## A Study on Daily Open-to-Close Volatility and Volume of Stocks in Some Asian Markets

Ruochen Zeng, Michael

Section 1 Data and Sources

This project considers the India and Singapore markets for illustration. Daily Open-Close-Volume data of stocks in these markets are required. Also needed are Open-to-Close Volatilities of the corresponding market indices as well as the Open-to-Close Volatility of S&P 500 index.

Python is used for scraping the stock data. First, a collection of stock tickers are obtained online through function downloadEverything(), which is modified from the package “Yahoo\_ticker\_downloader”. Second, this collection of tickers are used for downloading the stock data in OHLC format from Yahoo Finance via function pandas\_datareader.data.DataReader(). Third, KDB is set up and connected to Python. By q.sync(), Python can directly saves the data into KDB in memory and further saves them onto local disk. Finally, KDB is connected to R using library::rkdb. By execute(), R is able to load data from KDB in memory. On the other hand, the index series are downloaded manually from Yahoo Finance, saved as csv. files on the local disk, and imported into R using read.csv(). This project uses R, as some of the packages that are required for data analysis can only be found on R CRAN, e.g. the ‘fda’ for functional data analysis.

The selected period is from 01/01/2017 to 05/31/2017. The reasons are the following:

1. During this period, most listed companies will publish financial reports, which may cause some market movements.
2. Trump administration starts on 01/20/2017. It will be insightful to examine the market responses.
3. The period selected being too long could significantly slow down the data downloading process, and cause difficulties for debugging, due to the limited time constraint.

Open-to-Close volatility for stocks and indices are calculated through a 10-day moving window, such that the volatility on a particular trading day is obtained by computing the standard deviation of the Open-to-Close returns within the past 10 days. The choice of 10 is made, as it will generate sufficient data for the computation. On the other hand, a 10-day duration is still reasonably small.

Section 2 will report results on the India market, and Section 3 on the Singapore market. Section 4 concludes.

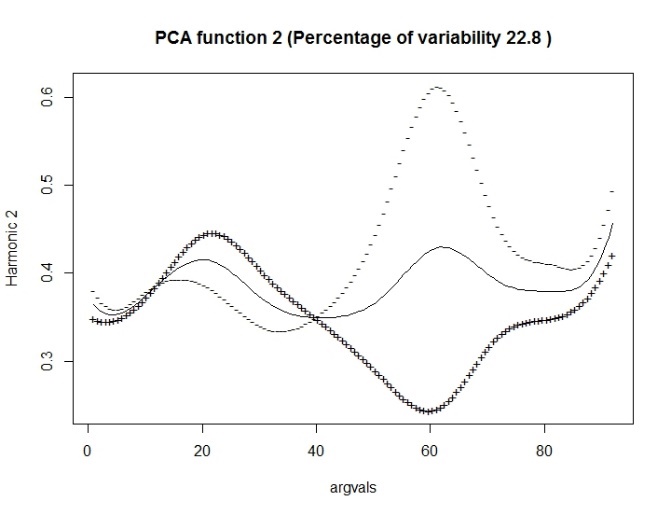
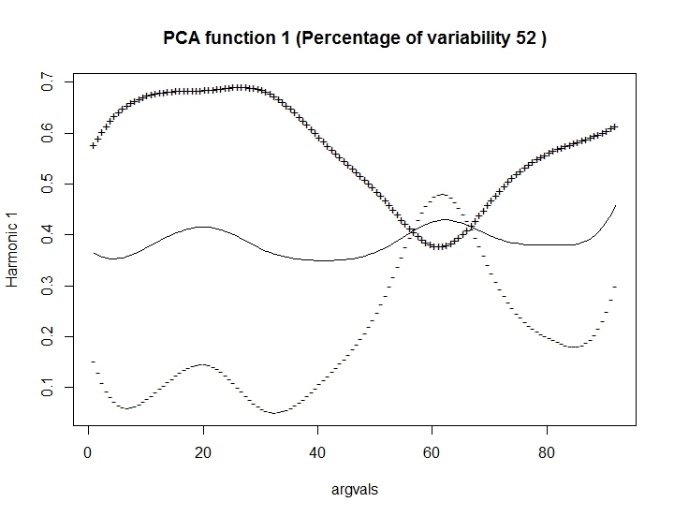
Section 2 India Market

2.1 Daily Open-to-Close Volatility of Stocks and Index

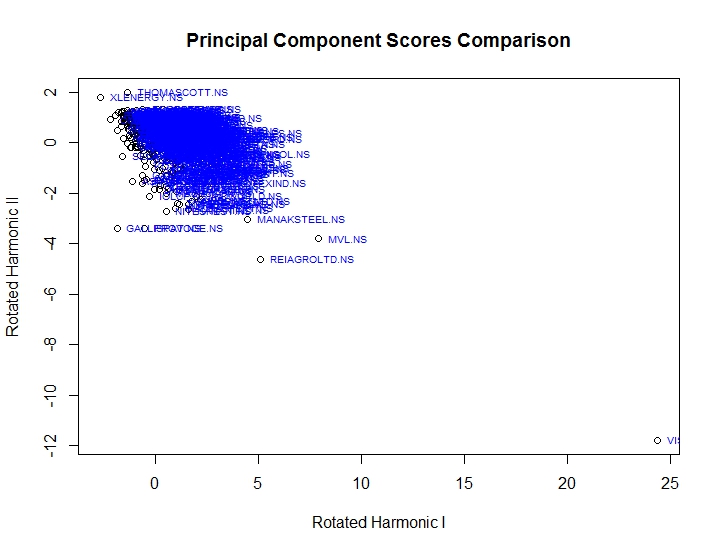
In this section, we aim to explore the relationship between daily open-to-close volatility of stocks in the India market and the open-to-close volatility of the corresponding market index, NSEI index, also known as NIFTY 50. For a given market, the dataset can be perceived in the following way: first, there are p trading days in the dataset, which can be treated as p variables; and there are n stocks in the dataset, which can be treated as n records for each one of the p trading days.

The relationship between stock volatility and the index volatility can be studied for each one of the p trading days. Yet this is not viable, since p could be very large. Moreover, the time t is generally considered continuous rather than discrete, which means the dimension of the variable space should be infinite. To account for the continuity of the time, functional data analysis is evoked, which constructs a (infinite dimensional) function to capture the temporal dynamics of a specific stock. Functional Principal Component Analysis (PCA) can then employed, similar as the multivariate PCA, to reduce the dimension of the (infinite dimensional) variable space. In brief, functional PCA constructs principal component functions, assigning weights to every possible value of time t, which should be continuously valued, so as to maximize the total variation across the time for all the daily stock volatilities.

