# FINANCIAL TIME SERIES MODELLING IN FREQUENCY DOMAIN

**WAI MAN TANG**

**M.Phil**

**The Hong Kong Polytechnic University**

**2017**

The Hong Kong Polytechnic University

Department of Applied Mathematics

# FINANCIAL TIME SERIES MODELLING IN FREQUENCY DOMAIN

WAI MAN TANG

A thesis submitted in partial fulfilment of the requirements
for the degree of Master of Philosophy
July 2016

## Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____________________ (Signed)

Tang Wai Man (Name of student)

**To My Family**

# Abstract

In financial time series modelling, one problem is to identify a small number of potentially important factors and incorporate them into a multi-factor model in order to explain the variable in consideration. In this thesis, we propose a new factor search methodology in frequency domain, and select factors based on frequency peak patterns to obtain the final model. This ensures the key patterns in dependent variable be found and suitable factors be selected based on the peaks in common. It performs well even when the number of factors is greater than the sample size. In addition, the frequency domain provides flexibility in dealing with independent variables with different timeframes, and this could be valuable in finance and economic when traditional models usually can handle data in single sampling frequency only.

Using the proposed method, we study three different types of applications. The first is to identify the constituents of an index or a mutual fund.  We demonstrate that our method can identify most of the constituents based on the frequency fingerprints (key patterns) in the variables. The second is to develop multi-factor models based on macroeconomic factors for economic and financial indices. We show that it is important to include factors with different timeframes to achieve better fit. Finally, we study the influential technical analysis indicators that investors might be using in their trading decisions as reflected in the transacted volume, and compared the indicators selected for the same company traded in Hong Kong and Mainland stock exchange markets.

## Acknowledgements

I would like to give my sincere thanks and gratitude to my supervisors, Dr. Yiu Ka Fai, Cedric, for his excellent guidance and valuable comments which are helpful to develop my research work. In the meantime, I would like to acknowledge Prof. Ouliaris for his kind work to verify my source code in frequency domain filter, and acknowledge Prof. Wong Heung for his suggestions on time series analysis. Furthermore, I really appreciate the assistance and advice from teachers and peers in the university. My parents and family should deserve a special note of thanks for the wise counsel and kind words which keep me motivated even at the depressed moment. Lastly, hope this thesis can contribute to future research in the similar topics and for the applications in different fields.

I hope you enjoy your reading.

Carter Tang

15 July, 2016.

## Table of Contents

# 1. Introduction

## 1.1 Background and Motivation

There is an overwhelming amount of economic and financial data which is available easily nowadays; this enhances the advantage to apply quantitative techniques in analyzing useful information from the data. However, when handling a large amount of data with different properties, we need more effort to identify a small number of the suitable factors, and incorporate these useful key factors into a model which is potentially important to explain the variable in consideration.

In fact, economic and financial data is the fundamental element in quantitative analysis in understanding the current state and the outlook of the economy and the financial market. Policy makers rely on data to identify unhealthy trends; and adjust their monetary and fiscal policies on time to prevent or at least minimize the impact of the economic downturn; this can be only achieved by effective use of data. In addition, investors and business sectors also require data to make decision in resource allocation and be alerted to the risk exposure by interpreting the macroeconomic information; for example, investors can manage their capital in different investment opportunities and they should be sensitive to the price and its movement in the financial market, such as yield curve, equity indices and commodity prices. Business sectors should analyze the macroeconomic information to guide them on the level of production, the trend of market demand and costing factors, and prepare for the downside risk of the economy. However, the lack of macroeconomic information or ineffective way in identifying the useful data may induce a skeptical view as people do not have full knowledge on the important factors.

Regression model can relate a dependent variable by other factors which have significant influence. There are different factor selection procedures developed based on regression model which facilitate the search of significant factors by using different criteria. However, traditional methods such as stepwise regression and selection by the correlation matrix may not be the effective way to identify the right model or may not be the reasonable model to explain the dependent variable we considered (Judd et. al., 2011), as they cannot match the useful patterns from the pool of independent variables which can easily exceed a few thousand. Another difficulty is that the model selected may not be the best or even not meaningful if independent variables are highly correlated, as variables cannot be effectively differentiated by the conventional methods due to redundant information in variables. These problems seem more significant in complicated environment with affluent information available nowadays.

Data format may also impose problem for conventional models, data may be collected in different time intervals, for example collected hourly (e.g. financial markets), daily, monthly, quarterly, or even annually. In the literature, consolidating the full information from different time sampling into one model has not been studied properly, and the usual way is to align the data by sampling in the lowest frequency for analysis. This is not an acceptable way anymore in the current dynamic environment, since every piece of information is crucial for the success in modeling.

In this circumstance, the solution should target at the common patterns between the dependent and independent variables. One possible direction worth our attention is analysis via frequency domain, this technique has a long history of successful applications in different areas including natural science, engineering, and economics. Spectral analysis has other properties that give its advantages in applications, namely the ability to determine dominant cycles and their peak amplitudes in the data set. Climate change and detection is a key topic in natural science, Tukey (1966) and Bath (2012) showed a success use of spectral analysis for earthquake detection; and Ghil et. al. (2002) correlated temperature data of the global sea surface with the climate data by spectral analysis. In engineering aspects, signals analyzed by spectral analysis could be used to detect leakage and support the maintenance decision (Mailhes, 2006). It could also be used in molecular imaging by nuclear magnetic resonance (NMR) (Williams and Fleming, 1995). Gresty and Buckwell (1990) showed that mechanism of tremor could be analyzed by spectral components of tremor records.

Spectral analysis also has extensive applications in analyzing economic and financial data, early applications can date back to Crum (1923) which indicated that a cycle was discovered in periodogram by using monthly commercial paper rate in 1874 – 1913; it was a cycle in 40 to 41 months found from the data. However, Wilson (1934) and Greenstein (1935) showed a skeptical view towards the applicability of spectral analysis in long monthly macroeconomic data. Greenstein (1935) found 9.14-year and 10-year periods from the periodogram of the data set, but the two cycles were not likely the exact periodicity in the business failure data from 1867 to 1932. Later, Nerlov (1964) and Granger (1966) investigated the spectrum shape of the macroeconomic data and discussed the control problem and model building in a reality problem by spectral analysis. Their works set the foundation for using the frequency domain method to analyze the dominant cycles in economic time series, and to relate factors that contain similar patterns by respective frequency contents. And Praetz (1979) used spectral analysis techniques to show that cycles were incorporated in the stock price data in the study, and Turhan-Sayan and Sayan (2002) reviewed the potential for representing financial time series by a spectral analysis technique. Wilson and Okunev (1999) analyzed the property and financial assets by the spectral analysis method and found that direct real estate markets showed significant cyclical pattern with the economy, but this pattern less correlated when comparison was based on securitized property markets and financial markets.

## 1.2 Summary of Contributions

In economic and financial modelling, it is advantageous to build up a model by selecting factors in a frequency domain environment, as this is a more flexible and efficient method to select independent variables. In this thesis, we develop a new methodology to obtain multi-factor model in frequency domain, and the main features and significant values are summarized in the following.

- Frequency domain provides flexibility in dealing with independent variables with different timeframes, and this could be valuable in finance and economic research since data collected is usually in different timeframes, while traditional models usually can handle data in single sampling frequency only.
- Matching by unique peaks for frequencies in common between the dependent and the independent factors can ensure that the key patterns are found and suitable factors are selected.
- Our method can perform well even when the number of factors is greater than the sample size, where the conventional methods usually face difficulty in this situation.

To the best of our knowledge, this is the first developed multi-factor model with variables in different timeframes. We study three practical applications in finance to illustrate the proposed method.

- In the first application, we relate portfolio performance to individual equites by a fingerprinting technique. For many financial products, they usually comprise of a collection of tradable assets, but the actual composition is usually unknown. It is advantageous for investors to understand performance attribution earlier for better investment purpose. We can quantitatively relate the price series to a wide pool of tradable assets (factors). As examples, we can identify HSI and its sub-indices successfully; for mutual funds, we can match most of the top 10 constituents disclosed in the regular reports.

- In the second application, we develop a multi-factor model based on macroeconomic data to explain the economic and financial market indices (Faff and Chan, 1998). Spectral analysis technique is deployed to facilitate the selection of factors by matching the dominant cycles in different frequencies between the dependent and the independent variables. Using the fitted multi-factor models, we can analyze the characteristics of the dependent variable by mean of the selected macroeconomic factors, and to reveal the impact of individual factor from the signs and magnitudes of the regression coefficients.
- In the last application, we target at the relation between transacted volume and technical analysis indicators in Hong Kong and Mainland stock exchange markets. We discover the influential indicators that investors might be using in their trading decisions as reflected in the transacted volume, which support the fact that technical indicators might be applying in making buy and sell decisions. In particular, we can separate the selected technical indicators into three main categories in the understanding of different strategies, and found that the spread indicator is a significant factor in both markets.

## 1.3 Organization of the thesis

This thesis has six chapters and is organized as follows:

- Chapter 2 is about the review of the frequency domain technique and the frequency spectrum properties; then we discuss the model formulation and the procedures to perform the factor search in a frequency domain environment.
- Chapter 3 puts a focus on using fingerprinting technique to relate portfolio performance to individual equites. We can quantitatively relate the price series to a wide pool of tradable assets (factors) by using our selection method.
- Chapter 4 discusses the application of our selection method to build up a multi-factor model based on macroeconomic data to explain the economic and financial market indices. Using the fitted multi-factor models, we discuss the characteristics of the dependent variable by mean of the selected macroeconomic factors, and to reveal the impact of individual factor from the signs and magnitudes of the regression coefficients.
- Chapter 5 mentions the application of our selection method to relate transacted volume and technical analysis indicators in Hong Kong and Mainland stock exchange markets. We show the influential indicators that investors might be using in their trading decisions as reflected in the transacted volume, which support the fact that technical indicators might be applying in making buy and sell decisions.
- Chapter 6 summarizes our findings and suggests possible future work.

# 2. Regression Model in the Frequency Domain

When the number of independent variables is relatively small to the sample size, forward, backward and stepwise selection procedures should be effective in searching for the independent variables in the model. However, when the number of factors in consideration is large relative to the sample size, backward procedure may not give meaningful results. For the forward and stepwise methods, they can only search through a subset of models and might not be able to obtain useful and appropriate pattern information from the pool of independent variables. In order to tackle this problem, we propose to relate the selection procedures by taking advantages of the frequency domain in regression model.

In this section, we first provide a brief review on some properties of the Discrete Fourier Transform in section 2.1; then the model formulation and the methodology that search for the best selection of independent variables will be discussed in section 2.2 & 2.3. A numerical example is included in section 2.4 to show that our proposed method is more effective than the tradition methods.

## 2.1 Review on the Properties of Discrete Fourier Transform

In scientific research, we hope to gain information by taking advantage of some stable cycles found in the data; however, we need a quantitative method to capture the dominant cycles found in the data which seems random in human interpretation. Since cycles depend on time, and we can measure and predict the changes. Indeed, these cyclical patterns may even provide insight to relate the phenomenon to other factors which we regard as usefully to understand the mechanism. The concept of spectrum analysis first came from the discovery of light spectrum in physical science, Fourier (1822) constructed the mathematical algorithm to transform data into a frequency spectrum in complex scales, and the two are interchangeable in suitable conditions under the Fourier inversion theorem. Michelson and Stratton (1898) developed a “harmonic analyzer” which related sine and cosine functions to variations in terms of the frequencies and proposed to represent these variations by a periodogram. It is a plot for power density against the frequencies in consideration, and this plot can deliver an instant idea by human interpretation if the sample has dominant peaks in certain frequency ranges; this is the key tool in frequency domain analysis nowadays, researcher can focus their work in the target frequency ranges. Later, Michelson (1913) employed the method to compute periodogram for the sunspot data from Kimura (1913). Einstein (1914) may be regarded as the first publication to fix the frequencies in the spectrum by $\omega' = \frac{n\pi}{T}$, where n is the position in the frequency spectrum, and *T* is the total number of sample in data. Data is transformed into the frequency spectrum based on this definition, and it is similar to the algorithm in Fourier transform. In 1940s, the works for Tukey and Hamming (1949), Bartlett (1950), and Tukey (1949) pushed the subject of spectrum analysis into the modern era. They established an effective computation method in estimating the power spectrum to support useful estimates and identifications of the periodicities in the applications. (Brillinger, 1993)

In general, the time series can be decomposed into cycles with different periods, which can be transformed into frequency domain. We can extract certain frequency ranges that we concerned for further analysis. We define a cycle with consistent oscillation over the period, such as sine or cosine function in a completed period, we can formulate a simple mathematical algorithm to represent this cyclical oscillation as

$$A(\omega_1)\sin(2\pi\omega_1 t + \phi_1), \tag{2.1}$$

where $t = 0, \pm 1, \pm 2, \dots$ , and $\omega$ is the frequency of the periodic oscillation which is equal to the number of completed cycles per unit time. $A(\omega_1)$ is the amplitude of the oscillation in frequency $\omega_1$, it controls the highest and lowest points of this cyclical movement; and $\phi_1$ is the cycle phase which determines the initial position when $t = 0$.

However, equation (2.1) is not a linear model, which makes it difficult in the estimation procedures. We can simplify this equation into

$$A(\omega_1)\sin(2\pi\omega_1 t + \phi_1) = a_1 \cos(2\pi\omega_1 t) + b_1 \sin(2\pi\omega_1 t)\,. \tag{2.2}$$

In equation (2.2), it is a linear model with $a_1 = A(\omega_1)\sin(\phi_1)$, and $b_1 = A(\omega_1)\cos(\phi_1)$. Amplitude can be calculated by $A(\omega_1) = \sqrt{a_1^2 + b_1^2}$, and cycle phase is $\phi_1 = \tan^{-1}(\frac{-b_1}{a_1})$. When the time series is explained by multiple frequencies and amplitudes, we can extend the equation by the summation of the equation (2.2) into

$$\sum_{n}^{c} a_n \cos(2\pi\omega_n t) + b_n \sin(2\pi\omega_n t), \tag{2.3}$$

where $a_n, b_n$ for $n = 1, 2, \ldots, c$ are the coefficients to determine the amplitude and phase of each cycle, while $\omega_n$ is the distinct frequencies, $c$ and is the number of cycles in consideration to explain the time series $x_t$. By using DFT (Discrete Fourier Transform) to convert a time series sample into a frequency domain spectrum, we set the Fourier or fundamental frequencies as $\omega_n = \frac{n}{T}$ for $n = -\frac{T-1}{2}, \ldots, 0, \ldots, \frac{T-1}{2}, \frac{T}{2}$ in the estimation, where $T$ is the total number of data points in the sample. We can define the linear model by the Fourier frequencies as

$$x_t = \frac{1}{T} \sum_{n=-\frac{T-1}{2}}^{T/2} (a_n \sin(\frac{2\pi nt}{T}) + b_n \cos(\frac{2\pi nt}{T})), \qquad for \qquad t = 1, \ldots, T, \tag{2.4}$$

$$S_x(\omega_n) = a_n + ib_n = \sum_{t=1}^{T} x_t \cdot e^{-\frac{i2\pi\omega_n t}{T}}. \tag{2.5}$$

After DFT, the time series $x_t$ is transformed into frequency domain with frequency spectrum $S_x(\omega_n)$ with amplitude $A_x(\omega_n)$ at frequency $\omega_n$ , where $\omega_n$ is in the frequency range $(-\frac{1}{2},\frac{1}{2}]$ , and $S_x(\omega_n)=\overline{S_x(\omega_{-n})}$ . Then, the spectral value and its amplitude can be calculated by the algorithm in equation (2.6)

$$\begin{aligned} &A_x(\omega_0)=a_0 \\ &A_x(\omega_{\frac{T}{2}})=b_{\frac{T}{2}} \\ &A_x(\omega_n)=2(a_n^2+b_n^2)^{\frac{1}{2}} \qquad for \quad n=-\frac{T-1}{2},\dots,0,\dots,\frac{T}{2}. \end{aligned} \tag{2.6}$$

In regression analysis, we consider the spectrum value with frequency $\omega_n$ within the range from $(-\frac{1}{2},\frac{1}{2}]$ ; however, we only need to consider frequency $\omega_n$ within the range from $[0,\frac{1}{2}]$ in the amplitude plot due to symmetry.

