Correlation Based Clustering of Stocks Traded in NSE

***Dr. S.Sampath**
****R.Senthil Kumar***

1. INTRODUCTION

Financial market can be regarded as a complex system. The macroscopic patterns in finance, such as exchange rate, stock prices etc. are made up by the collective behavior of companies and individuals. The characteristic of a financial market lies in the huge amount of data collected making the system well defined which allows detailed statistical analysis of the system. Similarities between traditional science subjects and financial markets make techniques originally constructed for traditional subjects applicable to the field of theoretical finance. (Mantegna and Stanley (2000)).

The problem of quantifying cross-correlation is important, not only from the point of view of understanding collective behavior between the constituents of a complex system, but also from the point of view of estimating the risk of an investment portfolio (Plerou et al. (2001)). Correlation between assets in a portfolio is of fundamental importance when one attempts to diversify investments, for example, when trying to reduce exposure to a sector or industry specific stocks.

The importance of correlation in portfolio optimization was first addressed by Markowitz (1959) in his Capital Asset Pricing Model (CAPM). The use of correlation also plays a fundamental role in more recent techniques in theoretical finance, such as value at Risk (Embrechts et al. (1999)) and Arbitrage Pricing Theory (Campbell et al. (1997)).

Strong correlation between certain groups of stocks would indicate that the financial market is affected by common economic factors. From an economic point of view, it would be of interest to have a classification, clustering of stocks based only on their correlation with other stocks. The co-movement plays an important role in asset allocation and a clustering based on this could, for example in the case of asset diversification, be more useful than a clustering based solely on specific business activities.

Most of the current research on correlation between individual assets focuses on large stock exchanges, such as the New York Stock Exchange (NYSE) and other world dominating markets**.** In this paper, a study is conducted regarding the inter-stock correlation and the underlying hierarchical structure for the National Stock Exchange of India.

2. CORRELATION AND HIERARCHICAL CLUSTERING OF STOCKS

2.1 Covariance and Correlation between stocks

Consider, for example, the two variables weight X_1, and length X_2, measured on a population consisting of random individuals. Two statistical measurements that help to assess the relationship between two random variables are covariance and correlation. The covariance is defined as

$$COV(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] \tag{2.1}$$

where the expected value of X_1, and $\mu_2 = E(X_2)$. If X_2 tends to be large when X_1 is large and small when X_1 is small, then X_1 and X_2 will have a positive covariance. If, on the other hand, X_2 tends to be large when X_2 is small and large when X_1 is small, then X_1 and X_2 will have a negative covariance. The most common measure of correlation is the Pearson's Product Moment Correlation. The coefficient of correlation between two variables X_1 and X_2 is given by

$$\rho_{ij} = \frac{COV(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}} \tag{2.2}$$

where, $V(X_i)$ is the variance of variable X_i. Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from -1 to +1. A correlation of +1 means that there is a perfect positive linear relationship between the variables; a correlation of -1 means that there is a perfect negative linear relationship (anti-correlation) between variables and 0 indicates that there is no correlation.

Co-movement between individual stocks and between stocks and specific market indices plays an important role in finance. Figure 2.1 shows a comparison between the daily (logarithmic) closing prices for Tata Motor's/Infosys and Reliance Industries Ltd/Grasim Industries Ltd. over the year 2007 – 2008.

* Professor of Statistics, Department of Statistics, University of Madras, Chennai-600005. E-mail:sampath1959@yahoo.com
** Department of Statistics, Loyola College, Chennai-600034

