tree (MST) is generated from the graph by selecting the most important correlations and it is used to identify clusters of companies. Here, we briefly review the method of the minimum spanning tree [10, 6]. The minimum spanning tree, as a strongly pruned representative of asset correlations, is found to be robust and descriptive of stock market events.

For analysis and smoothing purposes, the data is divided time-wise into $M$ windows $t = 1, 2, ..., M$ of width $T$ corresponding to the number of daily returns included in the window. Several consecutive windows overlap with each other, the extent of which is dictated by the window step length parameter $\delta T$, describing the displacement of the window, measured also in trading days. The choice of window width is a trade-off between too noisy and too smoothed data for small and large window widths, respectively. For some results reproduced here from the paper [10], $\delta T = 20.8$ days and $T = 1000$ days; the overall number of windows was $M = 195$ for $\tau = 1, 2, ..., 5056$ price quotes per stock, and there were $N = 477$ from NYSE.

A non-linear transformation

$$d_{ij} = \sqrt{2(1 - \rho_{ij})}$$

is then used to obtain “distances” with the property $2 \geq d_{ij} \geq 0$, forming an $N \times N$ distance matrix $D^t$, such that all distances are “ultrametric”. This hypothesis is motivated a posteriori by the finding that the associated taxonomy is meaningful from an economic point of view. The concept of ultrametricity is discussed in detail by Mantegna [9], while the economic meaningfulness of the emerging taxonomy is addressed later. Out of the several possible ultrametric spaces, the subdominant ultrametric is opted for due to its simplicity and remarkable properties. In practice, it is obtained by using the distance matrix $D^t$ to determine the minimum spanning tree (MST) of the distances, according to the methodology of [9], denoted $T^t$. This is a simply connected graph that connects all $N$ nodes of the graph with $N - 1$ edges such that the sum of all edge weights, $\sum_{d_{ij}^t \in T^t} d_{ij}^t$, is minimum. (Here time (window) dependence of the tree is emphasized by the addition of the superscript $t$ to the notation.) MSTs constructed for different time windows are not independent from each other, but form a series through time. Consequently, this multitude of trees may be interpreted as a dynamic asset tree.

Below are the plots of the distribution of (i) distance elements $d_{ij}^t$ contained in the distance matrix $D^t$ (Figure 4.1), and (ii) distance elements $d_{ij}$ contained in the asset (minimum spanning) tree $T^t$. In the plots,
there appears to be a discontinuity in the distribution between roughly 1986 and 1990. It seems that part has been cut out, pushed to the left and made flatter. This anomaly is a manifestation of Black Monday (October 19, 1987), and its length along the time axis is related to the choice of window width $T$.

This illustrates the fact that in constructing the minimum spanning tree, we are effectively reducing the information space from $N(N - 1)/2$ separate correlation coefficients to $N - 1$ tree edges, in other words, compressing the amount of information drastically. This follows because the correlation matrix $C^t$ and distance matrix $D^t$ are both $N \times N$ dimensional, but due to their symmetry, both have $N(N - 1)/2$ distinct upper (or lower) triangle elements, while the spanning tree has only $N - 1$ edges. This, of course, raises the key question of information theory, whether essential information is lost in the reduction. One can actually show that the minimum spanning tree is a strongly reduced representative of the whole correlation matrix, which bears the essential information about the asset correlations [10].

Once we have constructed an MST, we define what is called the “central vertex” which is considered to be the “parent” of all other nodes in the tree. It is used as the reference point in the tree, against which the locations of all other nodes are relative. Thus all other nodes in the tree are “children” of the central vertex. Although there is arbitrariness in the choice of the central vertex, any change in its price strongly affects the course of events in the market as a whole. One choice is the node with the highest vertex degree, i.e. the number of edges (neighbours) of the vertex.

### 4.3.2 Results

Mantegna’s idea of linking stocks in an ultrametric space, was motivated *a posteriori* by the property of such a space to provide a meaningful economic taxonomy. In order to visualize the grouping of stocks, we show a sample tree in Figure 4.2. This was obtained by studying a dataset of 116 S&P 500 stocks, extending from the beginning of 1982 to the end of 2000, resulting in a total of 4787 price quotes per stock [11]. The window width was set at $T = 1000$, and the shown sample tree is located time-wise at $t = t^*$, corresponding to 1.1.1998. The stocks in this dataset fall into 12 *sectors*, which are Basic Materials, Capital Goods, Conglomerates, Consumer/Cyclical, Consumer/Non-Cyclical, Energy, Financial, Healthcare, Services, Technology, Transportation and Utilities. The sectors are indicated in the tree with different markers, while the industry classifications are omitted for reasons of clarity.

We use the term *sector* exclusively to refer to the given third party classification system of stocks. The term *branch* refers to a subset of the tree, to all the nodes that share the specified common parent. In addition to the parent, we need to have a reference point to indicate the generational direction (i.e. who is who’s parent) in order for a branch to be well defined. In our case, the reference is the central node. There are some branches in the tree, in which most of the stocks belong to just one sector, indicating that the
branch is fairly homogeneous with respect to business sectors, see Figure 4.2.