### 2.1.1 Illustrative Example of Financial Time Series in Frequency Domain

In this section, we composed a time series similar to the price movement of an equity stock traded in stock exchange market. The time series were only formed by cycles with different periods. The specification for the price movements ($x_t$) is defined in equation (2.7), and the time series plots are displayed in Figure 2.1, and Figure 2.2 is the amplitude plot of its Fourier spectrum. The equation for $x_t$ is

$$x_t = 0.15\sin(\frac{2\pi t}{\lambda_1}+\phi_1)+0.4\sin(\frac{2\pi t}{\lambda_2}+\phi_2)+0.6\sin(\frac{2\pi t}{\lambda_3}+\phi_3)+\sin(\frac{2\pi t}{\lambda_4}+\phi_4)+0.05\varepsilon_t \ , \quad (2.7)$$

where,

$$\lambda_1 = 20, \qquad \phi_1 = \frac{\pi}{6}$$

$$\lambda_2 = 130, \qquad \phi_2 = \frac{\pi}{3}$$

$$\lambda_3 = 400, \qquad \phi_3 = -\frac{3\pi}{8}$$

$$\lambda_4 = 920, \qquad \phi_4 = -\frac{\pi}{7}.$$

$$\varepsilon_t = \mathrm{N}(0,1)$$

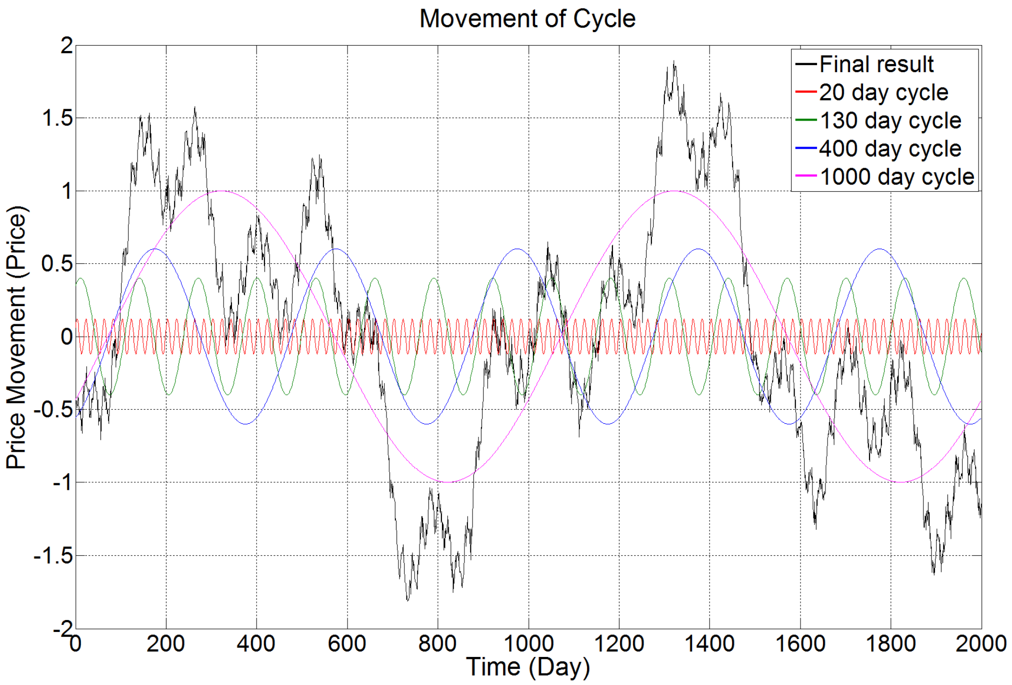

**Figure 2.1 Time series plot of price movement by cycles in different periods**

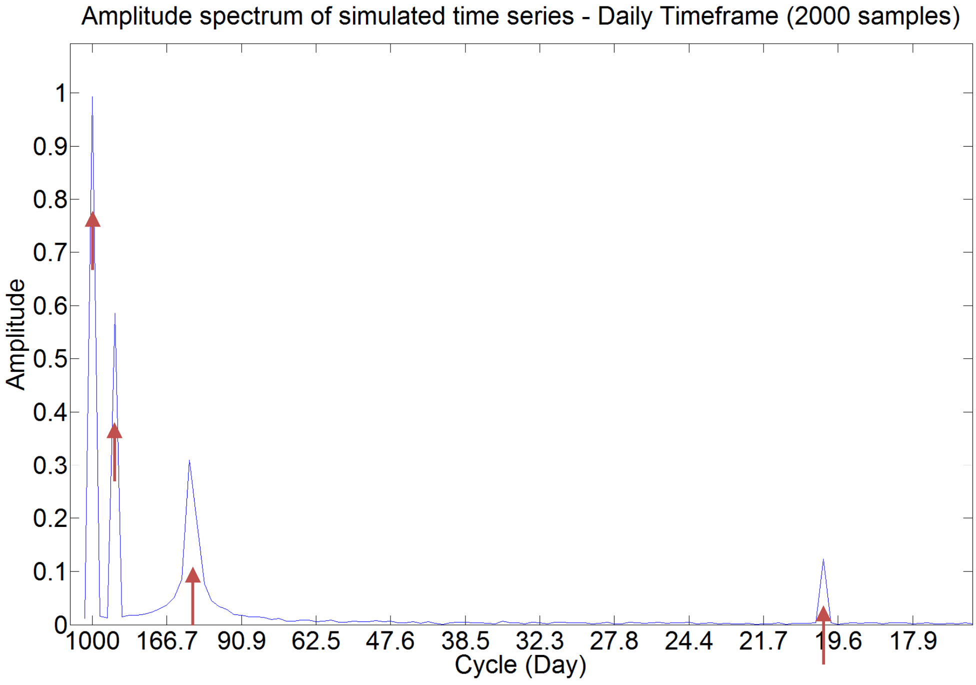

**Figure 2.2 Amplitude of Fourier spectrum for the price movement by different cycles (starting from the cycle period in 16.67 days)**

According to Figure 2.2, we identify four main frequencies, they were 1000, 400, 133 (the closest approximation for cycle 130) and 20 day cycle (as shown by red arrows), If the target cycles do not have the same frequency represented in the Discrete Fourier transform, their peaks can be only revealed by the nearby frequency grids in the Discrete Fourier Transform. Example in Figure 2.3 will review the impact of discretized problem when target frequencies are not the same in the frequency grid.

In this example, we will use three reference cycles in 666.7-day, 25-day, and 5-day periods with 1 unit of amplitude which are exactly matched to the frequency grids of 2000 data points in the example. Then, we review the problem by introducing another three cycles which are deviated from the frequency grids by the three reference cycles and the new target cycles should have their frequencies in between the two nearby frequency grids; they are 583.4 day, 24.85 day and 4.994-day period respectively with 1 unit of amplitude as well. In Figure 2.3, all the three cycles show the peaks same as to the reference cycles which mentioned before; thus, the deviation errors from the actual cycles are 12.5%, 0.6%, and 0.12% respectively. We observe a larger deviation percentage for lower frequency. Moreover, the amplitudes of the deviated cycles are only around 0.65 units which is far below the original one (i.e. 1 unit), while low frequency cycle subjects less impact to the amplitude reduction. The residual, which is not explained by the dominant cycle, is spread among the two sides with its amplitude reduced gradually.

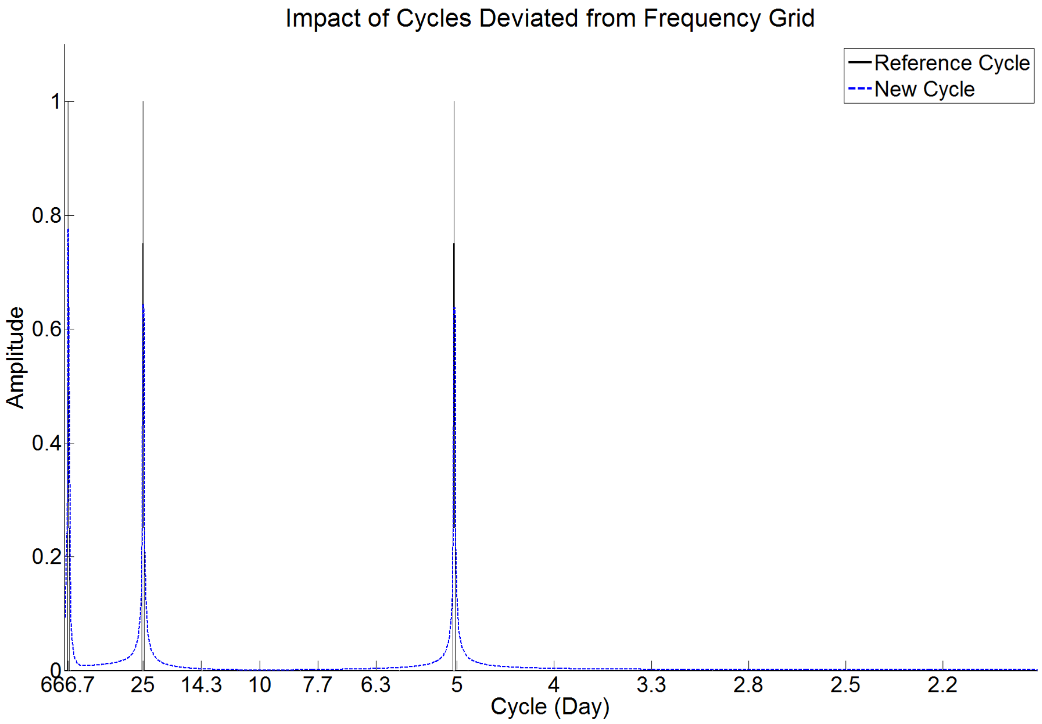

**Figure 2.3 Amplitude plot of three cycles to show the impact of cycles deviated from frequency grid**

Then, we consider different frequency grids by using different sample sizes to review the frequency peaks of the target time series. This should be more advantages to locate the frequency patterns of the target time series as some of the frequency grids can be more effective to identify the peaks. We tried to investigate the outcome of the merged spectrum based on the cycles as shown in Figure 2.3 by using different sample sizes, where we take a sample at every 10 data point starting from 1500 to 2000 points, and the plot of the merged spectrum is shown in Figure 2.4. The peak cycle periods are 583.33 days, 24.844 days, and 5 days, and the deviation errors from the actual cycles are 0.03%, 0.02%, and 0.12% respectively. Amplitude with the cycle period above 10 days can be recovered to the amplitude level similar to the reference cycle. Therefore, this method is effective for most frequencies.

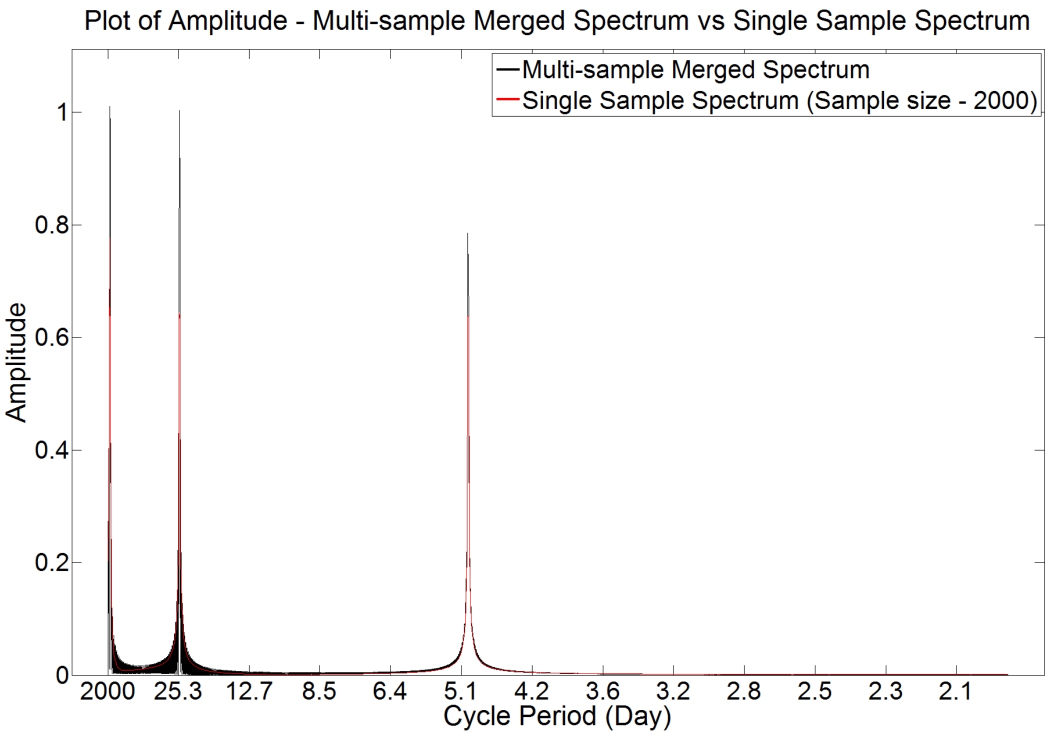

**Figure 2.4 Amplitude plot for multi-sample merged spectrum compared with spectrum in single sample**

## 2.2 Model Formulation for Fingerprinting

We match the independent variables to the dependent variable by using multiple linear regression model in frequency domain environment, and the model in time series is denoted as,

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \varepsilon_t \,,$$

where the notations of this time series regression model are as follows:

$y_t$ , $x_{it}$ are the time series for dependent variable and independent variables

$\varepsilon_t$ is the time series for the error term in regression

$b_i$ is the regression coefficient for the independent variable $x_i$

$r$ is the total number of independent variables in the model

Then, all variables should be transformed by Discrete Fourier Transform defined in (2.5), and the model becomes

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + S_\varepsilon(\omega) \,, \tag{2.8}$$

where the notations of this regression model are as follows:

$S_y(\omega)$ is the frequency spectrum for the dependent variable $y$

$S_{x_i}(\omega)$ is the frequency spectrum for the independent variable $x_i$

$S_\varepsilon(\omega)$ is the regression residual in terms of the frequency spectrum

$\omega$ is the pre-defined frequency range in regression modelling (e.g. $\omega = \{\omega_n = \frac{n}{T}, n = -\frac{T-1}{2}, \cdots, 0, \cdots, \frac{T-1}{2}, \frac{T}{2}\}$ ; where $T$ is sample size)

Regression is performed using the frequency spectrum between dependent and independent variables, and thus the frequencies represented in the spectrum should be the same for all the variables in the regression model.

In the next section, we will provide the model details and testing procedures for the time series modelling. Our methodology, which builds up the model by searching for suitable independent variables from their peaks in frequency domain environment, will be discussed in section 2.4.