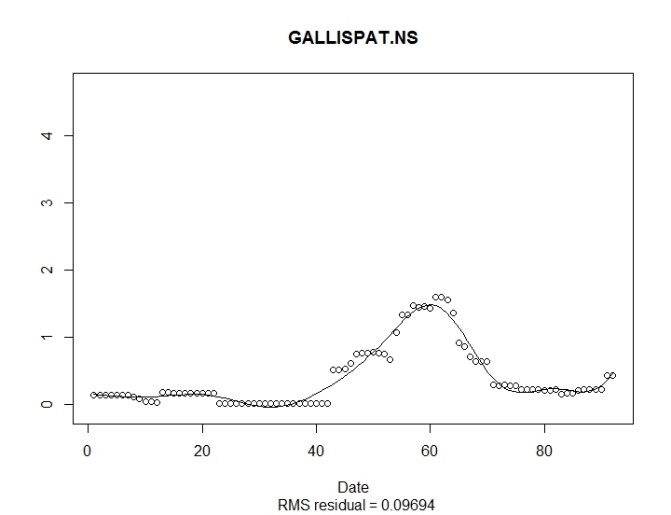
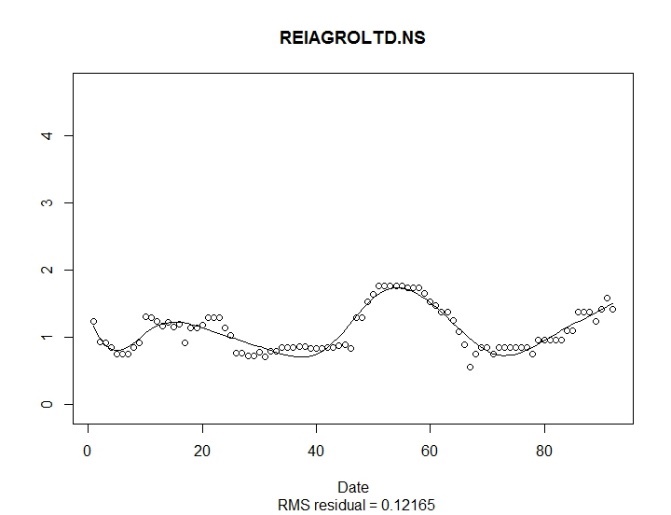
The data come from 330 stocks in National Stock Exchange of India, each with 93 trading records from 01/17/2017 to 05/31/2017. First, a spline basis is set up and used to smooth the volatility data. Then the functional PCA is applied to decompose the variation along the 92 days across the 330 stocks, seeking the most important mode of variation, and displaying it as ﬁrst, second PC functions (harmonic), etc. The next graph shows the first 2 principal component functions by displaying the mean curve along +’s and -’s indicating the consequences of adding and subtracting a small amount of each principal component function. They are presented in this way because a principal component represents variation around the mean, and therefore is naturally plotted as such.



We see that the ﬁrst harmonic, accounting for 52% of the variation, represents variation primarily other than records 50 – 70, i.e. 03/29/2017 to 04/28/2017. The second harmonic shows primarily variation during 03/29/2017 to 04/28/2017, which accounts for 23% of variation contained in the whole dataset. The following ﬁgure plots the principal component scores for pairs of harmonics for all of the 330 stocks.



It shows that most of the stocks are contained within a cluster around the origin (0, 0), indicating small 1st and 2nd PC values. These stocks have relatively small volatility for the whole period. A stock with very negative 2nd PC score has large volatility during the period from 03/29/2017 to 04/28/2017; and a stock with very positive 1st PC score has large volatility during the time other than 03/29/2017 to 04/28/2017. The stock REIAGROLTD has both 1st and 2nd PC score large in magnitude. A plot shows the reason: it has a large volatility across the whole period. The solid curve is the smoothing function, and the dots are original data representing the daily volatility.



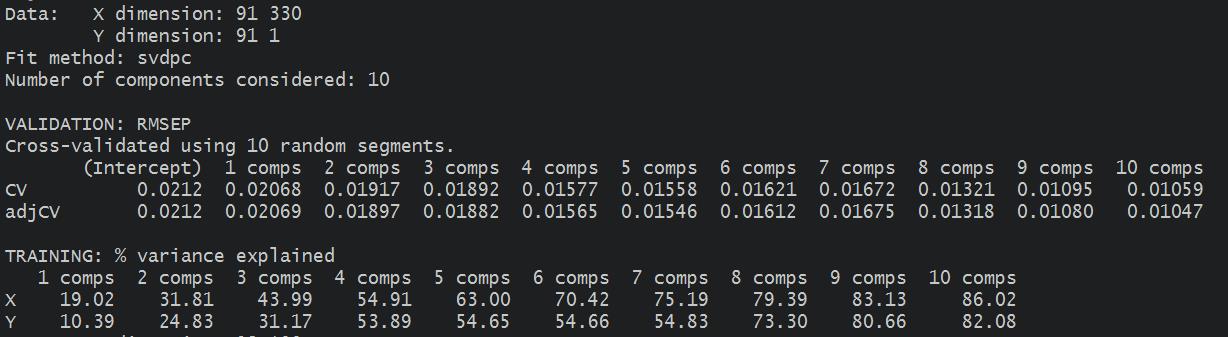
On the other hand, GALLISPAT has a very negative 2nd PC score with low 1st PC score, as confirmed by the above figure. Lastly, a comparison of the following plot of daily volatility of the NSEI index and the PC function plots show an interesting pattern: the period from 03/29/2017 to 04/28/2017, which corresponds to 2nd PC function, has a low volatility for the index, while during the period corresponds to 1st PC function the index, the index has relatively high volatility.



Hence, a comparison between daily Open-to-Close volatility of stocks and that of index can be drawn as:

1. A stock with high 1st PC score has the same high volatility period as the NSEI index.
2. A stock with high 2nd PC score has the same low volatility period (from 03/29/2017 to 04/28/2017) as the NSEI index.
3. A stock with both low 1st and 2nd PC scores have low volatility across the whole period.

