Stock market analysis using Multidimensional Scaling

Ravee Malla*
2008CS50224
CSC 410 Colloquium
Indian Institute of Technology Delhi
ravmalla@gmail.com

ABSTRACT

We describe the application of a statistical technique known as *Multidimensional Scaling*[1] to analyze & model a stock market[2]. We have reviewed a set of novel papers in this field, and report on the current state of the art. The strength of this approach is in it's ease of implementation, effective visualization and dynamic nature.We give a brief introduction to MDS, and state the assumptions & problems in stock market modeling. The approach developed is modified to study the time-evolving nature of the markets. Finally we introduce the notion of an *Asset Tree* which can be used to make quantified conclusions of the state of the market.

1. INTRODUCTION

It is commonly remarked that Economics is too serious a subject to be left to Economists. This is very true when one considers the plethora of parameters and variables that interact to result in macro & micro economic phenonmenon. This is easily observed in the context of stock markets. The New York Stock Exchange[3] is the world's largest stock market. On an average day, nearly $73 billion are traded on this exchange and 1.6 billion shares exchange hands. There are currently around 2,300 companies listed on the exchange worth roughly $12.4 trillion (September 2011). In India, there are around 7000 listed companies in total at various stock markets, the biggest being National Stock Exchange[4].

Stock markets are a vital component of any country's economic prowess. Economists use the stock markets as an indicator of the state of the Economy, and design policies to improve the state of affairs. Investors try to predict the movement of the market to decide which companies to invest in and maximise their profits. Some of the most fundamental problems in characterising any stock market are to

- Identify *trends* & *clusters* in the market
- Design efficient strategies for investor risk reduction
- Study the time evolution of the market
- Design an analysis framework that is easy to visualize
- Build models to predict future trends of the market

Organization: We summarize some related work in Section 2 and introduce some Economics Preliminaries in Section 3. We then go on to give a formal description of the problem and analyze it's hardness. We then introduce MDS & apply it to our problem. We conclude by showing some results.

2. LITERATURE SURVEY

Stock Price modeling is a well studied problem. Past work [3, 8, 11] has focussed on modeling stock prices as stochastic processes and random walks. Alternatively, one can look at a more *macro* effect by combining stock prices of various companies to make general qualified statements about the market itself. In this report we study the latter approach by reviewing two interesting papers [6, 7] which use the technique of MDS & Asset Trees to study stock markets. MDS [9, 2] is a well studied statistical technique to study similarities in large datasets. As we show later, the problem of interest here is NP-hard. Hence, rather than solve the problem completely, these papers have shown how these techniques can help tackle the problem to a fair extent.

3. ECONOMICS PRELIMINARIES

3.1 Stock

A stock (also known as a share) is a notional unit of ownership of a company. A company offers a part of it's ownership in units of stock to raise capital from the market.

3.2 Stock Market

In our model, we define a market to be a set $\mathcal{C}$ of all the listed companies that trade their stocks.

3.3 Stock Price

Stock price is a function $p : \mathcal{C} \times T \rightarrow \mathbb{R}_+$ which assigns a price to every stock traded in the market. It is a result of complex dynamics of market demand & supply. The set T is the time index set at which the stock is to be priced. Time can be measured on multiple granularities, however it is most commonly reported on a daily basis.

3.4 Return on Stock

Return on Stock models the *utility* of a particular stock with respect to the market. It can be thought of as a function $r : p \times \mathcal{C} \times T \rightarrow \mathbb{R}$. Given the price function p, it takes a stock and predicts it's utility around a time point τ. There are various ways of defining a return function, and it critically impacts the quality of the model. The daily return function

*Colloquiim Report supervised by **Dr. Amit Kumar**
[1]hereafter referred to as MDS
[2]`http://en.wikipedia.org/wiki/Stock_market`
[3]`http://nyse.com`
[4]`http://nseindia.com`

is defined as

$$r(p, c, \tau) = \ln \frac{p(c, \tau)}{p(c, \tau - 1)} \quad (1)$$

We take the logarithmic fractional increase of the stock price of day τ from it's previous day $\tau - 1$, which is the mostly widely used function.