## 2.3 Model Formulation for Time Series

Multiple linear regression model in frequency domain environment will be applied in our methodology for searching suitable independent variables to explain the dependent variable, and the model in time series is denoted as,

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \varepsilon_t \,.$$

Then, all variables should be transformed in Discrete Fourier Transform (defined in equation (2.5) of the section 2.1), and the model becomes

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + S_\varepsilon(\omega) \,.$$

Lag variables will be included if regression residual shows autocorrelation, and the final model in the frequency domain is

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + \sum_{p=1}^{P} a_p S_{y,t-p}(\omega) + \sum_{q=1}^{Q} \sum_{i=1}^{r} d_{i,q} S_{x_{i,t-q}}(\omega) + S_{\varepsilon'}(\omega) \,,$$

where $P$ is the largest time lag applied and $a_p$ is lag variable coefficients in dependent variable, and $Q$ is the largest time lag applied and $d_{i,q}$ is lag variable coefficients in independent variables. And the final model in time series can be written as

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \sum_{p=1}^{P} a_p y_{t-p} + \sum_{q=1}^{Q} \sum_{i=1}^{r} d_{i,q} x_{i,t-q} + \varepsilon_t \,.$$

The analysis outcome went through the processes as follows in order to fulfill the requirements in regression:

- Stationary: Frequency domain filter removes non-stationary components in dependent and independent variables if needed. (see section 2.3.1)
- Multiple Timeframes: Set the sample sizes of the independent variables in different timeframes to match their frequencies in the regression spectrums. (see section 2.3.2)
- Error Analysis: Regression residual should be free from autocorrelation by Durbin-Watson test; otherwise, lag variables will be introduced to solve this problem. (see section 2.3.3)
- Residual Normality Check: After the error analysis, we check the residual for normality by Dickey Filler test. If residual is non-normal, we use bootstrapping to ensure the model is significant in F statistics. (see section 2.3.4)

In order to choose a model with suitable number of factors, we search for the models by different *r* in initialization setting. Then, AIC and BIC value are also investigated in the model selection process by the formula below

$$AIC = 2r + T\ln(\sum \left|S_{\varepsilon}(\omega)\right|^{2}) ,$$

$$BIC = r\ln(T) + T\ln(\frac{\sum \left|S_{\varepsilon}(\omega)\right|^{2}}{T}) ,$$

where $T$ is the sample size in regression. Finally, we will discuss our search methodology in section 2.4 with some numerical examples to demonstrate our significant value over the traditional methods.

### 2.3.1 Frequency Domain Filter

In order to ensure the data represented in the frequency spectrum are stationary, we deployed a frequency domain filter after the Discrete Fourier Transform defined in (2.5). Non-stationary time series with the spectral amplitude concentrated in low frequency cycles may affect our factor search results as similar spectral shape appeared across all factors. However, we wish to avoid pre-filtering in time domain environment, as Corbae and Ouliaris (2006) points out that a potential approximation error by using some common approaches to filter out low frequency cycles, such as first difference, averaging or Hodrick-Prescott filter. Our analysis is based on the frequency domain environment, so we tend to approach this issue directly by adjusting the frequency spectrum. Frequency domain filter proposed by Corbae and Ouliaris (2006) can yield a filtered spectrum which is $\sqrt{n}$ consistent in a time series with deterministic and stochastic trends, and this is the merit over the pre-filtering methods. By deploying frequency domain filter, we can effectively remove the non-stationary component, since the approximation error by other filters may distorts the spectral estimate significantly. We determine the frequency range that can effectively remove the non-stationary component but keep most of the unique characteristic in the independent factors. In this section, we will briefly introduce the procedures of frequency domain filter proposed by Corbae and Ouliaris (2006).

$\tilde{x}$ is a non-stationary time series with the property of $\Delta\tilde{x}_t = v_t$, where $v_t$ is a random variable and start from $t = 0$. By taking the Discrete Fourier Transform for $\Delta\tilde{x}_t$, we have

$$\Delta\tilde{x}_t = \tilde{x}_t - \tilde{x}_{t-1} = v_t$$

$$S_{\tilde{x}}(\omega) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{x}_{t-1}e^{i2\pi\omega t} + S_v(\omega) \qquad (\because S_{\tilde{x}}(\omega) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{x}_t e^{i2\pi\omega t})$$

$$S_{\tilde{x}}(\omega) = e^{i2\pi\omega}[\frac{1}{\sqrt{T}}(\sum_{t=1}^{T}\tilde{x}_{t-1}e^{i2\pi\omega(t-1)} - \tilde{x}_0 + \tilde{x}_T e^{i2\pi\omega T}) - \frac{1}{\sqrt{T}}(\tilde{x}_T e^{i2\pi\omega T} - \tilde{x}_0)] + S_v(\omega)$$

$$S_{\tilde{x}}(\omega) = e^{i2\pi\omega}[S_{\tilde{x}}(\omega) - \frac{1}{\sqrt{T}}(\tilde{x}_T e^{i2\pi\omega T} - \tilde{x}_0)] + S_v(\omega).$$

Thus, the Discrete Fourier Transform of $\tilde{x}$ is given by

$$S_{\tilde{x}}(\omega)=\frac{1}{1-e^{i2\pi\omega}}S_v(\omega)-\frac{e^{i2\pi\omega}}{1-e^{i2\pi\omega}}\frac{[\tilde{x}_T-\tilde{x}_0]}{\sqrt{T}}. \tag{2.9}$$

When the discrete Fourier transform of a time series $\{x_t; t=1,\cdots,T\}$ is written as $S_{\tilde{x}}(\omega)=\frac{1}{\sqrt{N}}\sum_{t=1}^{N}\tilde{x}_t e^{i2\pi\omega t}$ , and $\omega=\{\omega_n=\frac{n}{T}, n=-\frac{T-1}{2},\cdots,0,\cdots,\frac{T-1}{2},\frac{T}{2}\}$ are the fundamental frequencies. Then, the second expression of (2.9) can be rewritten using

$$S_{(\frac{t}{T})}(\omega)=\frac{-1}{\sqrt{T}}(\frac{e^{i2\pi\omega}}{1-e^{i2\pi\omega}}).$$

Thus, the second expression is a deterministic trend in terms of the frequency domain multiplied by a random coefficient $\frac{[\tilde{x}_T-\tilde{x}_0]}{\sqrt{T}}$. We can obtain the spectrum with the frequency range $\omega$ after removing the non-stationary component over the spectrum by de-trending in the frequency domain. Random coefficient is estimated by regression of the second expression on the raw spectrum, and the residual of this regression is the unbiased estimate of $\frac{1}{1-e^{i2\pi\omega}}S_v(\omega)$.

The regression equation: $S_{\tilde{x}}(\omega)=bS_{(\frac{t}{T})}(\omega)+S_{\varepsilon}(\omega)$,

where $S_{\varepsilon}(\omega)=\frac{1}{1-e^{i2\pi\omega}}S_v(\omega)$ and this is the output of the frequency domain filter.

## Testing Example in Frequency Domain Filter

We tested the frequency domain filter by comparing it to the spectrum after disregarding the low frequencies. We used two examples to investigate the results of the frequency domain filter; the first example is the cycles from high and low frequencies with accumulated sum of random noises, the objective is to test the ability of frequency domain filter in removing low frequencies and non-stationary noise components. Then, we analyzed another example of high frequency cycle with accumulated sum for noises only to test the effectiveness to remove non-stationary components from random noise. Finally, we tested the SPX index by using frequency domain filter with the different cut-off points to remove the low frequency component in the spectrum.

We composed our examples by the accumulated sum of cycles and random noise component, $F_{1t}$ has cycles in high and low frequencies, and $F_{2t}$ only has a cycle in high frequency. The specifications of the two examples are listed below with sample size *T* as 2000 data points, and Figure 2.5 and 2.7 show the time series plot of the two examples respectively.

$$F_{1t} = F_{1,t-1} + \varepsilon_{1t} = F_{1,t-2} + \varepsilon_{1,t-1} + \varepsilon_{1t} = \sum_{n=1}^{t} \varepsilon_{1,n}$$

$$\varepsilon_{1t} = 0.1\sin(\frac{2\pi t}{100}) + 0.05\sin(\frac{2\pi t}{1200} - \frac{\pi}{6}) + 0.05\sin(\frac{2\pi t}{3600} + \frac{5\pi}{6}) + 0.3\mathbf{N}(0,1)$$

$$F_{1,1} = 0, \quad F_{1,T} = \sum_{n=1}^{T} \varepsilon_{1,n}$$

$$F_{2t} = F_{2,t-1} + \varepsilon_{2t} = F_{2,t-2} + \varepsilon_{2,t-1} + \varepsilon_{2t} = \sum_{n=1}^{t} \varepsilon_{2,n}$$

$$\varepsilon_{2t} = 0.1\sin(\frac{2\pi t}{100}) + 0.3\mathbf{N}(0,1)$$

$$F_{2,1} = 0, \quad F_{2,T} = \sum_{n=1}^{T} \varepsilon_{2,n}$$

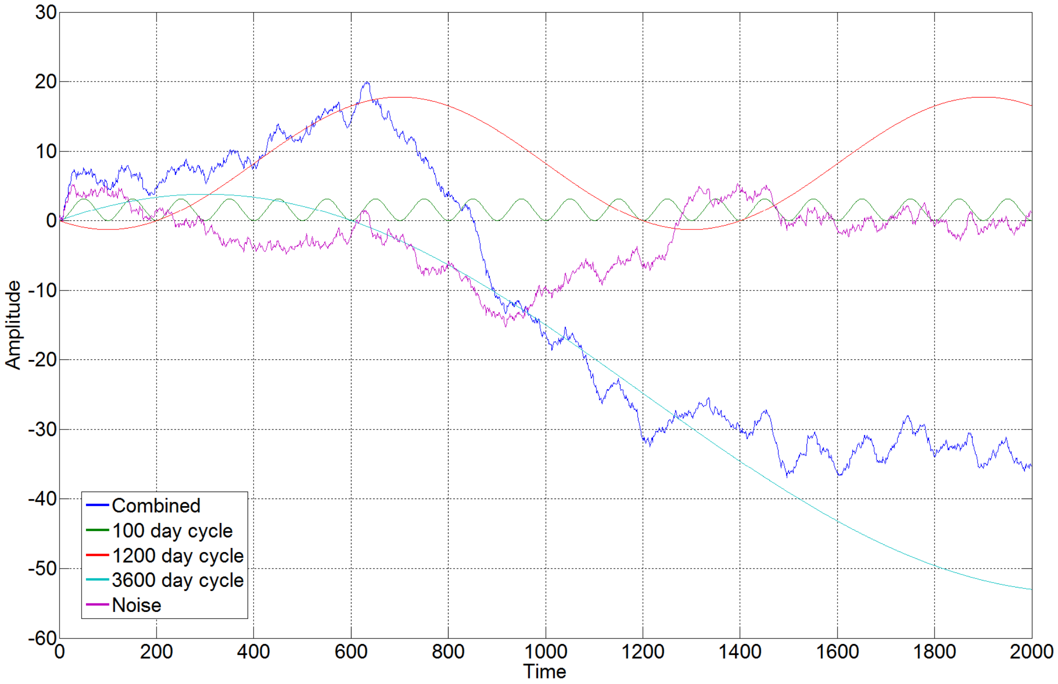

**Figure 2.5 Plot of combined cycles with high & low frequencies and random noises ( $F_{1t}$ )**

Considering the testing sample in 2000 days in Figure 2.5, the spectrum with frequencies smaller than 2000/3 (i.e. 666.67 days per cycle, the fourth spectral value in the spectrum) was disregarded after the frequency domain filter and DFT respectively for comparison. Then, we compared the two outcomes and the raw data in a time series plot (Figure 2.6), and Dickey-Fuller test was performed to test the raw data and filtered outputs if they are stationary or not, and the results are summarized in Table 2.1.

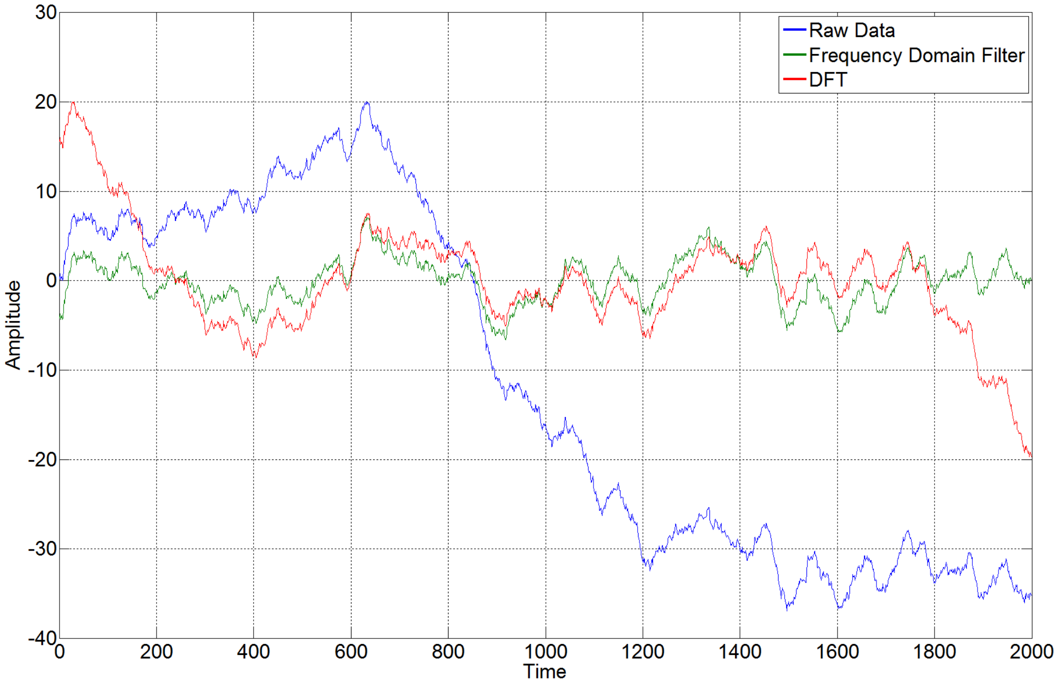

**Figure 2.6 Plot of combined cycles and random noises ( $F_{1t}$ ) after filtering**

**Table 2.1 Dickey-Fuller test of the time series output for $F_{1t}$**

| | **Raw** | **Frequency domain filter (For $\omega$>= 3/2000, i.e <= 666.67 days)** | **DFT (For $\omega$>= 3/2000, i.e <= 666.67 days)** |
|---|---|---|---|
| AR | Non-stationary | Stationary | Non-Stationary |
| p-value | 0.9779 | 0.0036 | 0.5591 |

According to Table 2.1, DFT cannot effectively remove the non-stationary component from the raw data, and distortion is found at the beginning and the end of the output time series. While, the low frequency component can be effectively removed by using frequency domain filter which gives the stable performance over the time series; this is supported by the p-value in Dickey-Fuller test as it rejects the non-stationary hypothesis at strongly significant level. Then, we investigated the high frequency cycle (100 days per cycle) only with accumulated random noises, and Figure 2.7 displays the time series plot for the example. We performed similar comparison in Figure 2.8 after removing the low frequencies, and Table 2.2 contains the Dickey-Fuller test of the raw data input and the filtered outputs in the second example.