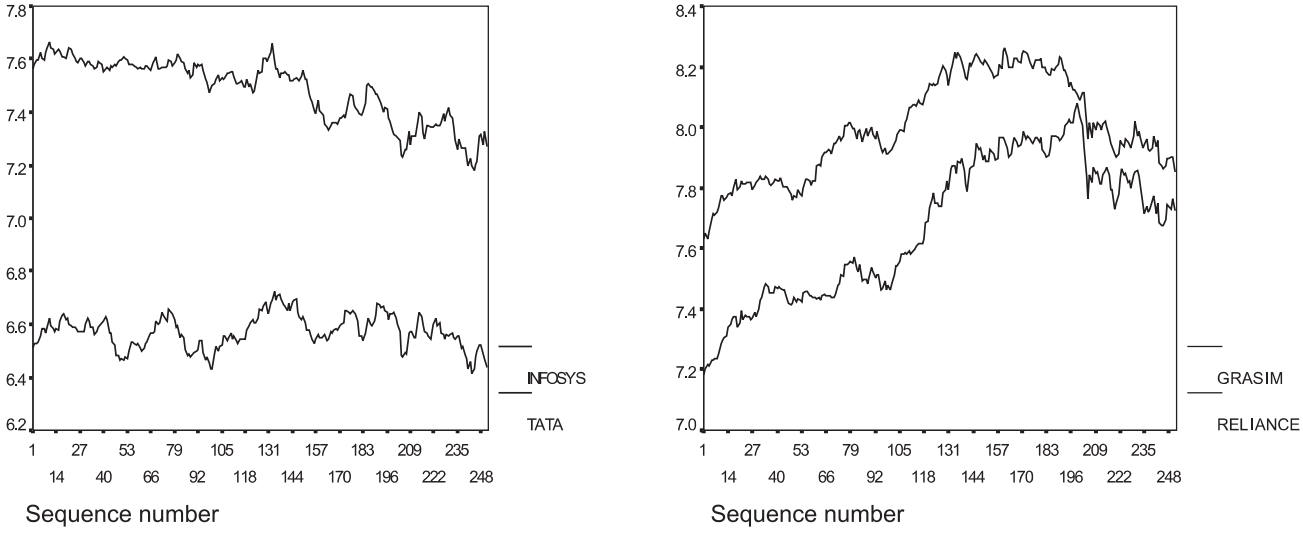

Figure 2.1

A comparison between the daily (logarithmic) closing price for Tata Motor's/Infosys (left) and Reliance Industries Ltd/ Grasim Industries Ltd (right) during the financial year 2007 – 2008.

The daily logarithmic price changes for a stock i is defined as

$$S_i = \ln Y_i(t) - \ln Y_i(t-1) \tag{2.3}$$

where S_i is independent of price scale. The correlation coefficient between the stocks in Figure 2.1 is determined over the 251 trading days of the financial year 2007 – 2008 using (2.2). The correlation between Tata Motor's and Infosys is estimated as 0.244 and the correlation between Reliance Industries Ltd and Grasim Industries Ltd is estimated as 0.861. The result agrees with the visual impression of the compared time series; and time evaluation of $\ln(Y)$ is more coherent for Reliance Industries Ltd and Grasim Industries Ltd compared to Tata Motor's and Infosys.

The difference in co-movement is hardly surprising, given the fact that for both Reliance Industries Ltd and Grasim Industries Ltd, the main business activity is Industrial and they are thus likely to be affected by common economical factors. Tata Motor's (Motor Company) and Infosys (Software Company) do not belong to the same industry sector and are thus less likely to have similar movements.

2.2 DISTANCE BETWEEN STOCKS

The most well known distance metric that we encounter in everyday life is the Euclidean distance (Anton (1994)). Given such a measure, it would be valid to talk about a "distance between individual stocks". For a distance function d_{ij}, to be a valid metric distance, the following four properties must hold:

$$\begin{aligned} &(i)\ d_{ij} \geq 0 \\ &(ii)\ d_{ij} = 0 \Leftrightarrow i = j \\ &(iii)\ d_{ij} = d_{ji} \\ &(iv)\ d_{ij} \leq d_{ik} + d_{kj} \end{aligned} \tag{2.4}$$

To obtain a metric distance based on the correlation coefficient that fulfills the properties of (2.4), Mantegna (1997) proposed the distance function

$$d_{ij} = \sqrt{2(1-\rho_{ij})} \tag{2.5}$$

Note that d_{ij} ranges from 0 for totally correlated stocks to 2 for totally anti-correlated stocks. For uncorrelated stocks, the distance is $\sqrt{2}$.

2.3 Distance Function of Hierarchical Clustering

The first step in performing a cluster analysis is to make an assumption about the topological space linking the objects together. The working hypothesis used by e.g Mantegna (1997) is that a distance space is an appropriate topological space for linking n stocks. The distance matrix must satisfy property (i) -(iii) of (2.4), while property (iv), the triangular inequality, is replaced by the stronger inequality.

$$\hat{d}_{ij} \leq \max\left[\hat{d}_{ik}, \hat{d}_{kj}\right]. \tag{2.6}$$

The property is, according to (2.4), defined as $d_{ij} \leq d_{ik} + d_{kj}$, i.e. the distance between i and is always less or equal to the distance between and , passing through some intermediate point . Equation (2.6) puts an even tighter constraint, in an ultrametric space, the distance between the points i and is always less or equal than the maximum of the distance between and any other point, , and the distance between j and any other point, .

One of the easiest ways of performing the hierarchical clustering is to obtain the Single linkage clustering from the metric distances that link together the objects to be clustered. The algorithm is conceptually described by Tola et al. (2005) in the following way:

Assume that we have a list D consisting of distances between pairs of elements (e.g. stocks) in the system to be clustered (e.g. a portfolio of stocks). Arrange all the distances d_{ij} (the distances between element i and element) in in increasing order. Different elements are iteratively included in clusters, starting from the first two elements of the distance measure ordered list. At each step, when two elements or one element and a cluster or two clusters p and merge in a wider single cluster , the similarity or distance between the new cluster t and cluster is determined as

(2.7)

This definition, where the distance between groups is defined as the distance between the closest pair of elements, is called single linkage clustering (nearest neighbor). Alternative ways of linking together separate clusters are average linkage and complete linkage clustering.