3.5 Portfolio

A portfolio $\mathcal{P}$ is defined as a 5-tuple $\langle \mathcal{B}, p, r, \mathcal{N}, \tau \rangle$ where $\mathcal{B} \subset \mathcal{C}$ & $\mathcal{N} : \mathcal{B} \rightarrow \mathbb{Z}_+$. $\mathcal{B}$ represents a subset of the listed companies (possibly a sector) that an investor or an economist is interested in. $\mathcal{N}$ determines the number of stocks bought of each element of $\mathcal{B}$. A portfolio is dynamic & time-evolving. In most practical cases, an additional parameter Γ is added to indicate the maximum investment that can be made into the market.

4. PROBLEM OVERVIEW

4.1 Portfolio Optimization

An instance of PORTFOLIO OPTIMIZATION consists of a portfolio $\mathcal{P}$ & a constant Γ. The objective is to design a portfolio $\mathcal{P}$ to

$$\begin{aligned} \underset{\mathcal{B} \subset \mathcal{C}}{\text{maximize}} \quad & \sum_{b \in \mathcal{B}} r(p, b, \tau) * \mathcal{N}(b) \\ \text{subject to} \quad & \sum_{b \in \mathcal{B}} p(b, \tau) * \mathcal{N}(b) \leq \Gamma \end{aligned} \quad (2)$$

The objective is to pick out a subset of companies $\mathcal{B}$ & buy a certain quantity of shares $\mathcal{N}(\mathcal{B})$ so as to maximize the total return obtainable at that time given the return on stock prices. Γ limits the total investment that can be made to buy the stocks of the set $\mathcal{B}$. PORTFOLIO OPTIMIZATION is NP-complete.

KNAPSACK $\leq_P$ PORTFOLIO OPTIMIZATION

1. PORTFOLIO OPTIMIZATION is in NP. A particular subset $\mathcal{B}$ serves as a certificate so that when we compute the total return on $\mathcal{B}$, we get a sum of amount atleast k.
2. PORTFOLIO OPTIMIZATION is NP-hard. The KNAPSACK problem can be reduced to the decision version of this problem by considering the companies to be the set of objects and the stock prices as the object values. We assign Γ as the total permissible capacity C of our knapsack.

4.2 Diversification

A related paradigm of portfolio designing is to chose companies from isolated & independent sectors. This reduces the overall risk involved as it is very unlikely that many uncorrelated sectors go down together. This is an additional constraint on the choice of the set $\mathcal{B}$. We want the companies to be forming an *independent set* of some sorts.

5. MULTIDIMENSIONAL SCALING

5.1 Overview

MDS is a statistical technique to explore similarities in large datasets and visualize them on a map. Given a geographic map of cities, one can easily determine the inter-city distances on the map. The problem of MDS is exactly the inverse of this problem. It is stated as: Given a set of distances $d_{ij} : \mathcal{C}^2 \rightarrow \mathbb{R}$ on an entity set $\mathcal{C}$, learn the best fit function $f : \mathcal{C} \rightarrow \mathbb{R}^m$. The function assigns m-dimensional coordinates to the entities which can be visualized in an m-dimensional Euclidean space. Figure 1 shows MDS working on a set of 10 American cities.

Table 1 Flying Mileages Between 10 American Cities

Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington, DC	
0	587	1212	701	1936	604	748	2139	2182	543	Atlanta
587	0	920	940	1745	1188	713	1858	1737	597	Chicago
1212	920	0	879	831	1726	1631	949	1021	1494	Denver
701	940	879	0	1374	968	1420	1645	1891	1220	Houston
1936	1745	831	1374	0	2339	2451	347	959	2300	Los Angeles
604	1188	1726	968	2339	0	1092	2594	2734	923	Miami
748	713	1631	1420	2451	1092	0	2571	2408	205	New York
2139	1858	949	1645	347	2594	2571	0	678	2442	San Francisco
2182	1737	1021	1891	959	2734	2408	678	0	2329	Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington, DC