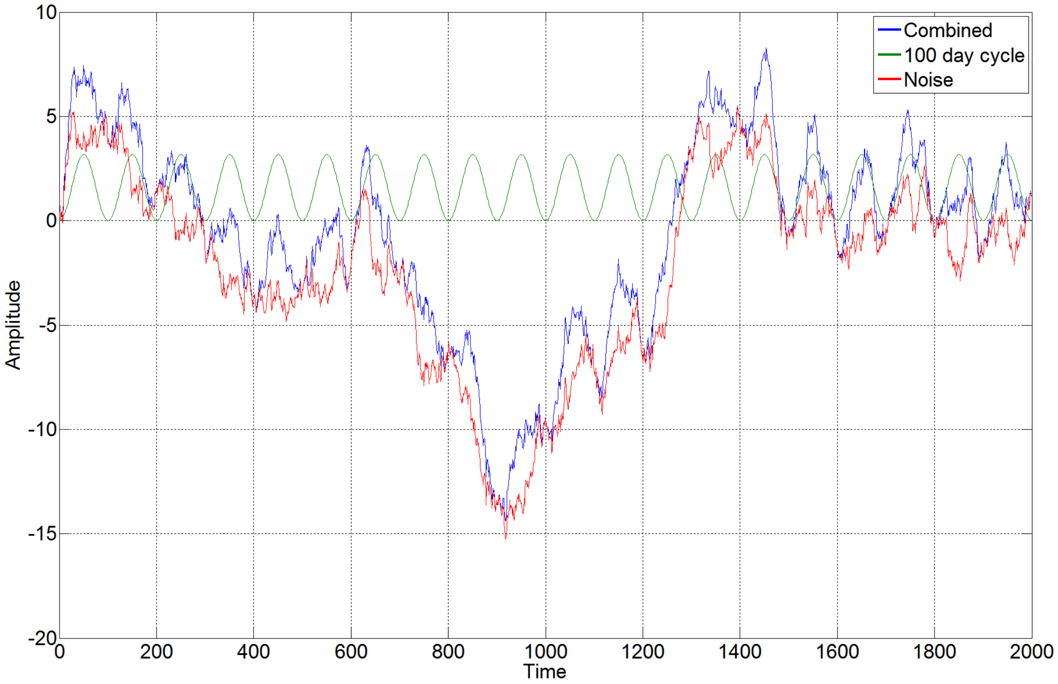

**Figure 2.7 Plot of 100 day cycle and random noises ( $F_{2t}$ )**

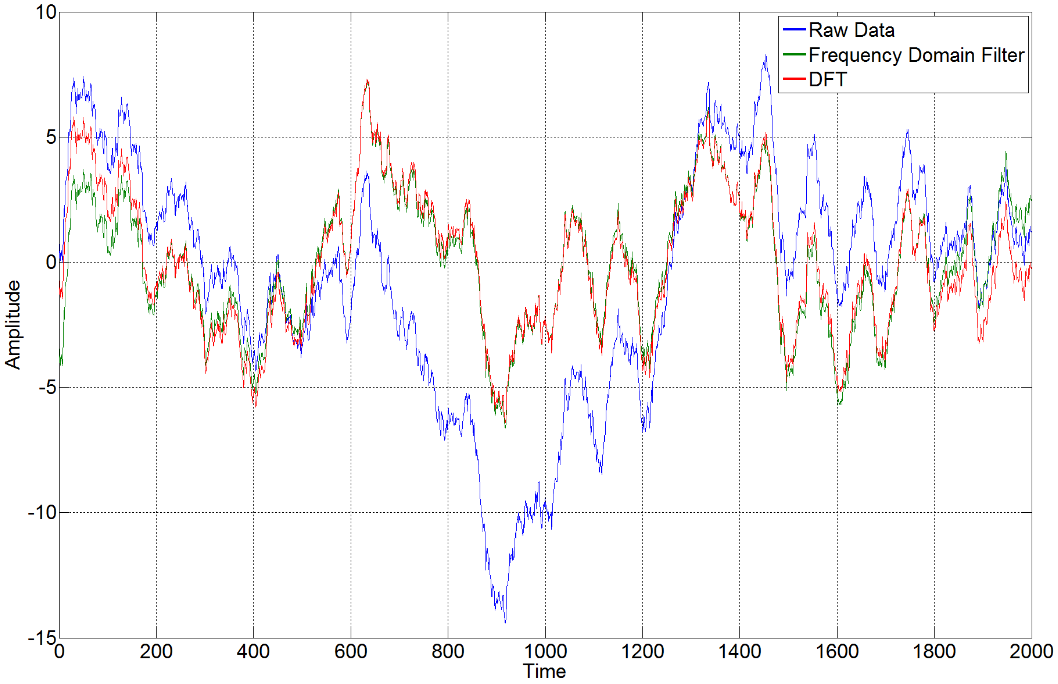

**Figure 2.8 Plot of 100 day cycle and random noises ( $F_{2t}$ ) after filtering**

**Table 2.2 Dickey-Fuller test of the time series output for $F_{2t}$**

| | **Raw** | **Frequency domain filter (For $\omega$>= 3/2000, i.e <= 666.67 days)** | **DFT (For $\omega$>= 3/2000, i.e <= 666.67 days)** |
|---|---|---|---|
| AR | Non-stationary | Stationary | Stationary |
| p-value | 0.1450 | 0.0071 | 0.0127 |

According to Figure 2.8, we have nearly the same output by using frequency domain filter or DFT in this example. Thus, DFT can only apply in removing the non-stationary component when input data without significant low frequency cycles (i.e. the cycle period is greater than the sample period considered). However, we cannot ensure the cycles in the data set, and we strongly advocate a filter should be stable enough for all situations and avoid the output by using DFT in Figure 2.6.

In the next step, we investigated a real example using a sample of S & P 500 index (SPX) from 7 May 2009 to 26 Mar 2013 (1000 trading days). Frequency domain filter and DFT are the two filters to be compared, we filtered the data by three different cut-off frequencies to test the results in removing the non-stationary component. They were frequencies smaller than or equal to 1/1000 (the second frequency grid in the spectrum), 2/1000 (the third frequency grid in the spectrum), and 3/1000 (the fourth frequency grid in the spectrum) respectively in the comparison. The filtered outcomes and raw data are plotted in the same diagram in Figure 2.9 – 11 for different cut-off frequencies in the filter; while, Table 2.3 displays the Dickey-Fuller test for the raw data input and the filtered outputs.

**Table 2.3 Dickey-Fuller test of the filter output by three cut-off frequencies**

| | **Result from Dickey-Fuller Test (p-value)** | | |
|---|---|---|---|
| | **Raw** | **Frequency domain filter** | **DFT** |
| **For $\omega$>= 1/1000, i.e <= 1000 days** | Non-stationary (0.9600) | Stationary (0.001) | Non-stationary (0.1802) |
| **For $\omega$>= 2/1000, i.e <= 500 days** | Non-stationary (0.9600) | Stationary (0.001) | Non-stationary (0.1060) |
| **For $\omega$>= 3/1000, i.e <= 333.33 day** | Non-stationary (0.9600) | Stationary (0.001) | Stationary (0.0323) |

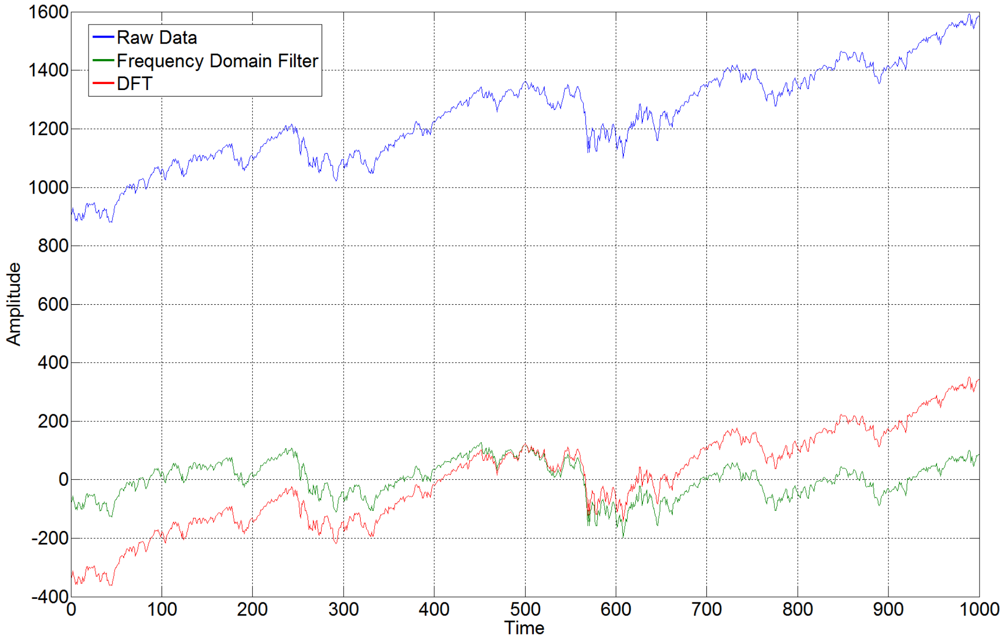

**Figure 2.9 Plot of time series for SPX index with frequencies <= 1/1000 by frequency domain and DFT filters**

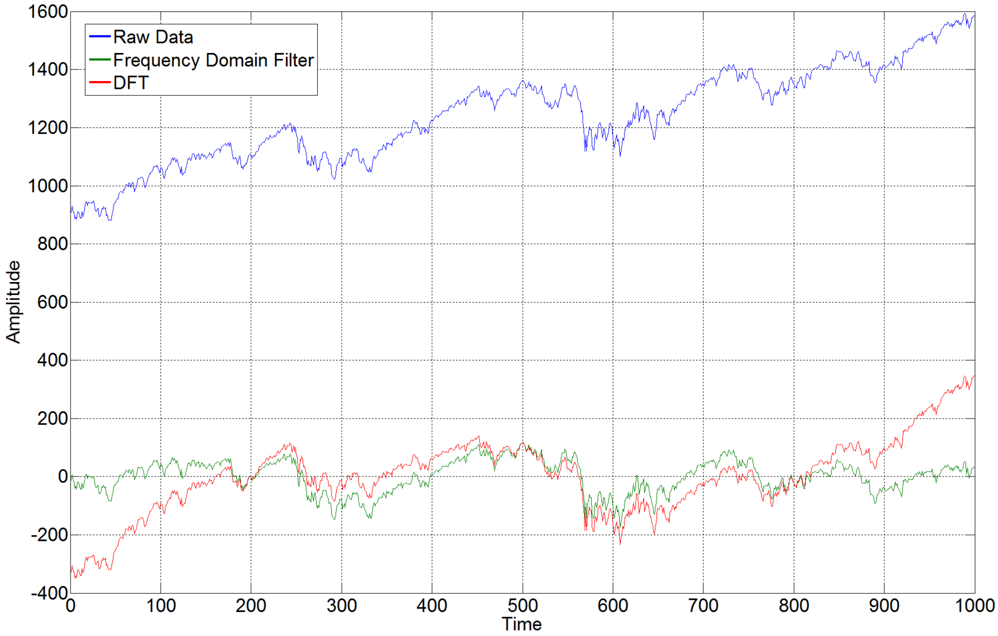

**Figure 2.10 Plot of time series for SPX index with frequencies <= 2/1000 by frequency domain and DFT filters**

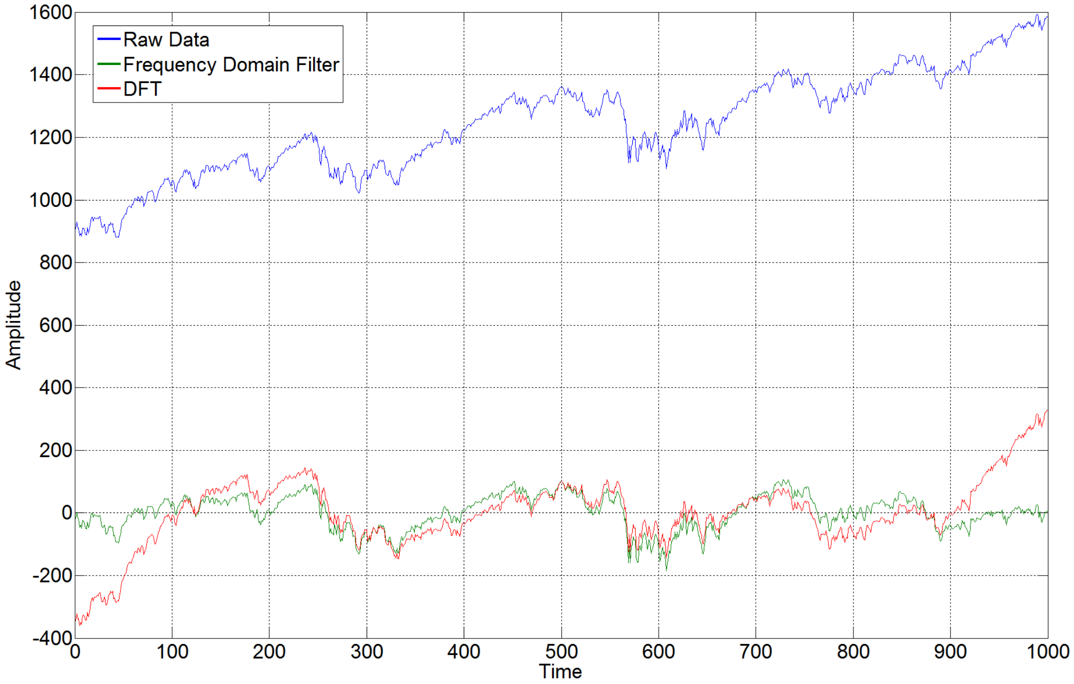

**Figure 2.11 Plot of time series for SPX index with frequencies <= 3/1000 by frequency domain and DFT filters**

According to Table 2.3, frequency domain filter can effectively remove the non-stationary component in the time series by disregarding the zero frequency only (i.e. it includes frequencies >= 1/1000, the second frequency grid in the spectrum). DFT can achieve stationary time series when the cycle period greater than 333.33 days ($\omega$ < 3/1000) is removed, that is the fourth frequency grid in the spectrum. Besides, the time series by DFT in Figure 2.9 – 2.10 have abnormal trending movements at the beginning and the end of the data set. Therefore, we can prove that frequency domain filter can retain more information by using cut-off with lower frequency and achieve stationary time series at the same time when compared with DFT.

Since the spectral amplitudes are mostly reduced in all frequencies after the frequency domain filter, so we multiply a scalar to the filtered spectrum and adjust the two spectrums with the same spectral mean. After this step, we can directly compare the raw spectrum and the filtered spectrum if it still keeps the major pattern of the raw spectrum; the two spectrums after adjustment are plotted in Figure 2.12.