3. NSE DATA

The data used in this study is collected from the National Stock Exchange website *http://www.nse-india.com*. Guided by the website *www.equitymaster.com*, top 50 companies (top gainers) were identified based on their return with respect to the financial year 2007 – 2008(251 days). Among the listed 50 companies available on the NSE website, the data was complete in all respects only for 42 of them and the remaining 8 of them had missing values. Hence, it is decided to include only those 42 companies for which full data is available. The closing price of stocks associated with these 42 companies for the 251 trading days during the financial year 2007-2008 are recorded for the analysis. The companies included in the study are given below.

Reliance Industries Limited, Infosys Technologies, Satyam Computer Technology, Tata Steel limited, Reliance Communication, Oil and Natural Gas Corporation Limited, State Bank of India, Tata Motors Limited, Indiabulls Finance Service Limited, Acc Limited, TCS, ICICI bank, ITC Limited, BHEL, Reliance Capital Limited, India Cements Limited, Century Textiles Limited, Sterlite Industries Limited, Steel Authority of India, Tech Mahindra Limited, IVRCL Infrast Limited, Suzlon Energy Limited, Siemens Limited, IFCI Limited, Zee Entertainment Limited, Bombay Dyeing & Manufacturing Company Limited, Bharti Airtel Limited, Bajaj Auto Limited, Unitech Limited, HDFC Limited, Mahindra & Mahindra Limited, Wipro Limited, Grasim Industries Limited, Jai Prakash Associated Limited, Ranbaxy Labs Limited, HDFC bank Limited, Mahanagar Telephone Nigam Limited, Dr Reddy's Lab Limited, Parsvnath Developer Limited, Bajaj Hindustan Limited, NTPC Limited and Sesa Goa Limited.

4. RESULTS ON CLUSTERING

4.1 Hierarchical Clustering of Stocks

In this section, the process of formation of clusters is illustrated by considering five arbitrarily chosen stocks where the distances between clusters are computed based on the average distances of cluster members. The starting point for this example is the daily logarithmic (closing) price changes for five stocks traded on the NSE during the financial year 2007-

2008. The stocks selected for the example are: Reliance Industries Ltd (RELI), Tata Motor's Ltd (TATA), Oil & Natural Gas Corporation Ltd (ONGC), ACC Ltd (ACC) and State Bank of India (SBI).

Using (2.2), the correlation matrix ρ_{ij} for the five stocks is computed as

	RELI	*TATA*	*ONGC*	*ACC*	*SBI*
RELI	1	.39	.88	.31	.95
TATA		1	.46	.45	.28
ONGC			1	.23	.79
ACC				1	.28
SBI					1

The associated distance matrix d_{ij} calculated using (2.5) is

	RELI	*TATA*	*ONGC*	*ACC*	*SBI*
RELI	0	1.10	0.48	1.17	0.31
TATA		0	1.03	1.04	1.2
ONGC			0	1.24	0.64
ACC				0	1.2
SBI					0

Treating each object as a cluster, the clustering commences by merging the two closest items. Since two stocks separated with the shortest distance are RELI and SBI $(d = 0.31)$, we merge them to form a cluster consisting of RELI and SBI. The total number of clusters now becomes 4 and the new distance matrix, where the distances are computed using average distances is

	$(RELI, SBI)$	*TATA*	*ONGC*	*ACC*
$(RELI, SBI)$	0	1.1	0.56	0.68
TATA		0	1.03	1.04
ONGC			0	1.24
ACC				0

The shortest distance in the new distance matrix is 0.56 which is the distance between (RELI, SBI) and ONGC $(d = 0.56)$. Hence we merge them to get the next cluster namely, (RELI, SBI and ONGC). The revised distance matrix for the next level of clustering is

	$(RELI, SBI, ONGC)$	*TATA*	*ACC*
$(RELI, SBI, ONGC)$	0	1.07	1.05
TATA		0	1.04
ACC			0

The shortest distance in the above distance matrix is 1.04 which leads to the merger of TATA with ACC.

	$(RELI, SBI, ONGC)$	$(TATA, ACC)$
$(RELI, SBI, ONGC)$	0	1.06
$(TATA, ACC)$		0

The dendogram picturing the hierarchical clustering just explained above is shown in Figure 4.1. The tree clearly shows that there are two groups of stocks (clusters) in this selected portfolio. In the first cluster we have the Tata Motor's Ltd and Acc Ltd. The second group consist Reliance Industries Ltd and State Bank of India. Out of these five companies, the remaining company is Oil and Natural Gas Corporation Ltd that is the one that is connected with the others.