Functional PCA is an unsupervised learning method that helps to discover underlying patterns, disguised by the excess number of variables in the feature space. Next, we use the supervised learning tools to study the relationship between index and stock volatilities. For a particular market, the dataset can also be perceived in the following way: there are n trading days in the dataset, which can be treated as n observations; and there are p stocks in the dataset, which can be treated as p variables, each has n observations. In addition, there is one variable with n observations, which is the daily volatility of NSEI index, whose correlation structure with the group of p variables in the dataset is to be explored. This fits the classical context of multiple correlation, which is a measure of the strength of linear relationship between one variable Y and a group of others X, by searching for the linear combination *a* for the variable group X that maximize the simple correlation between the linear combination *a*TX and variable Y. The linear regression model is utilized here, as it can be shown that the square root of the coefficient of multiple determination R2 of the linear regression of Y on X, is the multiple correlation between Y and X. Yet we cannot directly fit the linear regression model, as the dataset has more variables than observations, i.e. p > n. Hence, PCA method should be applied to reduce the dimension from p to a reasonable small number. Combing the PCA method with the linear regression model, we are essentially performing the Principal Component Regression (PCR) analysis. Note that this approach assumes independence between different observations, which may not be true, but we can still apply this model to compute the multiple correlation. The results are summarized as:



Hence, 7 PCs explain over 75% of variation in the X variables (stock volatilities), and over 54% of variation in the Y (index volatility), indicating a moderate multiple correlation.

2.2 Relationship between the S&P 500 Daily Open-to-Close Volatility and Daily Open-To-Close Volatility of Country Indices of India and Singapore.

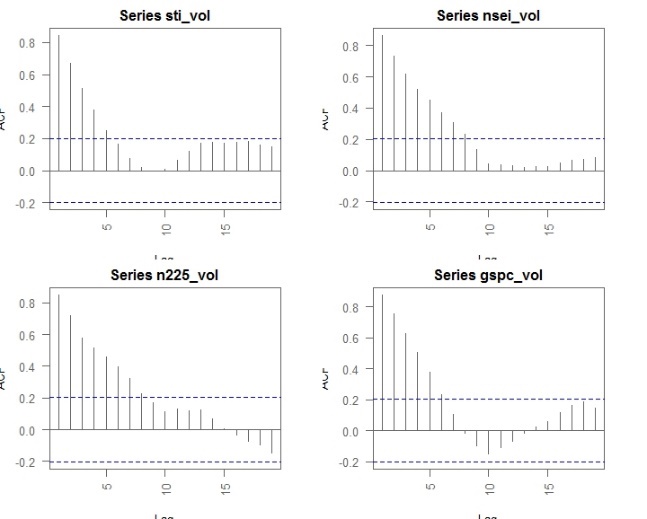
To study the relationships among the daily volatility of country indices, i.e. S&P 500, NSEI index, STI index and Nikkei 225, we first visualize the series by graphs:



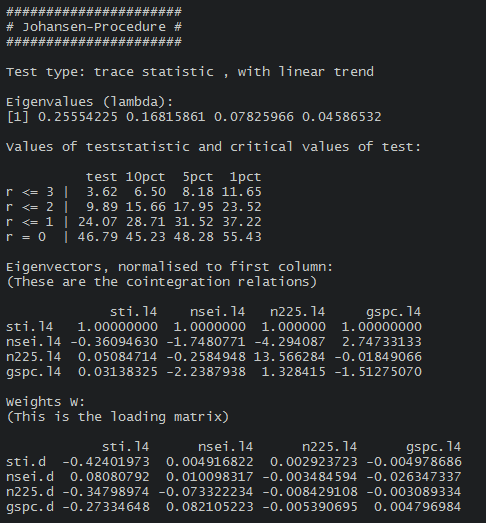


It seems Asian country indices may have certain lead-lag relationship. Statistical methods will be employed to further explore. Multivariate time series can help us to further explore the underlying relationships among these 4 series: first, Johansen procedure can be evoked to figure out whether there exists any cointegration relationships; second, a vector autoregressive (VAR) model can be built to explore any lead-lag relationship among the series, and if there exists any cointegration relationship, a vector error correction model (VECM) should be established; third, a multivariate GARCH model can be built to explore the volatility structure of the multivariate time series system. Note that the VAR model should only be applied to stationary series, e.g. log return series. Yet under cointegration, taking log differencing would lead to a case known as "overdifferencing". Overdifferencing will lead to unit roots in MA matrix polynomial in the VARMA model, causing noninvertibility, and problems in parameter estimation (P431, Analysis of Financial Time Series; Tsay). This explains why the VAR model should be replaced by a VECM under cointegration.

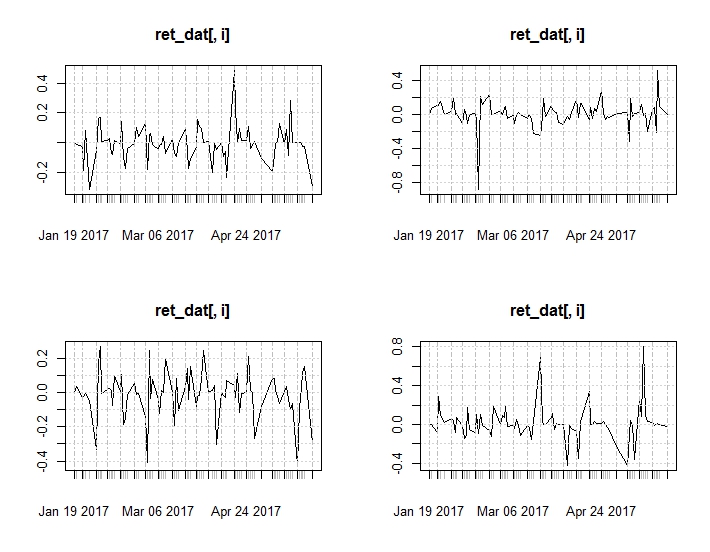
First, visualize the autocorrelation structure, and nonstationarity of these four series is confirmed, since the autocorrelation functions (ACF) decay slowly as the lag increases.