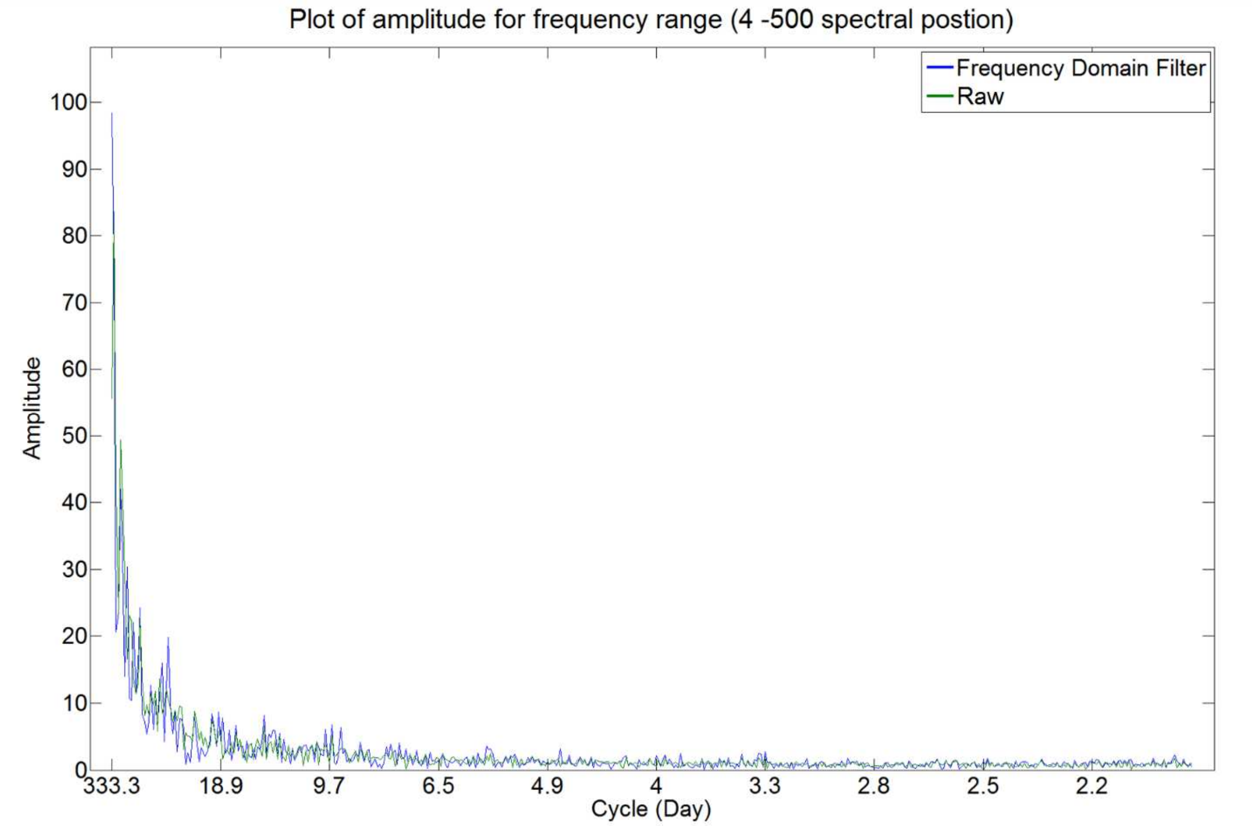

**Figure 2.12 Plot of rescaled amplitude spectrum for SPX index by frequency domain filter vs. raw spectrum**

According to Figure 2.12, the shape of the filtered spectrum with amplitudes resized is nearly the same with the spectrum by raw data, that means the frequency domain filter does not distort the shape of the spectrum.

### 2.3.2 Variables Setting in Multiple Timeframes

In our search algorithm, regression is performed using the frequency spectrum for dependent and independent variables, and thus the frequencies represented in the spectrum should be the same for all the variables in the regression model. Since we analyze factors in different sampling frequencies, variation of data sizes in input is expected. In order to match the frequency grids of variables in different timeframes in the regression spectrum, data length used in the analysis for each sampling rate should be considered carefully. Our regression model will follow the frequency grids in daily sampling, and the frequency grids of spectrums in other sampling rate will follow the sample size defined as below algorithm.

Daily Variable: The transformed data should be represented by cycles with period from 2 days per cycle to *T* days per cycle (i.e. frequency $\omega_n = \frac{n}{T}$), where *T* refers to sample size, and *n = 1 to T/2* is the grid position of the frequency spectrum.

Weekly Variable: There are five trading days in a week normally in Hong Kong, US, and China markets, and *T*/5 sample points in weekly sampling were included in the analysis. The spectrum represents cycles with period from roughly 2 weeks ($\frac{T/5}{T/10-1} \times 5 = \frac{10T}{T-10} = 10 + \frac{100}{T-10} \approx 10$ trading days) per cycle to *T*/5 weeks per cycle.

Monthly Variable: There are 52 weeks for a year (i.e. 12 months); thus, $\frac{12T}{5\times 52}$ sample points (nearest to integer) in monthly sampling will be used in the analysis. The spectrum represents cycles with period from roughly 2 months

(i.e. $\frac{\frac{12T}{5\times 52}}{\frac{12T/2}{5\times 52}-1} \times \frac{5\times 52}{12} = \frac{130T}{3T-130} = \frac{130}{3} + \frac{130(130/3)}{3T-130} \approx 43$ trading days) per cycle to $\frac{12T}{5\times 52}$ months per cycle.

Next we will test our method in multiple timeframes for regression analysis.

## Regression Analysis with Inadependent Variables in Multiple Timeframes

We tested the regression model with variables in different sampling intervals based on the method proposed, we denote *t* as daily sampling, and $t^M$ as monthly sampling by including data in every twenty day by daily sampling data. Finally, we compare the regression results for the data generated by two sampling intervals according to the model below.

Original model: $Y_t = X_t + \varepsilon_t \qquad \varepsilon_t = \mathbf{N}(0, 0.1)$

where, $X_t = 0.25\sin(\frac{2\pi t}{250} - \frac{\pi}{6}) + \sin(\frac{2\pi t}{1000} + \frac{3\pi}{8})$

Daily sampling: $t = \{0,1,2,....,1999\}$

Monthly sampling: $t^M = \{0, 20, 40, 60, ...., 1980\}$

After matching the frequencies for the monthly sampling data into the daily timeframe, we plot the time series of the two sets of data in Figure 2.13 and their frequency spectrum in Figure 2.14; it shows that the two sets of plots contain the same characteristic and information even in terms of two different sampling intervals. Table 2.4 shows the regression results for regression by $X_t$ (daily sampling) onto $Y_t$ (daily sampling), $X_{t^M}$ (monthly sampling) onto $Y_{t^M}$ (monthly sampling), and $X_{t^M}$ onto $Y_t$ after all variables' timeframes are matched in the daily sampling. It proves that nearly the same magnitude for the regression coefficient and R$^2$ in the original model and the model with all variables matched in the daily timeframe; residual tests also give the same conclusion in terms of normality, stationary and autocorrelation. From the regression results, the same coefficients by using the data in terms of time domain or frequency domain; thus, we can ensure that the same results obtained by time domain data or transformed data into the frequency domain.

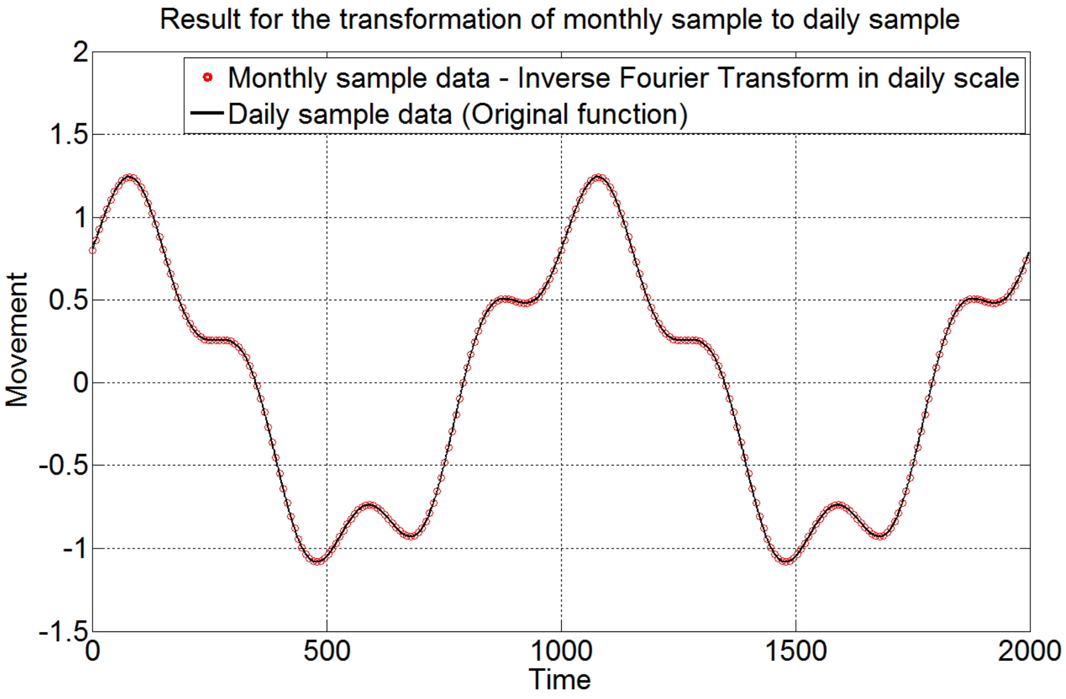

**Figure 2.13 Time series plot for $X_t$ & $X_{t^M}$ with variables matched in the daily timeframe**

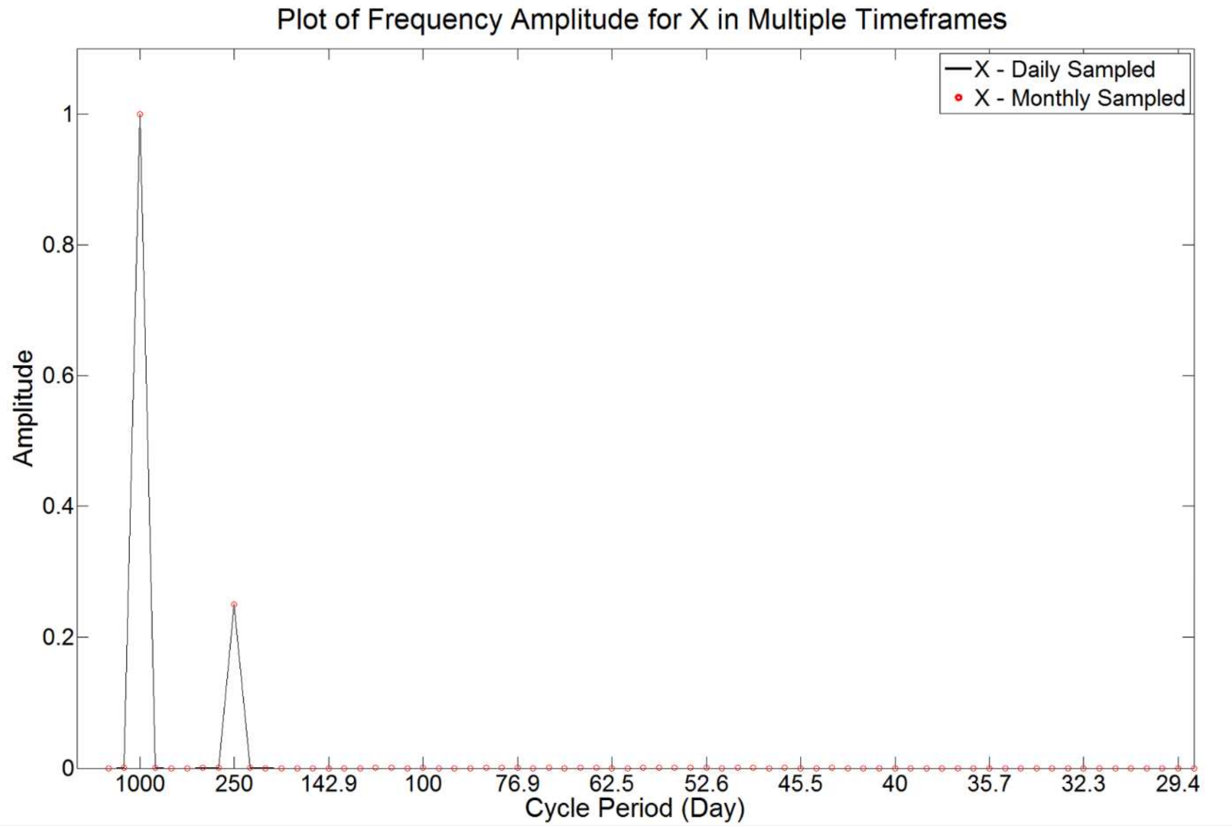

**Figure 2.14 Plot of amplitude for *X* in multiple timeframes**

**Table 2.4 Regression by dependent variables *X* & residual analysis**

| | $R^2$ | Coefficients | Normality (Null is stationary) | Stationary (Null is non-stationary) | Durbin-Watson value | Durbin-Watson p-value (Null has auto-correlation) | Maximum ACF value over 20 lag |
|---|---|---|---|---|---|---|---|
| Regression in daily sampling for all variables | 0.9815 | 0.9994 | 0.3118 | 0.0010 | 2.0451 | 0.3133 | 0.0439 |
| Regression in monthly sampling for all variables | 0.9810 | 0.9899 | 0.5000 | 0.0010 | 2.2929 | 0.1666 | 0.1805 |
| Regression in variables matched in the daily timeframe (In time domain) | 0.9815 | 0.9994 | 0.3118 | 0.0010 | 2.0451 | 0.3133 | 0.0439 |
| Regression in variables matched in the daily timeframe (In frequency domain) | 0.9815 | 0.9994 | 0.3118 | 0.0010 | 2.0451 | 0.3133 | 0.0439 |

### 2.3.3 Error Analysis in Frequency Domain

Based on the model in our proposed search method, regression residual spectrum $S_{\varepsilon}(\omega)$ will be transformed back to time domain ($\varepsilon_t$) for residual analysis via Inverse Discrete Fourier Transform. The formula are

$$S_y(\omega) = bS_X(\omega) + S_{\varepsilon}(\omega)\,,$$

$$\varepsilon_t = \frac{1}{N} \sum_{n=-\frac{T-1}{2}}^{T/2} S_{\varepsilon}(\omega_n) \cdot e^{\frac{i2\pi\omega_n t}{T}}\ .$$

In order to fulfill the assumption in regression analysis, we checked autocorrelation in the residual by using Durbin-Watson test. If autocorrelation exists in the residual, we will employ additional variables with lag in dependent and independent variables to the model. This can remove the autocorrelation property of the residual, and the final model should be

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + \sum_{p=1}^{P} a_p S_{y,t-p}(\omega) + \sum_{q=1}^{Q} \sum_{i=1}^{r} d_{i,q} S_{x_{i,t-q}}(\omega) + S_{\varepsilon'}(\omega)\,. \tag{2.10}$$

We tried to use less number of lag variables in the model to pass the Durbin-Watson test, different combinations of lag variables were tried and evaluated by ACF and PACF.

Then, we prepared one example below to demonstrate that the lag variables from dependent and independent variables can remove autocorrelation in the residual after regression.

## Regression Analysis for Lag Variables

We performed regression with lag variables to test the effectiveness in removing autocorrelation in residual, we first included residual $a_t$ with AR(1) property into $Y_t$ for our testing example $Y_t = X_t + a_t$ , then we performed regression analysis based on the formula $Y_t = \hat{b}X_t + \varepsilon_t$ .

Table 2.5 summarizes the test results on normality, stationary, and autocorrelation tests in the residual time series. The original model has a normal and stationary residual, but it is auto-correlated according to Durbin-Watson test, and the lag variable in *Y* can effectively remove the autocorrelation in the original model, as Durbin-Watson test cannot reject the null hypothesis, and we have the same result even the insignificant lag variable ( $X_{t-1}$) removed.