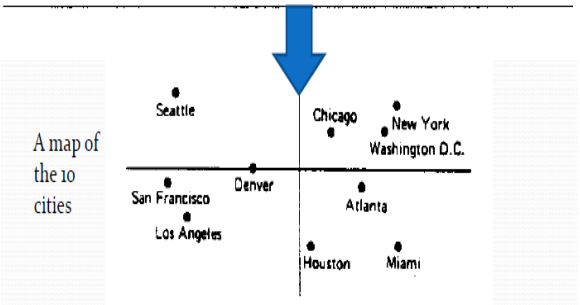

Figure 1: MDS implementation on a set of 10 American cities

5.2 MDS Stress function

MDS operates as a stress minimization algorithm which tries to match the distances in the coordinate space and the actual distance matrix. For each entity $c \in \mathcal{C}$, an m-dimensional vector $X^c = (x_1^c, x_2^c, ..., x_m^c)$ is chosen randomly. A stress function S_1 is defined as:

$$S_1 = \sum_{i,j \in \mathcal{C}, i \neq j} \sqrt{{d_{ij}}^2 - \|X_i - X_j\|^2} \quad (3)$$

Multiple stress minimization techniques like Simulated Annealing or Random Perturbation can be used in order to get an approximate solution.

5.3 Objective Minimization

MDS can be reformulated as a Linear Program and hence is in P. However, the Simplex algorithm may not give back optimal results in reasonable time. [1, 4] are two popular methods of optimizing MDS stress functions. Many Genetic Algorithm based techniques have also been tried out[5].

6. USING MDS ON THE MARKET

If one could apply MDS on the market, one can visualize companies and their interactions in an easy to visualize map.

6.1 Measure on the market

MDS requires a notion of distances between entities. Hence, we first define a measure on the market. There are certain properties that this measure must obey.

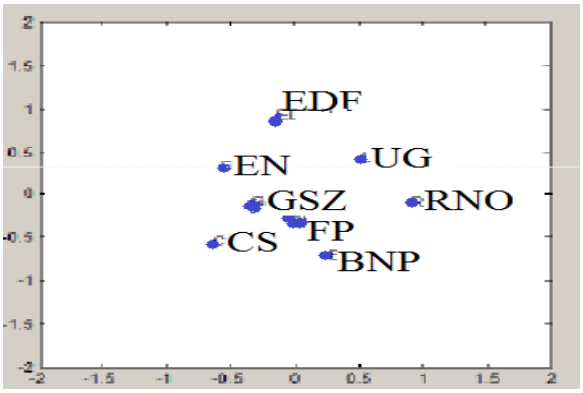

Figure 2: The subset of the European market visualized on a map. Proximity implies correlation.

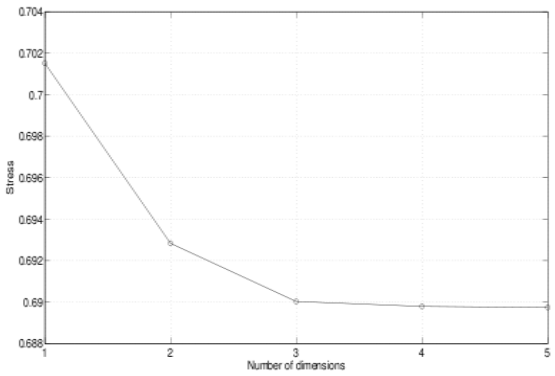

Figure 3: Minimum Stress versus the allowed degrees of freedom

- Model complex interactions and temporal evolution faithfully, while abstracting out needless details
- Helpful in finding clusters & trends in the markets
- Well behaved & easy to compute

Both papers note Pearson correlation as one such measure. It is defined as $\rho : \mathcal{C} \times \mathcal{C} \times T \to [-1, 1]$

$$\rho_{ij}^{\tau} = \frac{\langle r_i^{\tau} r_j^{\tau} \rangle - \langle r_i^{\tau} \rangle \langle r_j^{\tau} \rangle}{\sqrt{[\langle r_i^{\tau 2} \rangle - \langle r_i^{\tau} \rangle^2][\langle r_j^{\tau 2} \rangle - \langle r_j^{\tau} \rangle^2]}} \tag{4}$$

Here $\langle . \rangle$ represent a time-weighted average (expected value) in a window around τ. The window w is usually kept around 50 days, but can be varied as per requirement. r_i^{τ} represents the return on the stock of company i at time point τ. The paper justifies the use of simple correlation as a good measure by noting that

- Globally high ρ implies that market is generally correlated, so it is either in a recession or boom phase
- Treasure trove of rich probability theory applicable
- Easy to compute and analyze

6.2 Metric on the market

For the correlation measure to qualify as a distance, we need to transform it suitably. [7] suggests the following transform:

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

For highly correlated entities, $\rho_{ij} \to 1$ and hence $d_{ij} \to 0$. Hence the distance $d \in [0, 2]$ represents promixity in correlation space. One can verify that d preserves all the properties of a metric i.e. $\forall i, j, k$

- $d_{ij} \geq 0$
- $d_{ii} = 0$
- $d_{ij} < d_{ik} + d_{kj}$
- $d_{ij} = d_{ji}$
- Ultrametricity: Mapping preserves topology

6.3 MDS Static Maps

Figure 2 shows how the result of applying MDS to some market at one time instant. The paper notes that while MDS can create a geometric embedding of any dimension, it is favourable to restrict the problem to 2 dimensions for the following reasons

- Ease of visualization in 2 dimensions
- Stress doesn't decrease significantly in higher dimensions (illustrated in Figure 3)

6.4 MDS Time Evolving

A key component of market analysis is to study the temporal evolution of the market. We have a different distance function d^{τ} for each time point τ and hence the maps generated would be slightly different. However, if we were to create these maps in isolation, they may turn out to be completely unoriented. Hence, we add another component of stress to penalize different across maps along with component S_1 defined in (3). This serves to *stitch* consecutive maps.

$$S_T^{\tau} = S_1^{\tau} + \sum_i w_i * \|X_i^{\tau} - X_i^{\tau - 1}\| \tag{5}$$

There are various possibilities for the weighting factor w_i. One possibility is to set it as the market capitalization[5] M_i. The rationale is to have larger companies more rigid in the maps and smaller companies move around more freely.

7. ASSET TREE

Vandewalle et al [10] introduces the notion of an *Asset Tree* where essential characteristics and properties are described by measures on a tree. Define $G^{\tau} = (\mathcal{C}, \mathcal{C} \times \mathcal{C}, w)$ as a weighted graph on the market with nodes as companies & edge weights as the company distances. On this graph, a minimum spanning tree T^{τ} can be computed using Kruskal's or Prim's algorithm. The tree now has only $|\mathcal{C}| - 1$ most dominant edges (least distance, max correlation). Certain special properties of the tree can be studied.

[5] M_i = total stocks * stock price

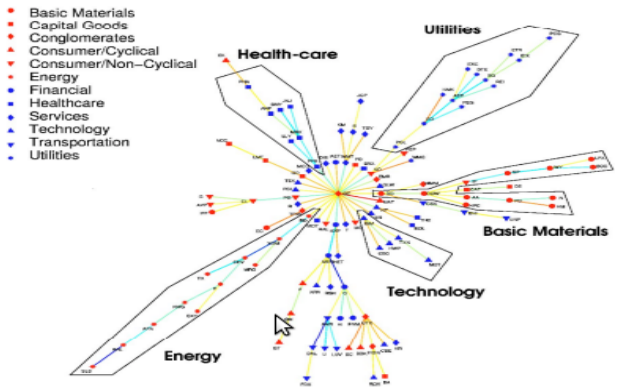

Figure 4: Asset Tree to examine 116 stocks of S&P 500 index. Reference [7]

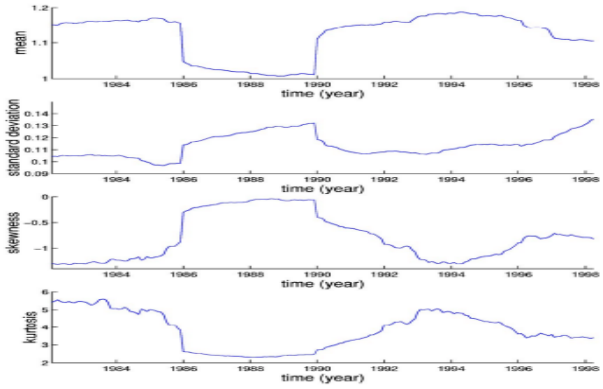

Figure 5: Statistical plots of the tree edge lengths as a function of time

- A central node v_c which can be defined as either the highest degree vertex or the most strongly correlated one. This node represents a company that is an important representative of the market as a whole. If the company falls, there is high chance the market falls as well.
- Mean occupation layer $l(\tau, v_c) = \frac{1}{|C|}\sum_i depth(v_i^\tau)$ reflects the level of correlation in the market by answering where the bulk of the tree is located.
- Center of mass $v_m = \{v | v \in V, l(\tau, v) \text{ is min}\}$

7.1 Clusters of the Tree

Onnela et al [7] suggest that clusters on the tree reflect the true economic sectors that companies belong to in the market. Under the assumption of market fundamentality[6], markets are naturally divided as clusters and this is reflected in Figure 4.