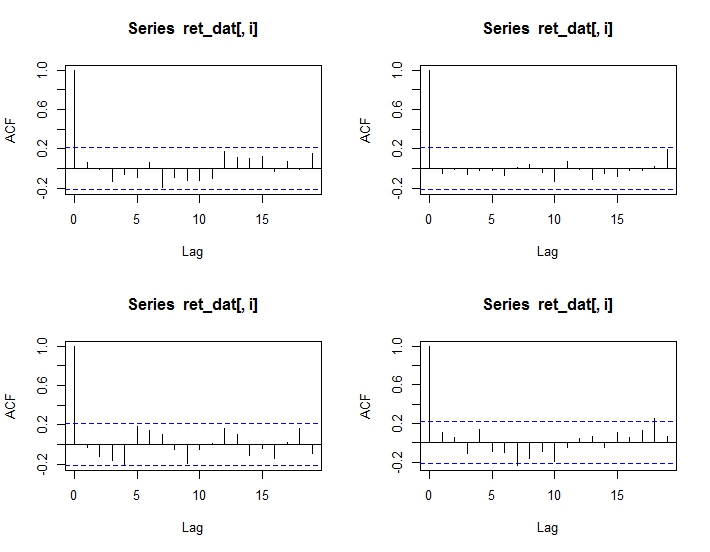
Second, Johansen procedure can hardly confirm any cointegration relationship at 5% level of significance, since the test statistics are consistently below the 5% critical values for the null hypotheses r = 0, r <=1, r <=2, r <= 3, where r denotes the number of cointegration relationships.



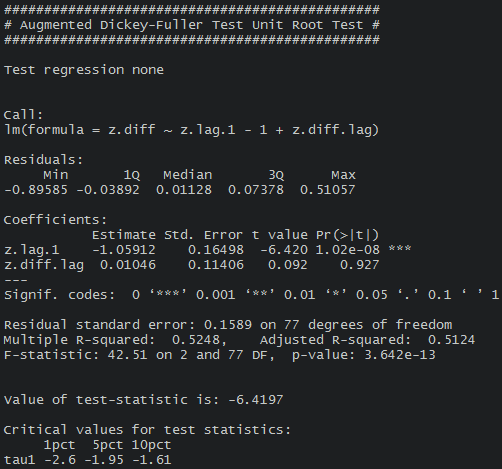
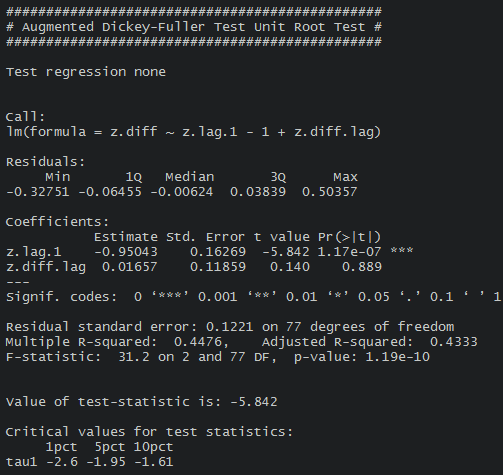
The lack of cointegration relationship in the multivariate time series system suggests that the VAR modeling is free from the potential overdifferencing issue. We thus proceed to extract daily log returns of the daily Open-to-Close volatility series for the 4 indices. The log return series are shown in the following figures (STI (upper left), NSEI (upper right), Nikkei 225(lower left), S&P 500 (lower right))::

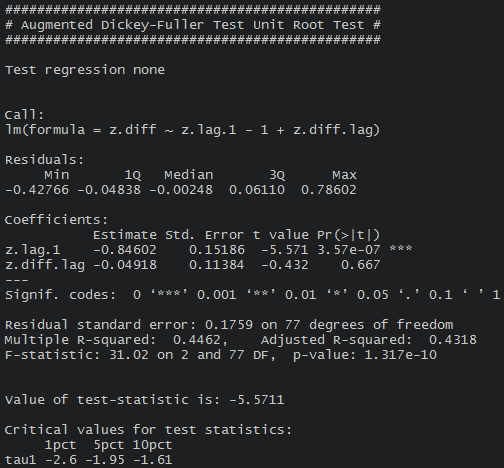
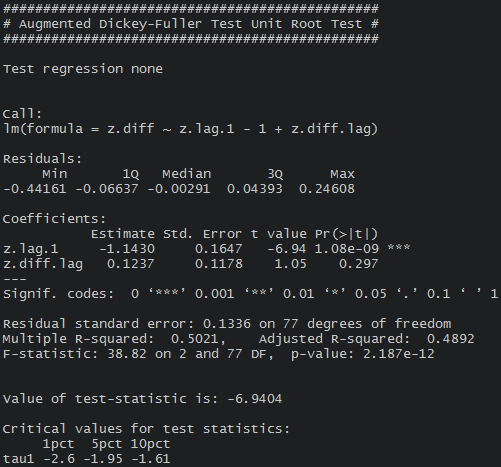


For each index volatility, the log return fluctuates significantly around 0 (the assumed mean), indicating rather week autocorrelation structure, and this can be further confirmed by the following ACF plots:

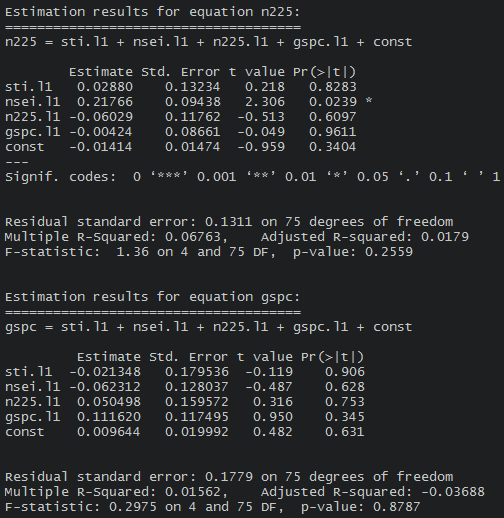
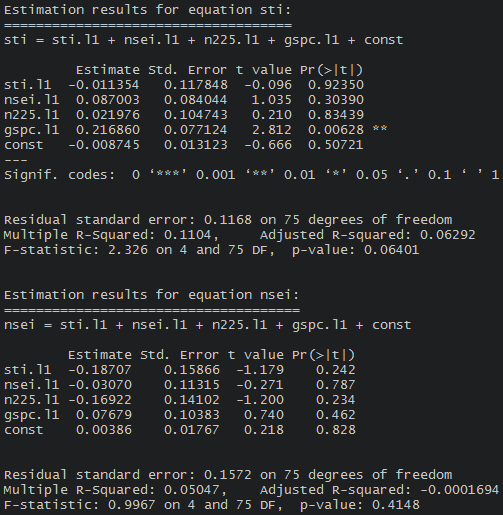


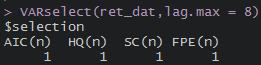
Statistical tests can also be conducted to rigorously test the nonstationarity of the log return series. Augmented Dickey-Fuller (ADF) unit root test is utilized. The results are summarized in the following figures (STI (upper left), NSEI (upper right), Nikkei 225(lower left), S&P 500 (lower right)):





All the test statistics are greater in magnitude than the critical values at 1% level of significance, so the null hypotheses of nonstationarity are all rejected.