### Table 2.5 Residual test result summary

| | $R^2$ | Normality (Null is stationary ) | Stationary (Null is non-stationary ) | Durbin-Watson value | Durbin-Watson p-value (Null has auto-correlation ) | Maximum ACF value over 20 lag | Coefficients | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | $X$ | $Y_{t-1}$ | $X_{t-1}$ |
| Regression without lag variables | 0.9763 | 0.5000 | 0.0010 | 1.0525 | 0.0000 | 0.4703 | 0.9991 | | |
| Regression with lag variables | 0.9816 | 0.2896 | 0.0010 | 1.9872 | 0.7387 | 0.0442 | 1.0965 | 0.4719 | -0.5688 |
| Regression with significant lag variable | 0.9815 | 0.3928 | 0.0010 | 1.9819 | 0.6662 | 0.0428 | 0.5291 | 0.4704 | |

### 2.3.4 Normality Check

We also checked the normality in residual for the final model which should not contain autocorrelation in the residual. If the residual is non-normal, testing procedures for the significance of the model will not be valid anymore, and thus we use bootstrapping technique to get the distribution of the F-test value in the regression model, we regard the model should be valid in the test result at over 99% significant. Bootstrapping is a technique for estimating sampling distribution and accuracy of the estimators; this is a useful tool in estimating the confidence intervals of the estimators by using a random sampling method with replacement when the distribution of the estimators are unknown. In our case, we take the approach of case resampling in following steps (Hesterberg et al., 2005):

1. Arrange dependent and independent variables as one data vector by each observation, such as $\mathbf{z}_t \equiv [y_t, x_{t,1}, ...., x_{t,r}]$.
2. It has *T* observations in total as $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T$, and they can be resampled with replacement to produce the regression data with *T* observations randomly
3. Perform regression analysis as defined in equation (2.10) to obtain the regression coefficients
4. Calculate the F-test value from the regression result by dividing the explained variance to the unexplained variance (variance of the error term in the model), that is

$$F = \frac{(\sum_{t=1}^{T}(\sum_{i=1}^{r} b_i x_{i,t} + \sum_{p=1}^{P} a_p y_{t-p} + \sum_{q=1}^{Q}\sum_{i=1}^{r} d_{i,q} x_{i,t-q})^2)/(r+P+Q-1)}{(\sum_{t=1}^{T} \varepsilon_t^2)/(T-r-P-Q)},$$

   where $P, Q$ is the number of lag variable included for dependent & independent in the model respectively.
5. Repeat step 2 & 4, and run 10,000 sets of F-test value by the bootstrap resampling
6. Take 1 percentile of the F-test value distribution based on the 10,000 bootstrapped sets.

## 2.4 Proposed Methodology

The proposed search method is based on the peak patterns in the spectrum to select suitable factors in explaining the dependent variable. It has four main parts, and they are initialization, initial factor selection, factor replacement and process termination. Table 2.6 shows the process flow of our search method, and we will discuss each part in details.

| | |
|---|---|
| Initialization | Step 1: Define required parameters<br>Step 2: Transform variables into frequency domain<br>Step 3: Check the amplitude peaks for each variable spectrum |
| Initial Factor Selection | Step 4: Based on a performance indicator by the peaks in the dependent variable spectrum for model selection, add one independent variable with the highest score to the model in consideration<br>Step 5: Repeat Step 4 until the model has pre-defined number of factors |
| Factor Replacement | Step 6: Rank each independent variable based on a performance indicator by the peaks in the dependent variable spectrum for model selection<br>Step 7: Add one independent variable each time from the ranked list (Step 6) to the model<br>Step 8: Eliminate one of the variables in the model by the factor with minimum change in $R^2$ after removal |
| Termination | Step 9: Repeat Step 6 – 8 if the model is changed; otherwise, repeat Step 7 – 8, and terminated until all independent variables are considered |

**Table 2.6 Flow chart for the peak search procedures**

## Initialization

We first initiated the required parameters used in this model selection methodology

- Step 1: Define the parameters, such as $r$ is the required number of factors in the model, $\mathbf{X}=(x_1,x_2,...,x_N)$ is a matrix containing all the *N* independent variables and $y$ is the dependent variable in time series, and $\omega$ is the frequency spectrum range for regression analysis
- Step 2: Transform variables into the frequency domain spectrum
  Used Discrete Fourier Transform to convert the time series samples of all dependent and independent variables into the frequency domain. $S_y(\omega)$ is the frequency spectrum for dependent variable $y$, and $S_{x_N}(\omega)$ is the frequency spectrum for the *N*-th independent variables.
- Step 3: Locate the amplitude peaks for each variable spectrum
  Identified the frequencies of the spectrum (local) peaks $\mathbf{P_y}=(p_{y,1},p_{y,2}....,p_{y,m_y})$.
  $m_y$ is the total number of the peaks identified for $y$.
  Where $m_y$ is determined by the condition $A_y(p_{y,k})>\mu_{\mathbf{p_y}}+c\sigma_{\mathbf{p_y}}$ for $k=1,...,m_y$.
  $\mu_{\mathbf{p_y}}$ is the mean amplitude of all the local peaks, $\sigma_{\mathbf{p_y}}$ is the standard deviations from the mean amplitude of all the local peaks, and $c$ is a constant to control the number of peaks in the variables.

## Initial Factor Selection

$X^i$ is the model contain $i$ selected independent variables, and factors are added to the model by step 4 & 5 until the model becomes $X^r$ , where $r$ is the predefined number of factors in initialization steps.

- Step 4: Evaluate the independent variables by their total score based on the peak positions of the dependent variable, so that the selected independent variable can fit the peak of the dependent variable. The score is calculated by using the independent variable itself when $i = 0$; if $i \geq 1$, explained spectrum ($S_{\hat{y}_{\tilde{x}}}(\omega)$) of the regression

  $S_y(\omega) = \sum_{j \in X^i} b_j S_j(\omega) + a_{\tilde{x}} S_{\tilde{x}}(\omega) + S_{\varepsilon_{\tilde{x}}}(\omega) = S_{\hat{y}_{\tilde{x}}}(\omega) + S_{\varepsilon_{\tilde{x}}}(\omega), \quad \forall \tilde{x} \in \mathbf{X} \setminus X^i$ is applied in the total score calculation. The score for each independent variable is calculated by $u_{\tilde{x}} = 100\phi_{\tilde{x}} + \rho_{\tilde{x},y}$ ; independent variable with the highest score will be added to the model which becomes $X^{i+1}$. In the total score calculation;

  - $\phi_{\tilde{x}}$ is the percentage of peak frequency matched between the dependent and the independent variable. Note that peaks are considered matched when both variables contain peaks at the same frequency grid points. That is $\alpha_{k,d} = 1$, if $p_{y,k} = p_{\hat{y}_{\tilde{x}},d}$ ; and $\alpha_{k,d} = 0$ , if $p_{y,k} \neq p_{\hat{y}_{\tilde{x}},d}$ , where $k = 1,...,m_y$ and $d = 1,...,m_{\hat{y}_{\tilde{x}}}$ .

    $\mathbf{p}_{\hat{y}_{\tilde{x}}} = (p_{\hat{y}_{\tilde{x}},1}, p_{\hat{y}_{\tilde{x}},2}, ..., p_{\hat{y}_{\tilde{x}},m_{\hat{y}_{\tilde{x}}}})$ is the peak frequency for the explained variable of the model $X^i + \tilde{x}$.

    And the percentage of peak frequency matched is $\phi_{\tilde{x}} = \dfrac{\sum_{d=1}^{m_{\hat{y}_{\tilde{x}}}} \sum_{k=1}^{m_y} \alpha_{k,d}}{m_y}$

  - $\rho_{\hat{y}_{\tilde{x}},y} = \dfrac{\overline{S_{\hat{y}_{\tilde{x}}}(\mathbf{p_y})}^{\mathrm{T}} S_y(\mathbf{p_y})}{\sigma_{S_{\hat{y}_{\tilde{x}}}(\mathbf{p_y})} \sigma_{S_y(\mathbf{p_y})}}$ is correlation between the dependent and the explained variable based on the peak spectrum in dependent variable only.

- Step 5: Repeat step 4 until the model reaches the pre-defined number of factors ($r$).

## Factor Replacement

One factor added to the model, then keep the resulting model with the smallest change in R$^2$ when one of the factor in the model is eliminated.

- Step 6: Rank independent variables for evaluation by their total score based on peak position of the dependent variable. Explained spectrum ($S_{\hat{y}_{\tilde{x}}}(\omega)$) of the regression $S_y(\omega) = \sum_{j \in X^r} b_j S_j(\omega) + a_{\tilde{x}} S_{\tilde{x}}(\omega) + S_{\varepsilon_{\tilde{x}}}(\omega) = S_{\hat{y}_{\tilde{x}}}(\omega) + S_{\varepsilon_{\tilde{x}}}(\omega), \quad \forall \tilde{x} \in \mathbf{X} \setminus X^r$ is applied in the total score calculation. The score for each independent variable is calculated by $u_{\tilde{x}} = 100\phi_{\tilde{x}} + \rho_{\tilde{x},y}$ (refer to step 5 for the calculation of $\phi_{\tilde{x}}$ and $\rho_{\tilde{x},y}$), and independent variables are ranked by their total scores.
- Step 7: Independent variable will be added to the model for factor replacement starting from the one with highest score, then one by one from the ranked list.
- Step 8: Eliminate one independent variable from the model $X^{r+1}$ respectively by regression analysis in equation (3.5).

  $$S_y(\omega) = \sum_{j \in X^{r+1} \setminus \tilde{x}}^{r} b_j S_j(\omega) + S_{\varepsilon_{\tilde{x}}}(\omega), \quad \forall \tilde{x} \in X^{r+1}$$

  (3.5)

  We have $r+1$ regression outcomes from the model $X^{r+1}$. $\Delta R_{\tilde{x}}$ is the change in R$^2$ when factor $\tilde{x}$ being taken out from the model $X^{r+1}$.

  $$\Delta R_{\tilde{x}} = \frac{\sum \left| S_{\varepsilon_{\hat{y}_{X^{r+1}}}}(\omega) \right|^2}{\sum \left| S_y(\omega) \right|^2} - \frac{\sum \left| S_{\varepsilon_{\tilde{x}}}(\omega) \right|^2}{\sum \left| S_y(\omega) \right|^2}, \quad \text{where} \quad S_{\varepsilon_{\hat{y}_{X^{r+1}}}}(\omega) = S_y(\omega) - \sum_{j \in X^{r+1}}^{r+1} b_j S_j(\omega)$$

  Change in R$^2$ is compared for all regression outcomes and $\tilde{x}$ is eliminated from the model $X^{r+1}$ with $\min_j(\Delta R_{\tilde{x}}) \quad \textit{for} \quad j = 1, \ldots, r+1$, and the model becomes $X^r$.

In order to proceed the elimination faster, we only consider those factors for elimination when their drop in R$^2$ to the original model ( $X^r$ ) is lower than 2 times of the minimum R$^2$ change when one of the factors eliminated from the original model. It is because if the drop is great when certain factor is eliminated, we can consider this factor is critical to explain the dependent variable, and usually this factor will not be removed in the replacement process.

## Termination

Stop the process if required conditions fulfilled in step 9

- Step 9: Terminate the peak search process when all independent variables are considered in step 6 but no further change in model (i.e. $X^r = X^{r+1} \backslash \tilde{x}$ , $\tilde{x}$ with $\min_j(\Delta R_{\tilde{x}}) \quad for \quad j = 1,...,r+1$)
    - Condition 1: If no change for the original model in step 8, then next candidate with the highest ranked in step 6 (e.g. $u_{\tilde{x}(2)}$) will go through step 7 – 8 accordingly
    - Condition 2: If the original model changed (i.e. $X^r \neq X^{r+1} \backslash \tilde{x}$ , $\tilde{x}$ with $\min_j(\Delta R_{\tilde{x}}) \quad for \quad j = 1,...,r+1$) in step 8, then proceed step 6 – 8 again

## 2.5 Numerical Example

In this section, we compared the effectiveness of our proposed factor search model to the conventional selection methods, such as forward stepwise regression and lasso regression. In the simulation, we first defined the method to generate dependent variables from all the independent variables available. Factor selection was performed by using different search methods stated in section 2.4.1. Finally, we investigated the success rate in matching the right factors for the dependent variables in different search methods, and we also considered the outcomes when error term is included in the dependent variables. The analysis results are shown and discussed in section 2.4.2.

### 2.5.1 Variables

We selected all equity stocks listed in the HKEx as the independent variables, but excluding the ETFs (e.g. Tracker Fund of Hong Kong – code 2800), for the numerical example. Dependent variables were generated according to the steps as followings:

1. Each independent variable is assigned a number for identification, and $r$ different independent variables were selected randomly by drawing random numbers in uniform distribution $\{a_1, a_2, ..., a_{r-1}, a_r\}$.
2. $r$ uniformly distributed random numbers are generated as the factor coefficient ($\beta_r$), which is ranged from -4 to -0.5, and from 0.5 to 4.
3. Generate the time series for dependent variable ($\hat{y}$) according to specification in point 1 & 2 by the equation
$$\hat{y} = \beta_1 x_{a_1} + \beta_2 x_{a_2} + ... + \beta_{r-1} x_{a_{r-1}} + \beta_r x_{a_r} \text{ .}$$
4. Error term ($\varepsilon$) was generated by a normally distributed time series with zero mean and standard deviation related to $\hat{y}$. The output of the dependent variable should follow the equation,
$$\varepsilon = \mathbf{N}(0, \frac{\sigma_{\Delta\hat{y}}}{100})$$
$$y = \hat{y} + \varepsilon.$$

### 2.5.2 Results and Discussion

In the simulation, we had 1,532 independent factors by 180 data points, then a list of dependent variables with 100 simulation cases in total was generated according to the procedures in section 2.5.1 for testing, and we included 5 constituents for each simulation case as the dependent variable (i.e. $r = 5$). Finally, we compare our proposed method and two other conventional search methods by counting the percentage of correct factors matched over the simulation cases. The details of the search methods in the efficiency test are listed as below, and the test results are summarized in Table 2.7.

- Forward stepwise method is deployed in time domain environment, and AIC is the criteria for factor selection and removal. $r$ variables are first selected, then one factor is added and removed continuously until no further change can be made by the model. This method is indicated as **Method 1** in Table 2.7
- Lasso (Least Absolute Shrinkage and Selection Operator) regression performs the variable selection by introducing penalty function to enhance the prediction accuracy. The Matlab program applies ten-fold cross validation in lasso fit, and the combined best result for the largest lambda with MSE within one standard error of the minimum, and lambda with minimum MSE. This method is indicated as **Method 2** in Table 2.7
- Our proposed peak search method in frequency domain (details in section 2.3) is used, $r$ variables are selected in the model. Regression was performed for cycle period equal to or below 45 day (i.e. starting at 4th position of the spectrum). This method is indicated as **Our Method** in Table 2.7

**Table 2.7 Results comparison for the 100 simulation cases by the three methods**

| | Dependent variable without random errors | Dependent variable with random errors |
|---|---|---|
| Method 1 | • 2.82 correct factors in average per simulation (56.4%)<br>• 21 correct out of 100 cases (21%) | • 3.09 correct factors in average per simulation (61.8%)<br>• 50 correct out of 100 cases (50%) |
| Method 2 | • 3.62 correct factors in average per simulation (72.4%)<br>• 29 correct out of 100 cases (29%) | • 3.62 correct factors in average per simulation (72.4%)<br>• 29 correct out of 100 cases (29%) |
| Our Method | • 5 correct factors in average per simulation (100%)<br>• 100 correct out of 100 cases (100%) | • 4.96 correct factors in average per simulation (99.2%)<br>• 96 correct out of 100 cases (96%) |

In conclusion, our proposed peak search method can effectively identify the correct independent variables from the simulation cases whether they contain error terms or not. According to the result in Table 2.7, the peak search method provides the correct answer for all simulation cases when no error term introduced, and only four cases with one factor missed that represents 99.2% of the factors can be identified. However, the stepwise and the lasso regression have relatively poor results even without error term introduced; it has around 60-70% of the factors can be searched, and with only 30 – 50% of the cases are all correct. The Lasso regression tends to have more stable outcome when compared with the stepwise regression method; while, the stepwise method shows more correct cases when error term introduced, but only slightly improved in the total number of factors identified.