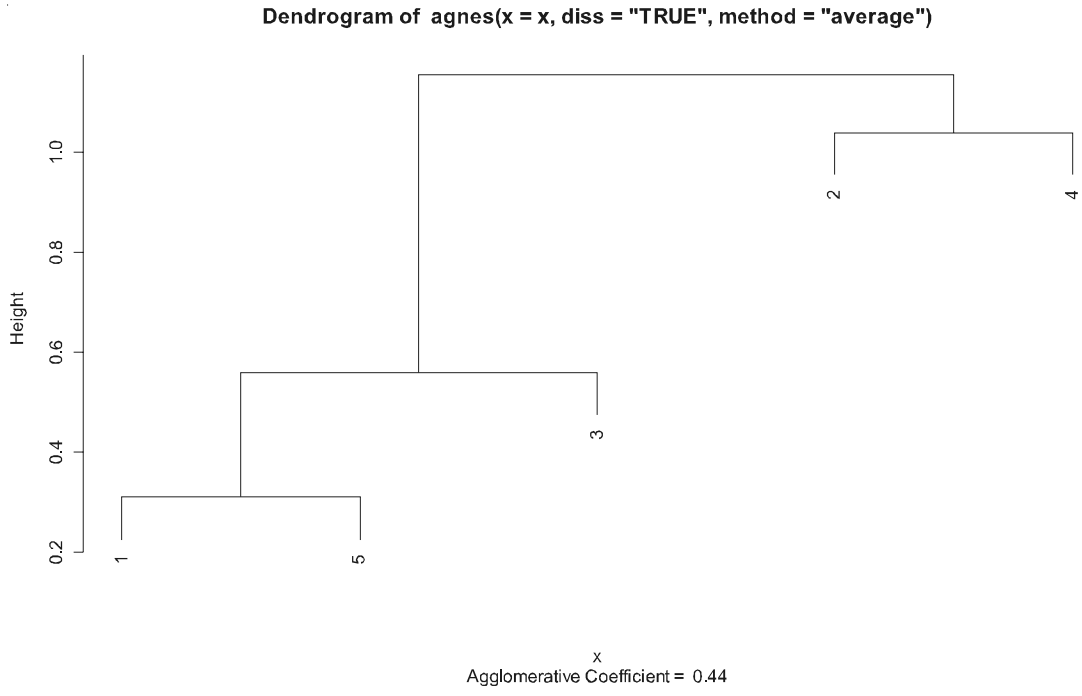

Figure 4.1

4.2 RESULTS AND ANALYSIS

In this section, the topological classification of the top 42 stocks for which the data has been collected is obtained using four methods of clustering, namely single linkage, complete linkage, average linkage and Ward's method. The hierarchical structure using the distance matrix is computed with the help of the process explained in the previous part of this section. The hierarchical structure obtained using single linkage clustering is given in the following diagram.

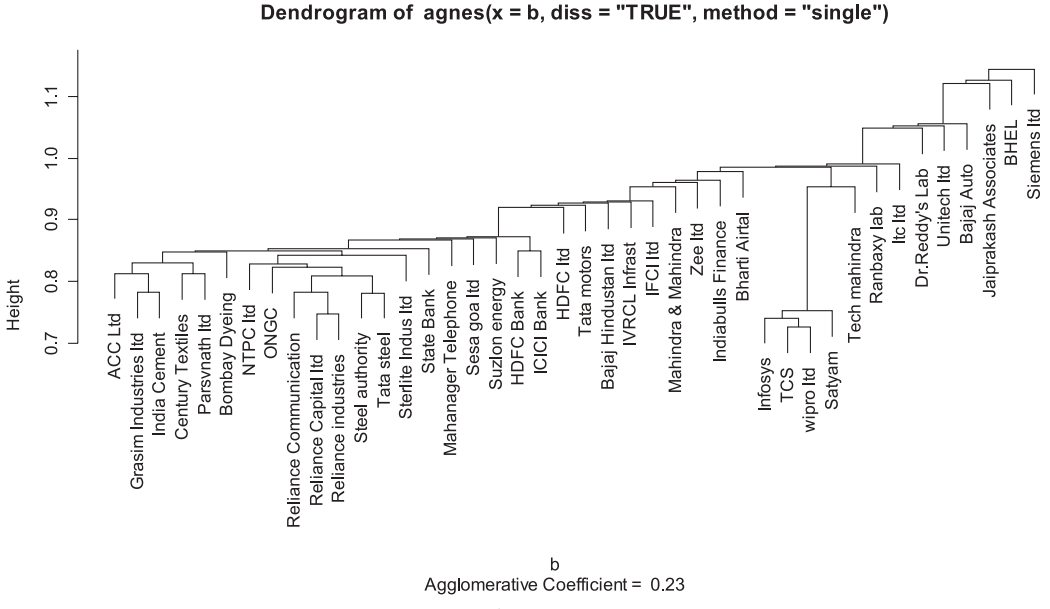

Figure 4.2

It may be noted that for $d \leq 0.85$ there are four clusters. The first cluster consists of the Industrial companies **India Cement, Grasim Industries Ltd and ACC Ltd**. The second cluster consists of the Reliance Group -**Reliance Communication, Reliance Capital Ltd and Reliance Industries**. The third cluster consists of the Commercial Bank companies **HDFC Bank, ICICI Bank and SBI.** The fourth cluster consists of the Software companies **Infosys, Satyam, Wipro and TCS.**