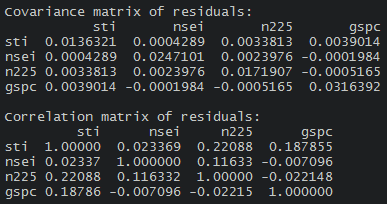




We can then proceed to build a multivariate time series model using the VAR model on these 4 log return series of the daily Open-to-Close volatility. Select the optimal lag using AIC, and lag of order 1 is selected as shown above. Fit the VAR model of order 1, and the results are summarized in above figures, from which the results on index volatilities can be concluded as follows:

1. The STI index is led by S&P 500 index, implying that Singapore market follows closely to the US market.
2. The NSEI index has no serial correlation with any indices including itself, implying that the India market is rather isolated.
3. The Nikkei 225 index is led by the NSEI index, but the reason needs further exploration.
4. The S&P 500 index has no serial correlation with any other indices including itself, which is reasonable, as the US market usually leads others.

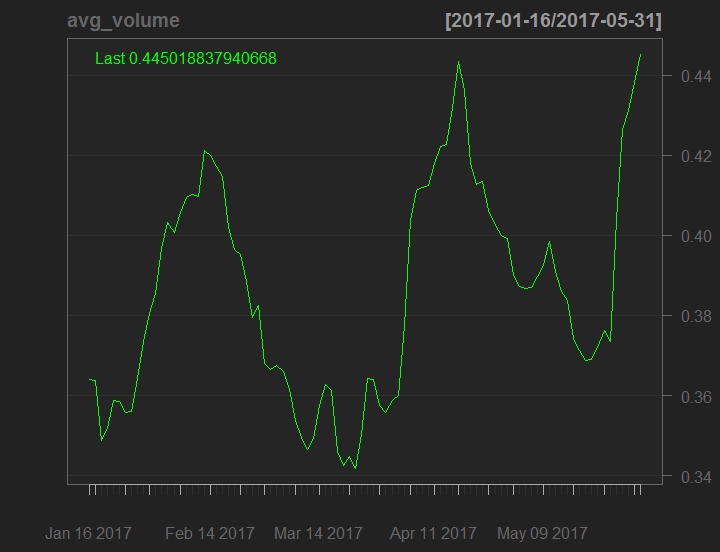
The adequacy of the VAR model can be verified by the covariance and correlation matrix of residuals, as no residual series are significantly correlated with others.



Last but not least, volatility modeling can be employed using multivariate GARCH to study the volatility relationship among these 4 time series of volatility of index. We don't elaborate here.

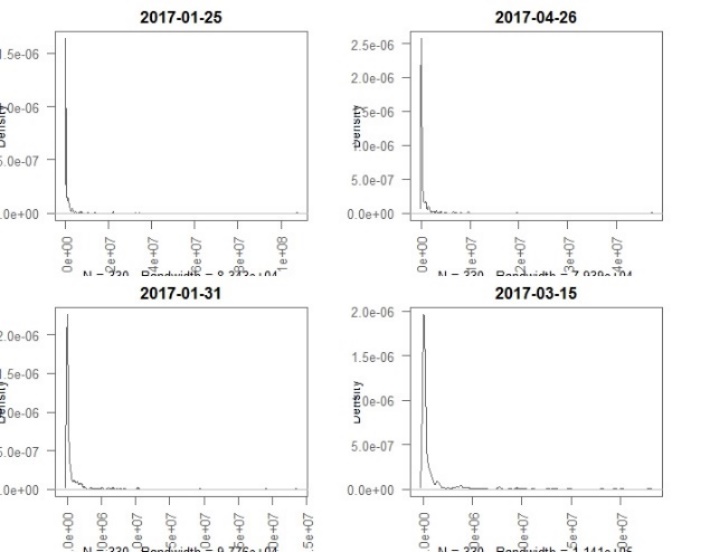
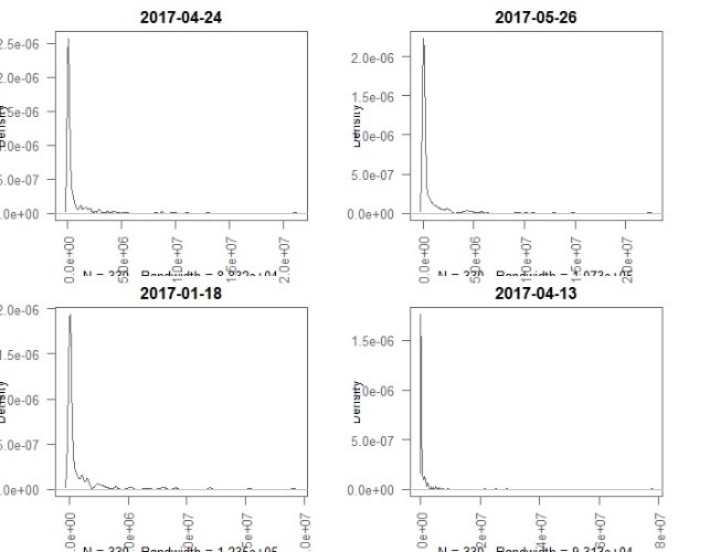
2.3 Daily Volume of the Stocks

The average daily volume of the stocks in the India market are computed and displayed by the following figure:



A side-by-side comparison of the average daily volume of the stocks in the India market and the daily volatility of the NSEI index shows a clear pattern: the distribution of the daily volatility of the NSEI index is closely related to the average daily volume of the stocks in India market. More specifically, we observe: when the average daily volume of the stocks in the India market is rising, the daily volatility of the NSEI index rises, and vice versa. An example can be shown: from 01/16/2017 to 03/14/2017, the average daily volume of the stocks in the India market rises and drops significantly, and this in turn leads to a similar trajectory of the index volatility.