# 3. Frequency Spectrum Fingerprinting

## 3.1 Motivation

Fingerprinting has many applications in engineering. For example, acoustic signal can be summarized into a unique fingerprint by considering its content-based compact signature. This can facilitate the matching of the target audio sample with available information in the audio database, and can be applied in audio identification and audio library or database management, such as songs, melodies, tunes, sound effect. Another technique is video fingerprinting, it identifies video contents, then extract the key characteristics by key frame analysis on color, and motion changes from the contents; unique fingerprint will be created for that video file based on the features extracted. This is an effective way to identify and manage digital video data. These fingerprinting methods are widely used in identification for copyright management, which includes the compliance to copyright law, and supporting licensing business. (Cano et al, 2002)

Similarly, frequency pattern features can be considered as a fingerprint for the individual variable, such as individual stock / constituents in a portfolio, and the frequency pattern features can be composed by the combination of key cycles in the data. Our case shares the same direction with an acoustic fingerprint in peak characteristic summary, where the model should be sensitive enough to the slight variation in the data, but tolerate those insignificant features to be included as the fingerprints.

In finance application, data for price or return time series is known to be related to a pool of tradable assets which is not known. It is important for investors to predict the future performance and the embedded risk exposure if we can scientifically relate the price or return time series to the movements by wide pool of factors, and sometimes it is also important to know the composition from the overall. For example, fund managers compose their own investment portfolios and make changes from time to time, but we seldom have full information regarding their investment categories or components. Some mutual funds may provide the detailed information by quarterly; however, this is just top 10 to 20 holdings in the funds, and this may just contribute less than 40% of total variation of the funds if they are widely diversified. We have the net asset values (NAV) of the funds announced daily, but it is difficult to utilize this information in a direct way since there are multi-securities in a portfolio.

Instead of matching a single factor in database like in the audio signal identification, we need to find a sum of stocks / constituents, which can yield the fund or index performance. Fourier spectrum has linearity additive properties, for any complex numbers $p, q$ , if $y = px_t^{(1)} + qx_t^{(2)}$ , then $S_y(\omega_k) = pS_{x^{(1)}}(\omega_k) + qS_{x^{(2)}}(\omega_k)$ , that means spectrums can be added together and it should be the same when adding their time series accordingly.

With this property, multiple regression can be applied to identify suitable stocks. We used an example in Figure 3.1 to demonstrate the peaks of two factors added (i.e. HSBC, HK & China Gas) should be inherited from their individual factors.

According to the Figure 3.1, the major peaks marked in red circles in the black line come from the corresponding peaks of the individual factors (green circles for HSBC and blue circles for HK & China Gas). The next section will discuss the implementation of our methodology to select corresponding securities based on their peak characteristic in the data.

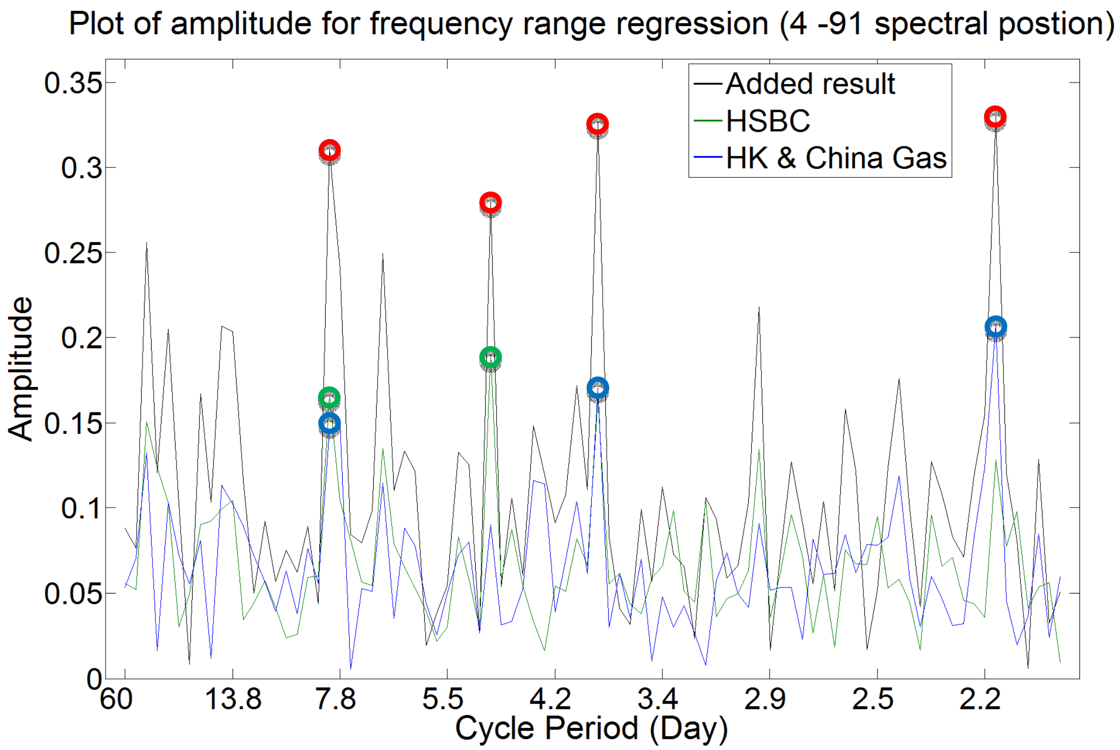

**Figure 3.1 Example of additive properties of Fourier spectrums**

## 3.2 Model Formulation

With reference to section 2.2 for fingerprinting modelling, we explain the portfolio value / market index movements (act as dependent variable) with the prices of individual equity stocks (act as independent variable) by using multiple linear regression model in frequency domain environment, and the model in time series is donated as,

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \varepsilon_t \text{ ,}$$

where the notations of this time series regression model are as follows:

$y_t$ , $x_{it}$ are the time series for the portfolio value / market index and prices of *i*-th equity stock respectively

$\varepsilon_t$ is the time series for the error term in regression

$b_i$ is the regression coefficient for the independent variable $x_i$

$r$ is the total number of equity stocks (independent variables) included in the model

Then, all variables should be transformed in Discrete Fourier Transform defined in (2.5), and the model becomes

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + S_\varepsilon(\omega)\,, \tag{3.1}$$

where the notations of this regression model are as follows:

$S_y(\omega)$ is the frequency spectrum for the portfolio value / market index as the dependent variable $y$

$S_{x_i}(\omega)$ is the frequency spectrum for the prices of *i*-th equity stock as independent variable $x_i$

$S_\varepsilon(\omega)$ is the regression residual in terms of the frequency spectrum

$\omega$ is the pre-defined frequency range in regression modelling (where in our examples, $\omega = \{\omega_n = \frac{n}{T}, n = -\frac{T-1}{2}, \cdots, -5, -4, 4, 5, \cdots, \frac{T-1}{2}, \frac{T}{2}\}$ ; where $T$ is sample size)

Regression is performed using the frequency spectrum between dependent and independent variables in the Discrete Fourier Transform, and thus the frequencies represented in the spectrum should be the same for all the variables in the regression model. In the next section, we will provide some examples to demonstrate mapping of stock constituents for market indices and mutual funds by using our proposed method in fingerprinting identification.

## 3.3 Results & Discussion

In this section, we explained the dependent variable by relating the factors that share the major common peaks between the dependent variable and independent variables. The major peaks in different frequency ranges over the frequency spectrum can be regarded as the unique fingerprint to represent the properties of that variable over time. We used the peak search algorithm to select the factors in the model to explain the prices of HSI index and its sub-indices, and the value of some mutual fund portfolio (dependent variable).

We considered all the equity securities listed in HKEx (excluding ETFs) as the independent variables, and we had 1,532 securities for the selection. The sample contained 180 data points ranged from 14 Jun, 2013 to 18 Mar, 2014. We considered this period because there was no significant change for the constituents of the HSI index. The time series for both dependent (indices) and independent (equities) variables should be ex-dividend, and we only adjusted the price series for stock split and stock dividend, because HSI and its sub-indices are quoted on ex-dividend basis; and thus, we can keep the value of both dependent and independent variables in the same condition. Since the equity data contains common low frequency cycles with the relatively large magnitude, and this may affect our peak search algorithm as it cannot distinguish the target constituents based on the major peaks in the spectrum. In this situation, we disregarded the cycles greater than 60 days (i.e. frequency <= 3/180) in the Fourier spectrum when performing the model selection and regression analysis. Moreover, our examples should not have negative weighting (short selling) for the indices and mutual funds, and we constrained the regression analysis for variables with positive coefficients only.

First, we used the peak search algorithm to identify the constituents in HSI Index and its sub-indices correctly; then, we extended our search to some mutual funds with unknown components, and we can effectively identify the top 10 constituents reported by the funds' management companies. We take $c = -1$ in initialization for the peak search algorithm.

### 3.3.1 Application 1: HSI and its Sub-Index as Dependent Variable

In this example, we used the raw price of HSI and its sub-indices time series to test our proposed search method; while the time series of the individual stock prices served as the independent variables. The total number of stock selected in model ($r$) should be equal to the number of constituents in HSI and its sub-indices. All the indices are correctly matched by using the search method, the sub-indices include Commerce, Utility, Properties and Finance sectors. Table 3.1 shows the fingerprint search results of the indices respectively.

**Table 3.1 Fingerprint search result for HSI and its sub-indices**

| **Sector of HSI and HSI Sub-Indices** | **Number of Constituents** | **Number of Correct factor identified** | **$R^2$** |
|---|---|---|---|
| Commerce | 25 | 25 | 0.999995519521508 |
| Finance | 12 | 12 | 0.999929138219101 |
| Properties | 9 | 9 | 0.999975292794337 |
| Utilities | 4 | 4 | 0.999999914224700 |
| HSI Index | 50 | 50 | 0.999996683378734 |

According to the Table 3.1, all the search results can match the constituents of their corresponding indices with $R^2$ closed to 1. In addition, Figure 3.2 and 3.3 point out the major peaks with the red circles that we can connect in both dependent and independent variables. Thus, the peak search algorithm can effectively select useful factors to explain the dependent variable based on the common peaks in the spectrums.

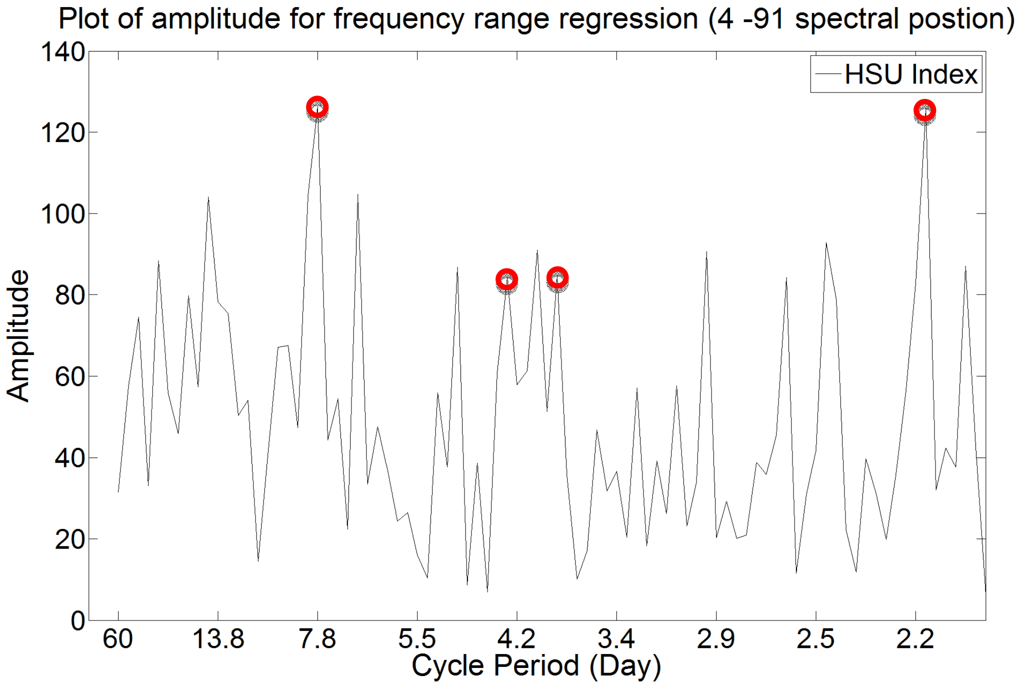

**Figure 3.2 Plot of HSU index spectrum with major common peaks to its two constituents**

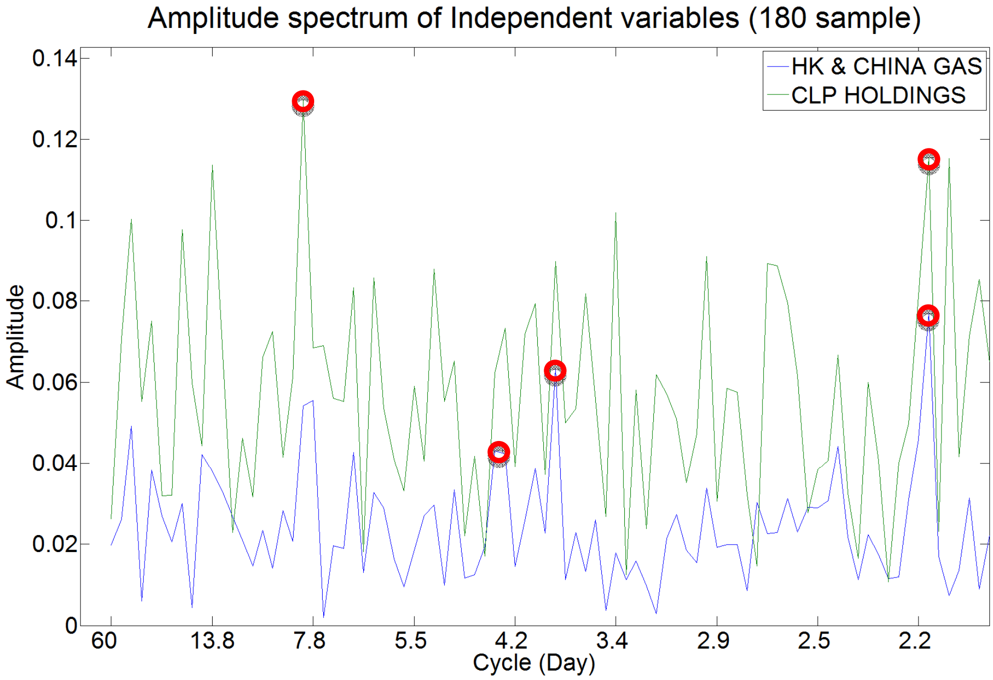

**Figure 3.3 Plot of two HSU index constituents' spectrums with major common peaks to HSU index**

### 3.3.2 Application 2: Mutual Fund as Dependent Variables

In this example, we took two mutual funds to demonstrate our method in selecting a portfolio with high similarity to the mutual fund performance. We chose a sample size in around half-year and just 1-2 weeks after the cut-off time, since the mutual funds usually report their return performance and top 10 holdings half-yearly, and we may expect more changes in their portfolio just before or after the cut-off time.