There are, however, some observed deviations to the classification, which could possibly be due to (amongst others): (i) Uncertainty in asset prices in the minds of investors causes some seemingly random price fluctuations to take place, and this introduces “noise” in the correlation matrix. Therefore, it is not reasonable to expect a one-to-one mapping between business sectors and MST clusters. (ii) Business sector definitions are not unique, but vary by the organization issuing them. In this work, we used the classification system by Forbes [12], where the studied companies are divided into 12 business sectors and 51 industries. Forbes has its own classification principle, based on company dynamics rather than size alone. (iii) In many classification systems, companies engaged in substantially different business activities are classified according to where the majority of revenues and profits comes from. For highly diversified companies, these classifications are more ambiguous and, therefore, less informative.

In conclusion, the success of the applied method in identifying market taxonomy is pretty remarkable.

4.4 Multidimensional scaling (MDS) approach

4.4.1 Methodology

Multidimensional scaling

Multidimensional scaling (MDS) is a set of data analysis techniques that display the structure of distance-like data as a geometrical picture.

Each object or event is represented by a point in a multidimensional space. The points are arranged in this space so that the distances between pairs of points have the strongest possible relation to the similarities among the pairs of objects. That is, two similar objects are represented by two points that are close together, and two dissimilar objects are represented by two points that are far apart. The space is usually a two- or three-dimensional Euclidean space, but may be non-Euclidean and may have more dimensions.

MDS is a generic term that includes many different specific types. These types can be classified according to whether the similarities data are qualitative (called nonmetric MDS) or quantitative (metric MDS). The
number of similarity matrices and the nature of the MDS model can also classify MDS types. This classification yields classical MDS (one matrix, unweighted model), replicated MDS (several matrices, unweighted model), and weighted MDS (several matrices, weighted model).

The identifying aspect of classical MDS (CMDS) is that there is only one similarity matrix. Fig. 4.3 is a matrix of similarity data suitable for CMDS. It contains the flying mileages between 10 American cities. The cities are the “objects” and the mileages are the “similarities”.

<table>
<thead>
<tr>
<th>ATLANTA</th>
<th>CHICAGO</th>
<th>DENVER</th>
<th>HOUSTON</th>
<th>LOS ANGELES</th>
<th>MIAMI</th>
<th>NEW YORK</th>
<th>SAN FRANCISCO</th>
<th>SEATTLE</th>
<th>WASHINGTON DC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta</td>
<td>0</td>
<td>985</td>
<td>1,018</td>
<td>1,836</td>
<td>614</td>
<td>2,123</td>
<td>2,502</td>
<td>951</td>
<td>510</td>
</tr>
<tr>
<td>Chicago</td>
<td>985</td>
<td>0</td>
<td>920</td>
<td>1,265</td>
<td>1,188</td>
<td>1,711</td>
<td>1,820</td>
<td>1,737</td>
<td>510</td>
</tr>
<tr>
<td>Denver</td>
<td>1,018</td>
<td>920</td>
<td>0</td>
<td>894</td>
<td>831</td>
<td>1,726</td>
<td>1,820</td>
<td>1,836</td>
<td>1,188</td>
</tr>
<tr>
<td>Houston</td>
<td>1,836</td>
<td>1,265</td>
<td>894</td>
<td>0</td>
<td>1,344</td>
<td>968</td>
<td>1,420</td>
<td>1,615</td>
<td>1,920</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>614</td>
<td>1,188</td>
<td>1,711</td>
<td>2,123</td>
<td>0</td>
<td>2,239</td>
<td>2,611</td>
<td>3,37</td>
<td>951</td>
</tr>
<tr>
<td>Miami</td>
<td>2,123</td>
<td>1,820</td>
<td>1,344</td>
<td>2,239</td>
<td>0</td>
<td>1,032</td>
<td>1,294</td>
<td>1,808</td>
<td>921</td>
</tr>
<tr>
<td>New York</td>
<td>2,502</td>
<td>1,820</td>
<td>2,611</td>
<td>3,37</td>
<td>2,239</td>
<td>0</td>
<td>629</td>
<td>2,534</td>
<td>2,502</td>
</tr>
<tr>
<td>San Francisco</td>
<td>951</td>
<td>1,188</td>
<td>1,820</td>
<td>3,37</td>
<td>2,239</td>
<td>0</td>
<td>629</td>
<td>2,534</td>
<td>2,502</td>
</tr>
<tr>
<td>Seattle</td>
<td>551</td>
<td>1,344</td>
<td>1,615</td>
<td>1,920</td>
<td>2,611</td>
<td>0</td>
<td>2,534</td>
<td>2,502</td>
<td>0</td>
</tr>
<tr>
<td>Washington DC</td>
<td>2,123</td>
<td>2,611</td>
<td>1,820</td>
<td>3,37</td>
<td>2,239</td>
<td>0</td>
<td>2,534</td>
<td>2,502</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4.3: Distance matrix

Figure 4.4: MDS map

An MDS of these data gives the picture in Fig. 4.4, a map of the relative locations of these 10 cities in the United States. This map has 10 points, one for each of the 10 cities. Cities that are similar (have short flying mileages) are represented by points that are close together, and cities that are dissimilar (have large mileages) by points far apart.