Alternatively, we can examine the daily volume of the stocks in the India market from a cross-sectional perspective. Cross-sectional analysis studies the distribution of stock volumes for a given day. For simplicity, 8 days are randomly selected from the 92 trading days, and we plot the density function estimated by kernel method as a visualization of the volume distribution. The results are depicted as following:



All of the 8 graphs demonstrate rather similar pattern:

1. the distribution of the volume of stocks in the India market on a given day is highly skewed with a large proportion clustering around 0;
2. the distribution is also quite heavy-tailed, as the right tail extends very far away from 0.

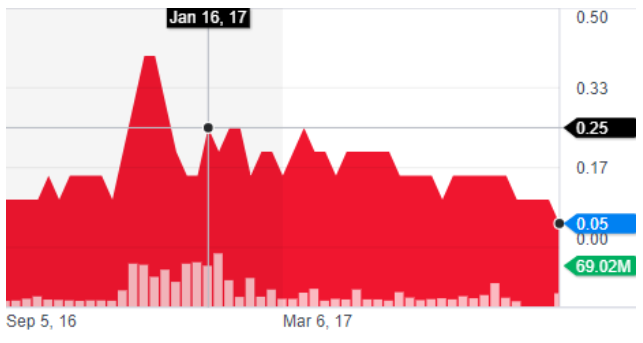
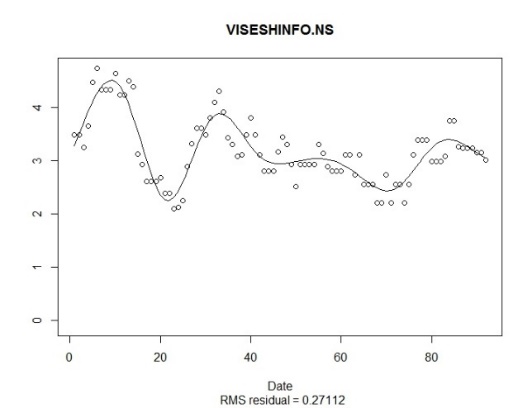
This pattern indicates that the market is binary in a sense that most stocks are rarely traded, while a proportion of stocks in the market are significantly more liquid, representing the blue chips. Moreover, this small portion of stocks are also significantly more volatile in daily returns, and this also explains why most of stocks have 1st and 2nd PC scores clustering around 0.

In light of the highly asymmetric and heavy-tailed distribution of the daily volume of stocks, average daily volume cannot serve as a fairly reliable measure of volume to go through on any given day, as the average cannot represent either side of the binary market. The generalized extreme value (GEV) distribution can be a relatively good approximation to the volume distribution on a given day.

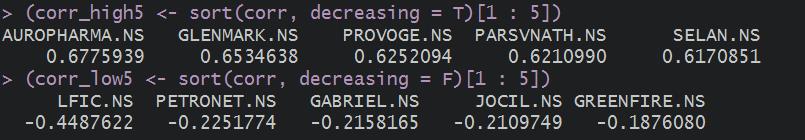
2.4 Daily Volume and Open-to-Close Volatility of the Stocks

In Section 2.3, we have already made a detailed comment on the relationship between the average daily volume of the stocks in the India market and the daily volatility of the NSEI index. In general, the relationship between the average daily volume of one stock and the daily volatility of this stock will be similar. As explained in Section 2.3, the market is binary in terms of volume turnover on a given day, and thus for a given stock it may have the following patterns:

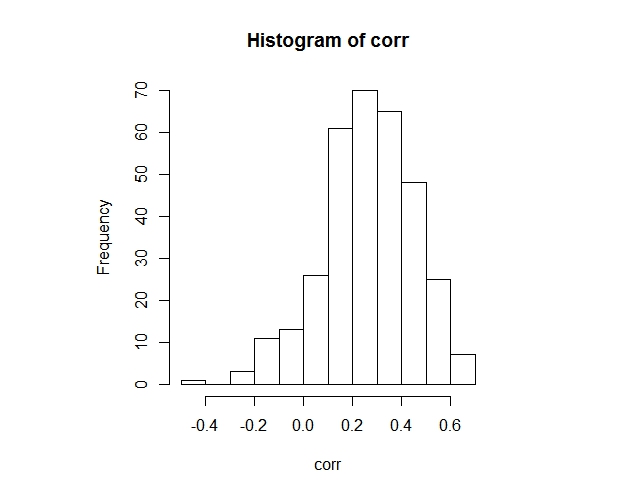
1. If the stock is blue chip and liquid, it has high volume turnover and volatility on a given day. And they will have similar relationship as the average daily volume of the stocks in the India market and the daily volatility of the NSEI index. See Section 2.3 for details.
2. If the stock is not liquid and probably has a quite low price on any given day, then it is likely that both of its volume and volatility are low. However, it may also be possible that its volume is low in value but the volatility is high, as a tiny move in the price would generate a large volatility. One example is VISESHINFO, whose daily volatility is highest among all stocks, but price never exceeds 0.25 in INR for entire period considered.



We compute the simple correlation between the daily volatility and volume turnover for all the stocks to measure the strength of linear relationship between these two quantities for one stock. The results are ranked, and the highest 5 and lowest 5 are demonstrated.



The distribution of the correlations is summarized as the following histogram plot. The proportion of correlations above 0 is 91.5%. Most stocks have moderate (0.2 ~ 0.4) correlation.