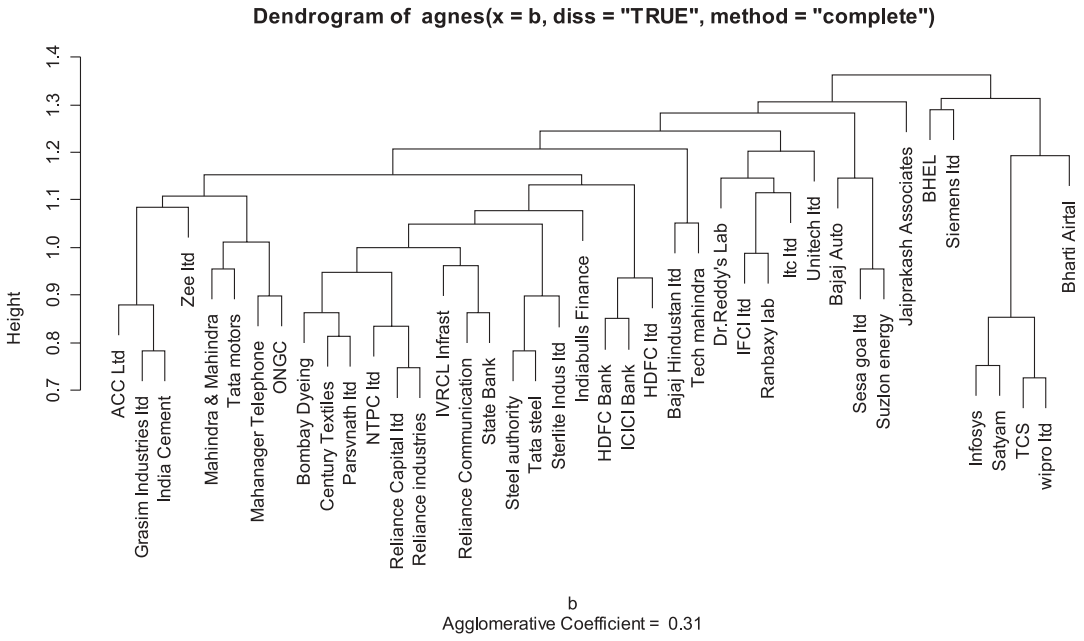

Figure 4.3

The indexed hierarchical tree based on complete linkage clustering (farthest neighbor) is given Figure 4.3. Choosing $d \leq 0.85$ results in six clusters. The first cluster consists of the Industrial companies **Grasim Industries Ltd, India Cement and ACC Ltd.** The second cluster consists of the Textiles companies **Century Textiles, Parsvnath Ltd and Bombay Dyeing.** The third cluster consists of the Reliance group **Reliance Industries Ltd and Reliance Capital Ltd.** The fourth cluster consists of the Steel Companies **Steel Authority and Tata Steel.** The fifth cluster consists of the Bank group **HDFC Bank and ICICI Bank.** The sixth cluster consists of the Software companies **Infosys, Satyam, TCS and Wipro Ltd**.