**k-means clustering**

The simplest definition of clustering would be “grouping together of similar data items”. As a more elaborate definition, the clusters obtained should reflect a mechanism that causes some instances to bear a stronger resemblance to one another, than they do to the remaining instances. The goal is to partition X instances into K groups such that members of the same group are more “alike” than those in different groups. Each of these K groups is called a “cluster”. We refer to instances as “points” henceforth.
One of the first clustering algorithms ever proposed, the \( k \)-means algorithm is based on a very simple idea: Given a set of initial clusters, assign each point to one of them, then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence (see Fig. 4.5). A point is assigned to the cluster which is close in Euclidean distance to the point.

![Figure 4.5: (a) The \( k \) initial “means” (in this case \( k=3 \)) are randomly selected from the data set. (b) \( k \) clusters are created by associating every observation with the nearest mean. (c) The centroid of each of the \( k \) clusters becomes the new means. (d) Steps (a) and (b) are repeated until convergence has been reached. Reproduced from Wikipedia [13].](image)

### 4.4.2 Results

As in earlier section, we begin by computing the correlation coefficients defined by Eq. (4.1), that fulfill the condition \(-1 \leq \rho_{ij} \leq 1\) and form an \( N \times N \) correlation matrix \( C^t \). We then make the non-linear transformation

\[
d_{ij} = \sqrt{2(1 - \rho_{ij})}
\]

to obtain “distances” with the property \( 2 \geq d_{ij} \geq 0 \), forming an \( N \times N \) distance matrix \( D^t \), such that all distances are “ultrametric”. This matrix is then used for running the classical MDS algorithm and then do clustering.

#### Synthetic data

First we simulate synthetic data where we can control the correlations amongst the \( N \) assets. We then use the classical MDS algorithm and \( k \)-means clustering to verify that we get the expected results.

From Fig. 4.7 it is quite clear that as the mean cross-correlations amongst the assets in a cluster increases, and the correlations between the clusters decreases, the dots that represent assets come closer to each other, but each cluster moves farther away from each other.
Figure 4.6: MDS maps using synthetic data: We generate returns for 40 assets (represented by dots), such that there are 4 clusters (represented by four colours) of 10 assets each. Within each cluster the mean cross-correlation coefficient is $\bar{\rho}$. Correlations increasing from (a) to (d): $\bar{\rho} = 0.13, 0.21, 0.29, 0.37$, respectively. The centroids of each cluster are plotted in magenta filled circles.

**Empirical data**

Next, we work with empirical data. For this purpose we use daily closure prices for 40 assets in 5 years span. We use a time window size of 30 trading days to compute the returns and correlation coefficients.

We then observe different MDS maps for different days and see their time evolution using a GUI prepared in MATLAB (see Fig. 4.7).
Figure 4.7: Screenshots that demonstrate the GUI to plot MDS maps using empirical data: We generate returns for 40 assets (represented by dots) and then plot for different days.
Chapter 5

Brief outlook and future work

5.1 Analyses using Moving averages

The entire study is currently with respect to the MA(2) rule. We could extend it to other members of the MA() family, specially the highly popular MA(21, 12) rule, and use empirical data.

In the model with no-memory (where $S_t$ does not depend on $S_{t-1}$), $\pi$ seems to be the controlling factor for $E(RR)$. It would be interesting to understand better this effect of $\pi$ with respect to $\alpha_0$ and $\alpha_1$, i.e. the two drifts. This would give us an insight as to which returns distribution goes best with our strategy.

In the model with memory (where $p_{00}$ and $q_{11}$ exist, and $S_t$ depends on $S_{t-1}$), the effect of the two said probabilities seems to be microscopic. Further in-depth study of this effect would improve the overall understanding of the rule and strategy.

Now that we have established that the said strategy seems to work for a certain type of returns, another interesting angle would be to compare it with the most basic buy and hold strategy (which always seems to work on any and all kinds of returns).

5.2 Analyses using Correlations

As pointed out in Ref. [14], where they worked with different markets instead of individual stocks, this type of graphical method to visualize possible time-varying correlations between stock returns, can be useful for observing stable or emerging clusters of stocks with similar behavior. The graphs using the MDS, may also guide the construction of multivariate econometric models.

In future, we would like to specially consider the following:

- Avoid overlapping windows, so that the price history is not mixed.
- Use high frequency intra day data.
- Capture non-linear effects in correlations.
- Make a comparative study of the different methods for studying correlations and clustering.
Bibliography