7.2 Scale Free Structure

Vandewalle et al [10] report that Asset Trees have a scale free power law degree distribution. The degree distribution of the vertex degrees $f(n)$ follows $f(n) \sim n^{-\alpha}$ with $\alpha \approx 2.2$.

8. RESULTS

We present below a few of the results demonstrated in the paper of using MDS & Asset Trees in markets.

8.1 Asset Tree

In Figure 4, we find that different sectors are well separated as different branches of the tree and larger companies appear closer to the root node, which is the central node of the tree. The paper also reports how well the tree properties correspond to real events in the market. In particular, Figure 5 shows how the period from 1986 to 1990 stands with the market more correlated than normal. This immediately reflects the possibility of a recession in the market at that time which is true. We find that individual markets got correlated & came together due to a recession. We can define global market score looking at the maps and do a regression analysis to predict future trends.

[6]All investors have access to uniform information

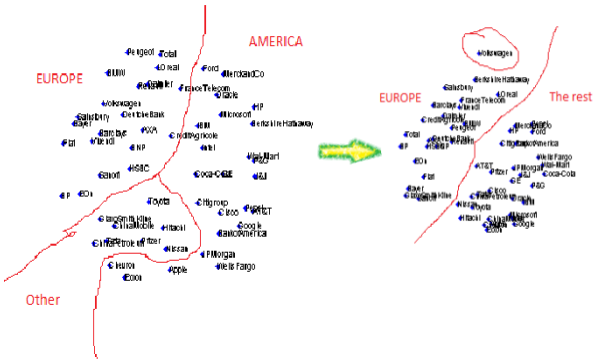

Figure 6: MDS Map evolution from 2006 to 2008 when it becomes very correlated

8.2 MDS Maps

In Figure 6, we can see how companies get placed on a map and how they change their coordinates with time. We can see how companies that are well clustered accordingly to their sector & market, get organized into one strong cluster as we move towards 2008. If one was to define measures on the map such as average entity distance or moment of inertia about a fixed point, one could get extract more useful information.

9. CONCLUSION

In this report, we summarized the technique of Multidimensional Scaling and how it has been applied in the domain of Stock Market analysis. We motivated the problems faced by investors & economists, and introduced PORTFOLIO OPTIMIZATION. We proved the problem to be NP-complete and then suggested how MDS can be used to make market analysis more intuitive and visual. By finding a geometric embedding of companies over time, we can immediately view companies that are far & close. A simple investment strategy could be to invest in companies as far away as possible. This helps restrict our choices of the set $\mathcal{B}$ and make the problem tractable. We then introduced the notion of an Asset Tree and report on some findings of the paper. We finally showed some results that were presented in the papers.

10. REFERENCES

[1] Seung-Hee Bae, Judy Qiu, and Geoffrey C. Fox. Multidimensional scaling by deterministic annealing with iterative majorization algorithm. 03/31/2010 2010. Submitted to InfoVis 2010.

[2] I. Borg and P. Groenen. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 40(3):277–280, 2003.

[3] L Harris. Stock price clustering and discreteness. *Review of Financial Studies*, 4(3):389–415, 1991.

[4] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[5] Abdullah Konak, David W. Coit, and Alice E. Smith. Multi-objective optimization using genetic algorithms: A tutorial.

[6] J. Tenreiro Machado, Fernando B. Duarte, and Gonçalo Monteiro Duarte. Analysis of stock market indices through multidimensional scaling. *Communications in Nonlinear Science and Numerical Simulation*, 16(12):4610 – 4618, 2011.

[7] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. 68(5):056110, November 2003.

[8] Ping-Feng Pai and Chih-Sheng Lin. A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6):497 – 505, 2005.

[9] Warren Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952. 10.1007/BF02288916.

[10] N. Vandewalle, F. Brisbois, and X. Tordoir. Self-organized critical topology of stock markets. *eprint arXiv:cond-mat/0009245*, September 2000.

[11] Walter Willinger, Murad S. Taqqu, and Vadim Teverovsky. Stock market prices and long-range dependence. *Finance and Stochastics*, 3:1–13, 1999. 10.1007/s007800050049.