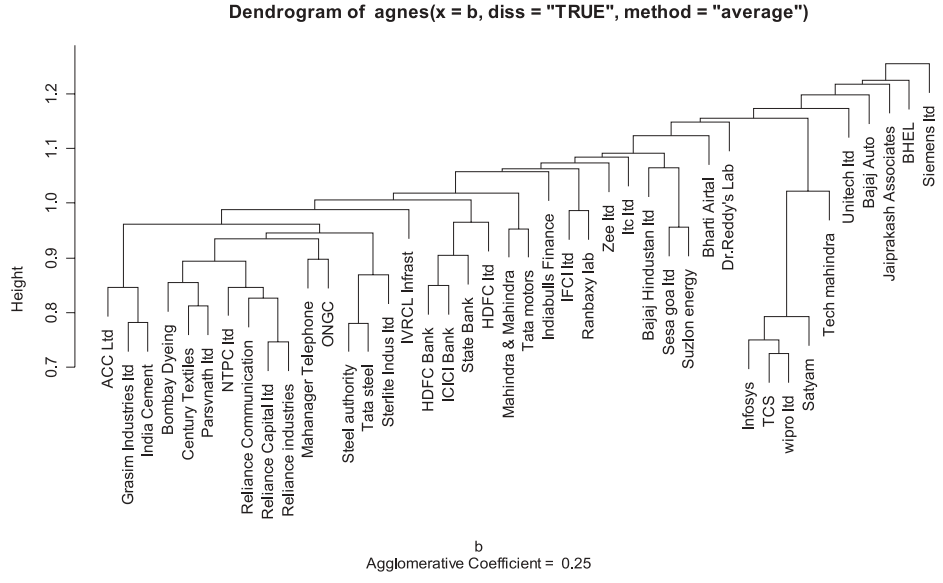

Figure 4.4

The indexed hierarchical tree based on average linkage clustering (nearest neighbour identified using average distances) is given in Figure 4.4. Choosing $d \leq 0.85$ results in five clusters. The first cluster consists of the Industrial companies **Grasim Industries Ltd, India Cement and ACC Ltd.** The second cluster consists of the Textiles companies **Century Textiles, Parsvnath Ltd and Bombay Dyeing.** The third cluster consists of the Reliance Group **Reliance Industries Ltd, Reliance Communication and Reliance Capital Ltd.** The fourth cluster consists of the Steel Companies - **Steel authority and Tata Steel.** The fifth cluster consists of the Software companies **Infosys, Satyam, TCS and Wipro Ltd**.

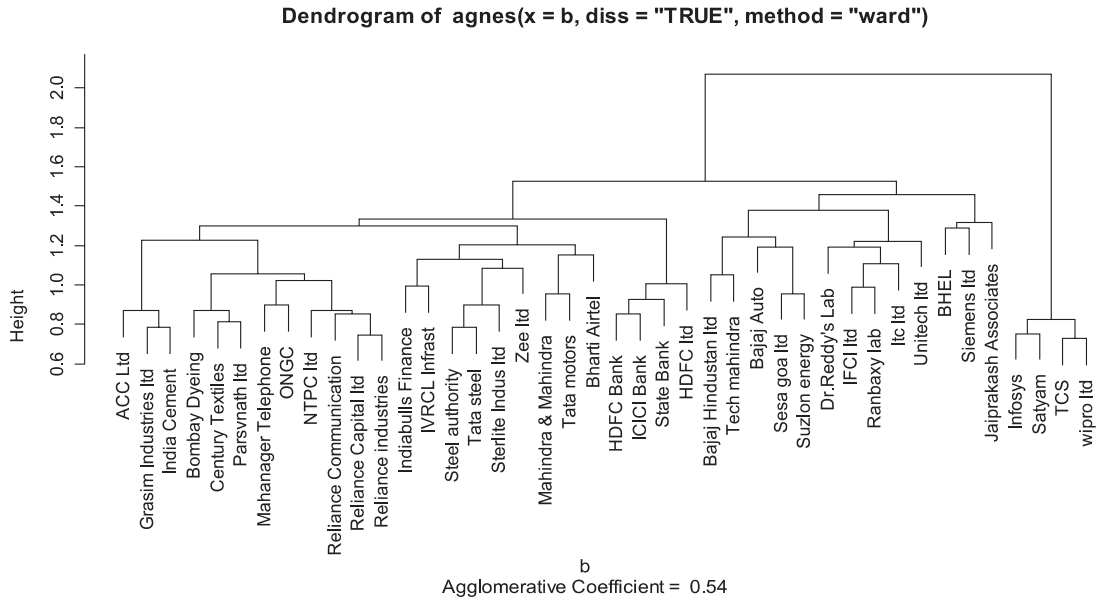

Figure 4.5

The indexed hierarchical tree based on Ward's clustering (based on mean square errors) is given Figure 4.5. Choosing $d \leq 0.85$ results in five clusters. The first cluster consists of the Industrial companies **Grasim Industries Ltd and India Cement.** The second cluster consists of the Textiles companies **Century Textiles and Parsvnath Ltd.** The third cluster consists of the Reliance group **Reliance Industries Ltd and Reliance Capital Ltd.** The fourth cluster consists of the Steel Companies **Steel authority and Tata Steel.** The fifth cluster consists of the Software companies **Infosys, Satyam, TCS and Wipro Ltd**.

Thus we have identified hierarchical structures of stocks based on four clustering methods, namely, "Single linkage", "Complete linkage", "Average linkage" and "Ward's method". In the following section, further analysis is carried out to know the dynamics of the system using the correlation between stocks.

5. RESULTS ON CORRELATION STRUCTURE

The clustering techniques adapted in this paper rely on the correlation between stocks. The spread and time dependence of the correlations are thus of fundamental interest. Figure 5.1 shows the probability density function $P(\rho_{ij})$ observed between stocks in the NSE 42 during the financial year 2007 and 2008. The density was estimated using a kernel smoothing method (Silverman (1986).