The stock AUROPHARMA has the highest correlation, and the high correlation can be easily confirmed by the following graph from Yahoo Finance:



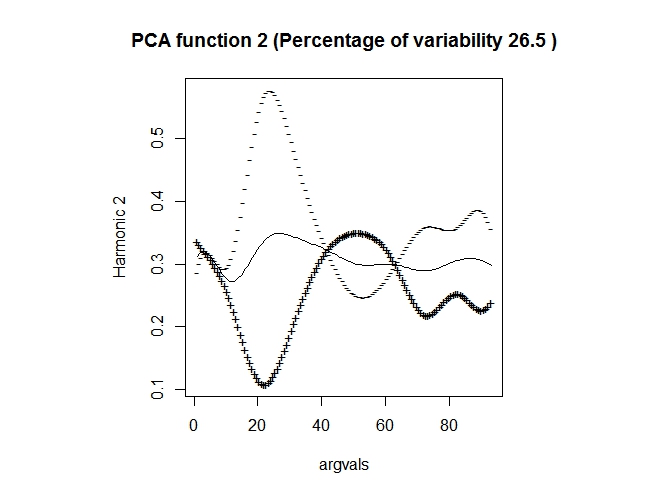
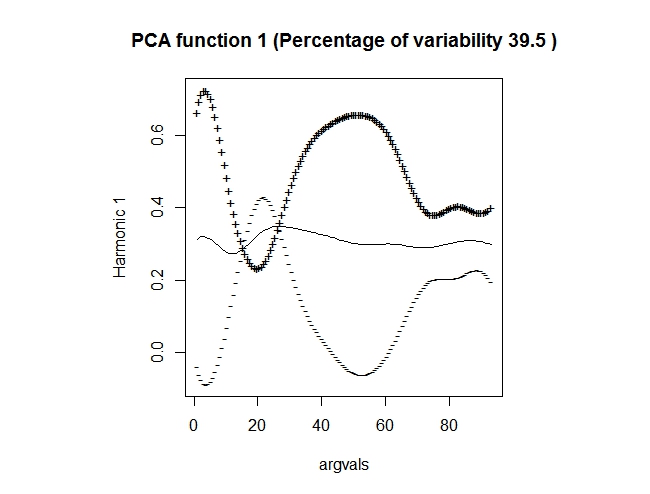
Functional Canonical Correlation Analysis (CCA) can also be employed to explore the relationship between the daily volume turnover and volatility of a stock. To measure the correlation between two group of variables X1 and X2 (volume and volatility for each one of the 92 trading days), CCA searches for the linear combination for both groups that maximizes the correlation between *a*TX1 and *b*TX2. Functional CCA produces canonical function pairs that each assign weights to the (infinitely dimensional) functions X1(t) and X2(t) (here volume and volatility are viewed as continuous functions in terms of time t). By exploring the values of canonical pairs for each stock, one can recover the underlying patterns between the two quantities. However, the interpretation of the results of the functional CCA is more of an art, and thus skipped. Detailed codes have been included in the R script attached.

Section 3 Singapore Market

The discussion on the Singapore market follow the same logic of the India market except Section 2.2. We thus only present the results here without much discussion.

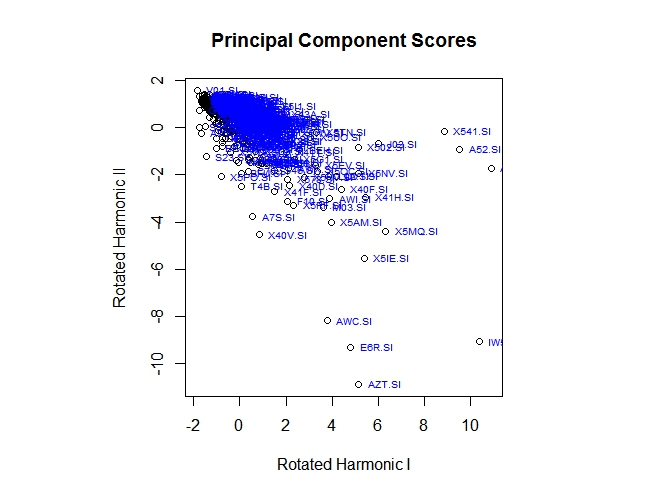
3.1 Daily Open-to-Close Volatility of Stocks and Index

The data comes from 339 stocks in Stock Exchange of Singapore, each with 93 trading records from 01/17/2017 to 05/31/2017. The results of the functional PCA are illustrated as:



The 1st and 2nd PC functions account for over 65% variation in the whole dataset, which is moderate. The 1st PC function explains the variation within periods from 01/17/2017 to 02/10/2017 and from 02/28/2017 to 04/26/2017. The 2nd PC function explains the variation within period from 02/10/2017 to 02/28/2017. The remaining periods are not well explained by either these two PC functions.

The PC scores are plotted. Again, for any stock, a high PC score for 1st /2nd PC indicates large variation within the periods whose variation are mainly explained by 1st /2nd PC functions. Most stocks are clustering around 0, but there are some quite a few deviate away from the origin. The clustering pattern for the Singapore market is similar as the India market.

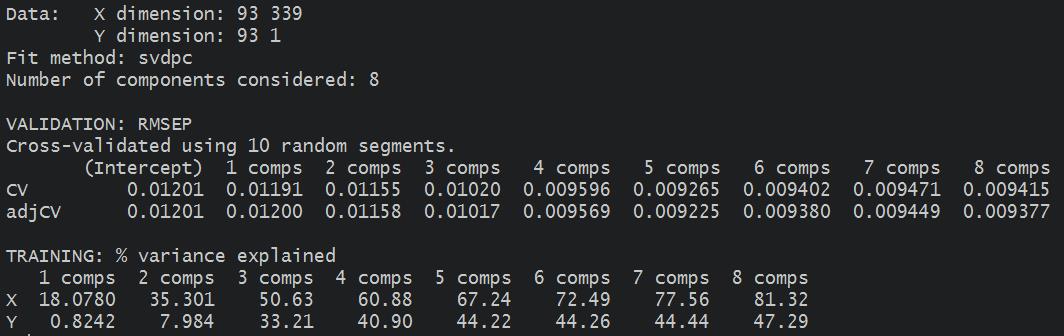




Like the India market, a comparison between daily Open-to-Close volatility of stocks in the Singapore market and that of index can be drawn as:

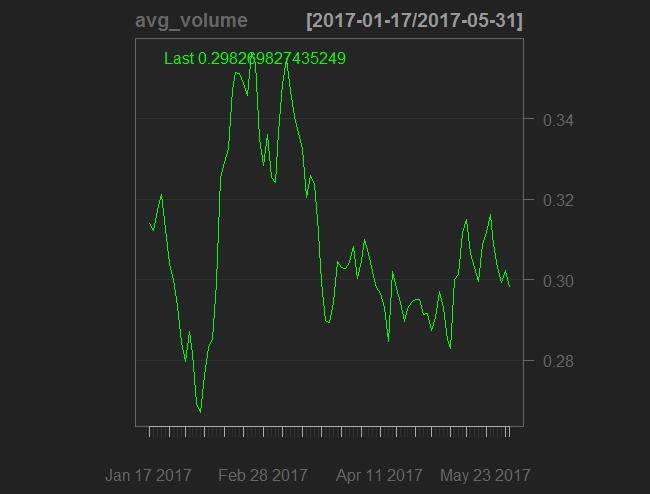
1. A stock with high 1st PC score has the same high volatility period (01/17/2017 to 02/10/2017) or the same low volatility period (02/28/2017 to 04/26/2017) as the NSEI index.
2. A stock with very negative 2nd PC score has the same volatile (02/10/2017 to 02/28/2017 and 04/26/2017 to 05/25/2017) volatility period as the NSEI index. A volatile volatility period refers to the time when the volatility moves up and downs within a band.
3. A stock with both low 1st and 2nd PC scores does not follow the patterns in 1) and 2).