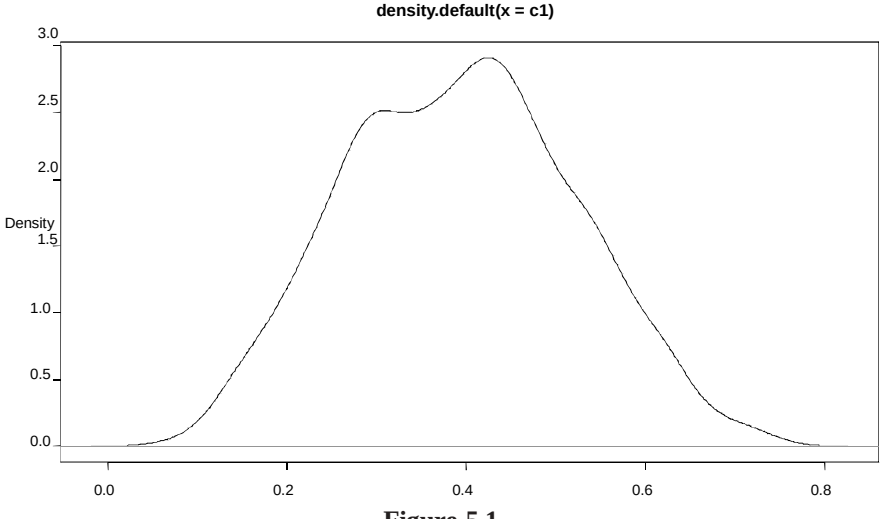

Figure 5.1

Figure 5.1 Probability density function $P(\rho_{ij})$ of the correlation coefficients ρ_{ij} observed between stocks in the NSE 42 during the financial year 2007 to 2008.

The probability density is rather bell shaped and there is a clear overweight towards positive ρ_{ij} . To show the dynamics of the system, Table 1 summarizes the minimum and maximum values of the ρ_{ij} observed between stocks during the financial year 2007 – 2008. The correlations given in the table are computed from daily logarithmic returns using (2.2).

Table 1 minimum and maximum correlation coefficients ρ_{ij} observed between stocks in the NSE 42 during the financial year 01-04-2007 to 01-03-2008.

Time Period	Minimum	Maximum
1^{st} *Quarter*	-0.21	0.73
2^{nd} *Quarter*	-0.15	0.72
3^{rd} *Quarter*	-0.19	0.76
4^{th} *Quarter*	-0.005	0.81
2007 – 2008	0.07	0.71

The maximum correlation $\rho_{ij} = 0.81$ is observed between IVRCL Infrastructure Limited and Zee Entertainment Limited during the fourth quarter. The high degree of correlation is surprising because the companies belong to two different sectors. IVRCL belongs to infrastructure developmental activities whereas Zee Entertainment belongs to media industry. The largest anti-correlation $\rho_{ij} = -0.21$ is observed between BHEL and Steel Authority of India during the first quarter even though both of them belong to the industrial sector.

The maximum correlation observed over the total time period 2007 – 2008 was between Infosys and Satyam Ltd, both Software companies. The minimum correlation for the period 2007-2008 happens to be 0.07. The minimum correlation is between Dr.Reddy's Lab and Siemens which have diversified business interest.

This shows that there is a dynamic behavior of the degree of correlation between stocks for the observed time period. In shorter durations, stocks belonging to even same sector have uncorrelated behavior whereas; over the long run, stocks have good degree of correlation among stocks related in terms of their industry.

6. CONCLUSION

The empirical results show that it is possible to obtain a meaningful taxonomy based solely on the co-movements between individual stocks and the fundamental distance metric assumption, without any presumptions of the companies' business activity. The obtained clusters indicate that common economical factors affect certain groups of stocks, irrespective of their NSE industry classification. The outcome of the investigation is of fundamental importance for e.g. in asset classification and portfolio optimization, where co-movement between assets is of vital importance. It is pertinent to note that the agglomeration coefficient (a measure of clustering validity) is maximum in the case of Ward's method.

The probability density function of the correlation coefficient showed that the distribution is approximately bell shaped and that there is a clear overweight toward positive ρ_{ij} . The minimum and maximum correlations also vary on a quarter-to-quarter basis. However, over the long run, stocks have a good degree of correlation among stocks related in terms of their industry.

It is to be mentioned that the entire computational process has been carried out using R-package for statistical computing. The functions used in the study are available in the packages *cluster* and *MASS* of R 2.5.1.

7. BIBLIOGRAPHY

Anton, H. (1994), *Elementary linear algebra*, 7:th edn, John Wiley Sons, pp. 169-171.

Bernhardsson, J. (2002), *Tradingguiden*, Bokforlaget Fischer Co, Stockholm.

Bonanno, G., Lillo, F. and Mantegna, R. N. (2001), 'Leverls of complexity in financial markets', Physica A **299**.

Campbell, J. Y., W, A., Lo, A. and MacKinaly, C. (1997), *The Econometrics of Financial Markets*, Princeton University Press, Princeton.

Embrechts, P., McNeli, A. and Straumann, D. (1999), 'Correlation and dependence in risk management: Properties and pitfalls', Risk **69-71**.

Johnson, D. E. (1998), *Applied Multivariate Methods for Data Analysts*, 1st ed,. Duxbury Press.

Lo, A. (1991), 'Long-term memory in stock market prices', *Econometria* **59**, 1276-1313.

Mantegna, R. N. (1997), Degree of correlation inside a financial market, in J. Kadtke, ed., Proc. Of the ANDM 97 International Conference', AIP Press.

Mantegna, R. N. (1999), 'Hierarchical structure in financial markets', Eur. Phys. J. B. **11**, 193-197.

(Contd. Page on 31)

$$\ln liq_{BSE} = 2.415 + 2.665\ PC_1 - 0.450\ \text{D-E Ratio} + 0.968\ \text{CA/CL Ratio} \quad \text{……..(1)}$$
$$(2.815^*)\ (10.593^*)\ (-2.846^*)\ (1.749)$$

and

$$\ln liq_{NSE} = 3.536 + 2.527\ PC_1 - 0.633\ \text{D-E Ratio} + 0.583\ \text{CA/CL Ratio} \quad \text{……..(2)}$$
$$(3.665^*)\ (8.932^*)\ (-3.561^*)\ (0.936)$$

Figures in brackets are t- statistic,

* Significant at 1% level

8. CONCLUSION

In an attempt to measure the empirical relationship between scrip level market liquidity of firms listed both in the BSE and NSE with some selected accounting variables, the present study has employed two variable models and multivariable models. The results of two variable models are straight forward. In general, a positive impact of Growth of Assets, Growth of Sales and change in PBIT could be observed on the percentage change in scrip level market liquidity for both the exchanges when Amivest Liquidity ratio has been used. But in case of leverage, no consistent results are found during the study period. Moreover, the study has not found any concrete evidence to establish any relation ship between Current Ratio and Stock market liquidity.

When multiple regressions are taken into consideration, although the predictive power of the model has increased through out the study period but the results of the regressions faces the problem of multi co-linearity. Since the sources of the accounting information are either the Income Statements or the Balance Sheet, the existence of such relationships may be obvious. Therefore, the method of Principal Component Analysis has been adopted to find Principal Component from correlated variables. Results achieved from this mechanism give a conclusion that the joint effect of log of Assets, log of Sales and PBIT on scrip level market liquidity is significant and positive but no specific relationship has been found between the latter and other two accounting variables under consideration.

BIBLOGRAPHY:

1. Amihud, Y. (2002), "Illiquidity and stock returns cross-section and time-series effects", *Journal of Financial Markets*, 5, 31-56.
2. Amihud, Y., H. Mendelson, (1986), "Asset pricing and the bid-ask spread", *Journal of Financial Economics*, 17, 223-249. *The Journal of Finance*, 42, 533-553.
3. Ball, R and Brown, P (1969) , " Portfolio Theory and Accounting Theory ", *Journal of Accounting Research*, 7, 300-323.
4. Barber, B. M. and J. D. Lyon (1997) "Detecting long-run abnormal stock returns: The empirical power and specification of test statistics", *Journal of Financial Economics*, 43, 341-372
5. Beaver B, P. Kettler and M.Scholes (1970) , " The Association Between Market Determined and Accounting Determined Risk Measures", *The Accounting Review* 45, 654-682.
6. Breton, G. and R. J. Taffler (1995) 'Creative accounting and investment analyst response', *Accounting and Business Research*, 25, 81-92
7. Kulkarni, M.,M. Powers and D. Shannon,(1991) "The Use of Segment Earnings Betas in the Formation of Divisional Hurdle Rates," *Journal of Business Finance and Accounting* 18,. 497- 512.
8. Malik, A & Ghosh, S, (1996), " Financial Management Decisions and Risk: A Case with Indian Share Price Data". *Finance India*, X(3), 617-632.

(Contd. from page 19)

Mantegna, R. N. and Stanley, H.E. (2000), *Introduction to Econophysics: Correlations Complexity in Finance*, Cambridge University Press, Cambridge.
Markowitz, H. (1959), *Portfolio Selection: Efficient Diversification of Investment*, J.Wiley, New York.
Mezard, M., Parisi, G. and Virasoro, M. (1987), *Spin Glass theory and Beyond*, World Scientific, Singapore.
Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N. and Stanley, H. E. (2001) 'Collective behavior of stock price movements: A random matrix theory approach', *Physica* A **299**, 175-180.
Schaeffer, R. and McClave, J. (1995), *Probability and Statistics for Engineers*, 4th edition edn, Duxbury Press.
Silverman, B. W. (1986) Density estimation, Chapmen and Hall, London.
Tola, V., Lillo, F., Gallegati, M. and Mantegna, R. N. (2005), Cluster analysis for portfolio optimization. Preprint arXiv:physics/0507006.
West, D. (1996), *Introduction to graph theory*, Prentice-Hall, Englewood Cliffs.