The following displays the PCR analysis. 7 PCs explain over 75% of variation in the X variables (stock volatilities), and over 44.4% of variation in the Y (index volatility), indicating a moderate multiple correlation.



2.3 Daily Volume of the Stocks

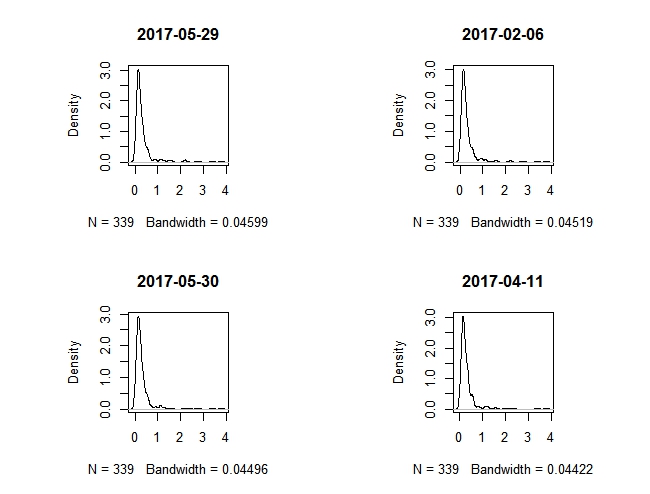
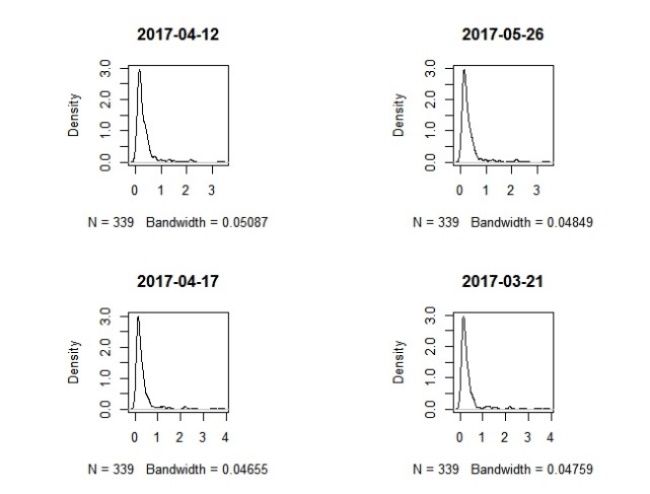
The average daily volume of the stocks in the Singapore market are computed and displayed by the following figure:



A side-by-side comparison of the average daily volume of the stocks in the Singapore market and the daily volatility of the STI index shows a clear pattern: the distribution of the daily volatility of the STI index is closely related to the average daily volume of the stocks in Singapore market. More specifically, we observe:

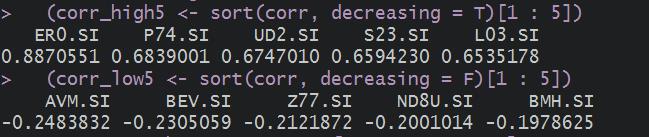
1. when the average daily volume of the stocks in the Singapore market is going up (down), the daily volatility of the STI index goes up (down); for instance, from 02/28/2017 to 04/11/2017, both go down; from 02/01/2017 to 02/15/2017, both go up;
2. though the trend mimics each other, the magnitude does not; sometimes the volume moves by a larger percent, and sometimes the opposite happens.

Alternatively, we can examine the daily volume of the stocks in the Singapore market from a cross-sectional perspective. Again, 8 days are randomly selected from the 93 trading days, and we plot the histogram as a visualization of the volume distribution. Based on the following plot, we can see that the pattern is quite similar to that of the India market, and average daily volume cannot serve as a fairly reliable measure of volume to go through on any given day.

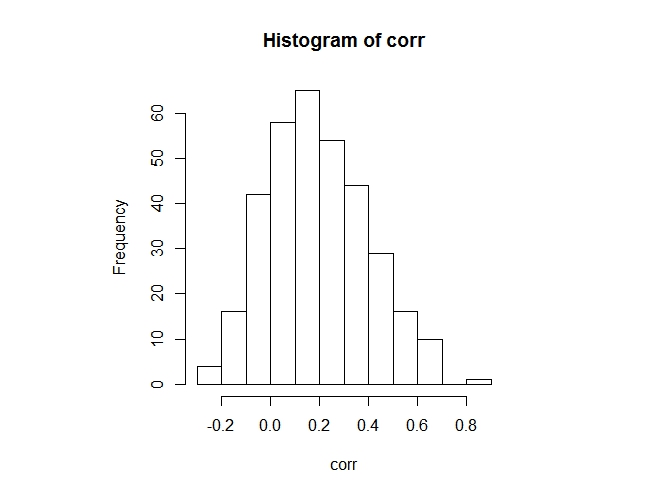


2.4 Daily Volume and Open-to-Close Volatility of the Stocks

The simple correlations between the daily volatility and volume turnover for all the stocks are computed, to measure the strength of linear relationship between these two quantities for one stock. The results are ranked, and the highest 5 and lowest 5 are demonstrated.



The distribution of the correlations is summarized as the following histogram plot. The proportion of correlations above 0 is 81.7%. Most stocks have moderate (0.1 ~ 0.5) correlation.



Section 4 Conclusions

This report presents a detailed analysis on daily Open-to-Close volatility and volume series of stocks and index for the India and Singapore markets. The functional PCA and PCR methods have been utilized to explore the relationship between daily volatility of individual stocks and the market index. Multivariate time series models are employed to study the relationships among the indices of these two markets, together with Nikkei 225 and S&P 500. Further research directions and possible tools have also been discussed.