The price series of the individual stocks served as the independent variables. In order to increase the density of the frequency grids available to represent the constituent peaks, we merged the frequency grids of different sample sizes using Discrete Fourier Transform. Figure 3.4 shows that merged spectrum by different samples can better reflect the characteristic of the cycle patterns in low frequency as sparse frequency can be filled by other sample sizes. The merged spectrum should be the input for the peak search algorithm to select the constituents for the mutual funds, and we set the parameter $r$ in step 1 of section 2.4 to select 30 constituents in the model.

We downloaded the price series of 160 sample points from 2 Jul, 2013 to 21 Feb, 2014 for Eastspring Investments – Hong Kong Equity Fund, and merged 16 spectrums for this mutual fund from 130 – 160 sample points by using the spectrum with one sample point taken out from the start and the end of the data. When the final model was obtained, regression analysis with 160 sample points was performed by the constituents identified, the regression coefficients were used for the calculation of the weighting in each constituent in the mutual funds' portfolio.

In the second mutual fund example (Allianz Hong Kong Equity), we downloaded the price series of 178 sample points from 18 Jun, 2013 to 6 Mar, 2014, and merged 7 spectrums for this mutual fund from 166 – 178 sample points by using the spectrum with one sample point taken out from the start and the end of the data. When the final model was obtained, regression analysis with 178 sample points was performed by the constituents identified.

Table 3.2 – 3.3 show the fingerprint search results of the two mutual funds respectively, and the spectrum plots of the explained model in the two mutual funds are shown in Figure 3.5 & 3.6 respectively.

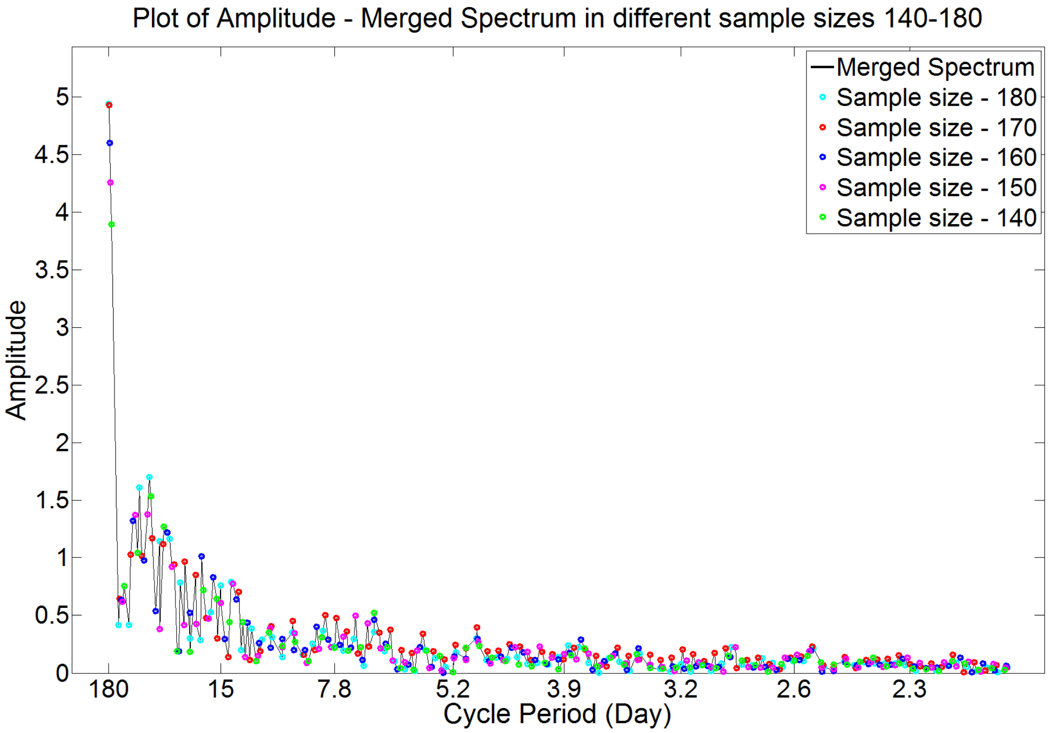

**Figure 3.4 Plot of amplitude spectrum by different sample sizes**

### Table 3.2 Fingerprint search result for the Eastspring mutual fund

| **Eastspring Investments – Hong Kong Equity Fund (瀚亞投資 香港股票基金)** | | |
|---|---|---|
| $R^2$: | 99.99996% | |
| Adjusted $R^2$: | 99.99995% | |
| VIF | 109.35 | |
| **Equity (HKEx code)** | **Weights of Top 10 holding (As of 31 Dec 2013)** | **Regression Coefficient (Estimated Weighting)** |
| AIA (1299) | 9.79% | 10.85% |
| CHEUNG KONG (1) | 9.74% | 10.59% |
| HUTCHISON (13) | 9.38% | 10.46% |
| SHK PPT (16) | 6.06% | 9.33% |
| BOC HONG KONG (2388) | 4.90% | 6.62% |
| WHARF HOLDINGS (4) | 4.16% | 3.04% |
| SJM HOLDINGS (880) | 3.24% | 2.82% |
| LINK REIT (823) | 2.83% | 3.38% |
| HKEx (388) | 2.82% | 4.32% |
| FRANSHION PPT (817) | 2.81% | 3.62% |
| POWER ASSETS (6) | | 1.96% |
| HK & CHINA GAS (3) | | 5.23% |
| CKI HOLDINGS (1038) | | 3.76% |
| NWS HOLDINGS (659) | | 3.53% |
| WUMART (1025) | | 3.33% |
| OOIL (316) | | 2.75% |
| CHINA MOTOR BUS (26) | | 2.64% |
| GALAXY ENT (27) | | 1.92% |
| CATHAY PAC AIR (293) | | 1.86% |
| INTIME (1833) | | 1.33% |
| PETROASIAN (850) | | 1.32% |
| I.T. (999) | | 1.10% |
| CIMC (2039) | | 1.07% |
| SHUNFENG PV (1165) | | 0.69% |
| CHI MER LAND (978) | | 0.55% |
| GEMINI INV (174) | | 0.52% |
| RUI KANG PHARM (8037) | | 0.43% |
| INT'L STD RES (91) | | 0.39% |
| CHINA PPT INV (736) | | 0.38% |
| GLOBAL ENERGY (8192) | | 0.16% |
| **Grand Total** | **55.73%** | **99.94%** |

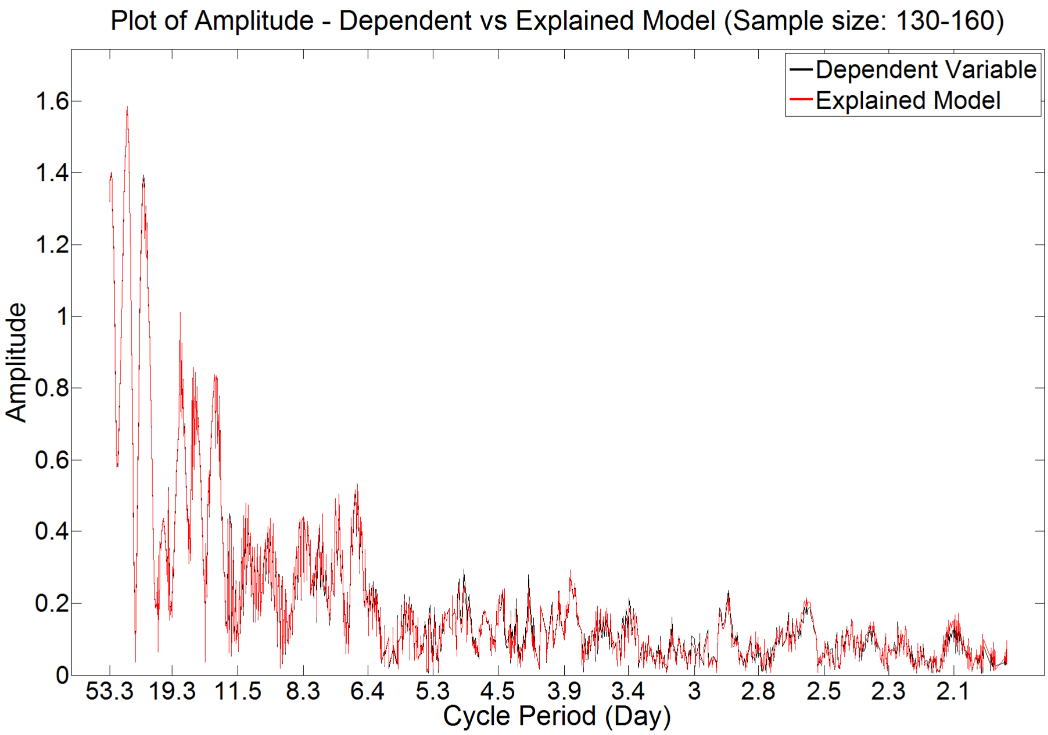

**Figure 3.5 Plot of amplitude spectrum for the Eastspring Investments – Hong Kong Equity Fund and its regression model**

## Table 3.3 Fingerprint search result for the Allianz mutual fund

| **Allianz Hong Kong Equity (安聯香港股票基金)** | | |
|---|---|---|
| $R^2$: | 99.8593819865275 | |
| Adjusted $R^2$: | 99.8593819787876 | |
| VIF | 51.5244 | |
| **Equity (HKEx code)** | **Weights of Top 10 holding (As of 31 Dec 2013)** | **Regression Coefficient (Estimated Weighting)** |
| HSBC HOLDINGS (5) | 7.00% | 5.54% |
| HUTCHISON (13) | 5.00% | 5.54% |
| ICBC (1398) | 4.00% | 11.00% |
| TENCENT (700) | 4.00% | 7.41% |
| CHINA EB INT'L (257) | 4.00% | 7.26% |
| SANDS CHINA LTD (1928) | 4.00% | 5.44% |
| BEIJING ENT (392) | 4.00% | 3.15% |
| CNOOC (883) | 3.00% | 7.30% |
| AIA (1299) | 4.00% | 0.00% |
| BOC HONG KONG (2388) | 3.00% | 0.00% |
| CITIC BANK (998) | | 7.11% |
| NEW WORLD DEV (17) | | 4.76% |
| CIMC (2039) | | 3.27% |
| YIP'S CHEMICAL (408) | | 3.05% |
| CHUANG'S CHINA (298) | | 2.82% |
| CHINA WINDPOWER (182) | | 2.58% |
| PROSPERITY REIT (808) | | 2.57% |
| HENGAN INT'L (1044) | | 2.50% |
| SA SA INT'L (178) | | 2.28% |
| CHINA RES LAND (1109) | | 2.01% |
| ZHONGSHENG HLDG (881) | | 1.97% |
| NATURAL BEAUTY (157) | | 1.80% |
| TSE SUI LUEN (417) | | 1.53% |
| VINCO FINANCIAL (8340) | | 1.39% |
| ND PAPER (2689) | | 1.35% |
| CHINA SCE PPT (1966) | | 1.23% |
| ZHI CHENG H (8130) | | 1.15% |
| BAIYUNSHAN PH (874) | | 1.01% |
| C TAIFENG BED (873) | | 0.93% |
| CHINA GAMMA (164) | | 0.92% |
| **Grand Total** | **42.00%** | **99.86%** |

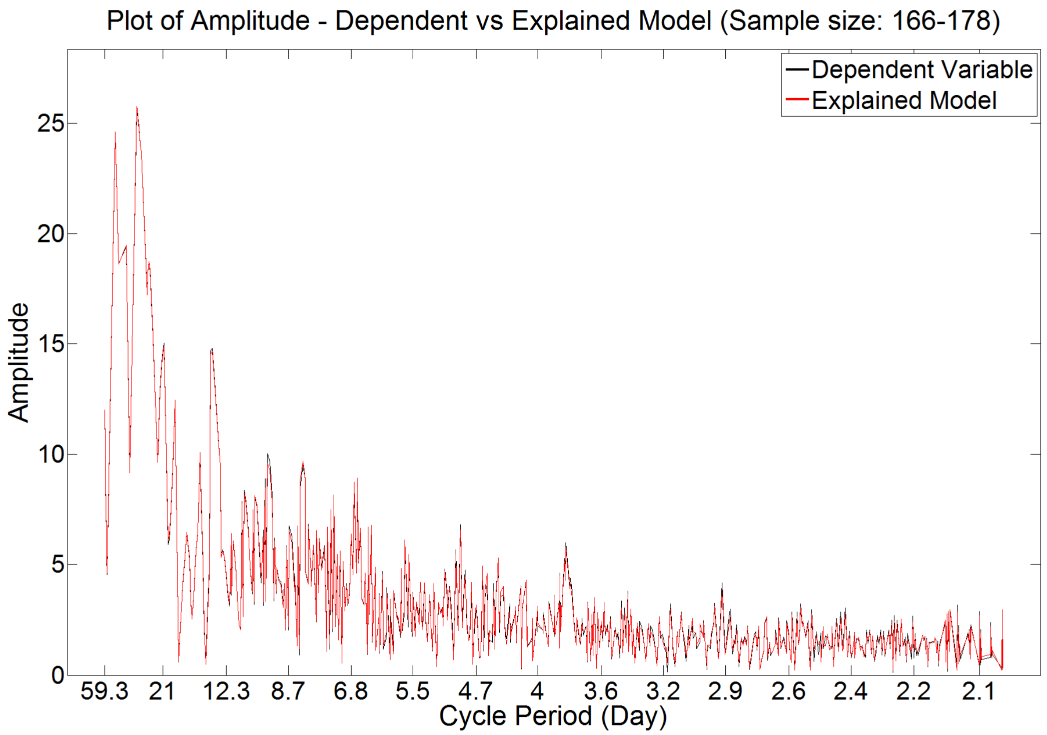

**Figure 3.6 Plot of amplitude spectrum for the Allianz Hong Kong Equity and its regression model**

Considering the regression results in Table 3.2 and 3.3, both fitted models have very high R$^2$ (over 99%), and all the top 10 holdings reported by the respective mutual funds can be selected by using our search model to the Eastspring Mutual Fund (Table 3.2); while the algorithm can match 8 of the top 10 holdings of the Allianz Hong Kong Equity Fund (Table 3.3).

In the Eastspring Mutual Fund result, portfolio weightings from the regression result are quite similar to the weightings of the top 10 holdings, while other equities selected in the regression model, which does not appear in the top 10 holdings, come from different industries. For example, our search model consists of constituents in Utility (HK & CHINA GAS, CKI HOLDINGS, NWS HOLDINGS), Transportation (OOIL, CATHAY PAC AIR), and Consumer goods (WUMART, I.T.); although they do not include in the top 10 holdings, their industries and total weightings are consistent with the major industry groups reported by the funds in the quarterly report.

In Allianz Hong Kong Equity Fund result, portfolio weightings from the regression result tend to be over-weighted when comparing to the top 10 holdings reported, and only two equities (HSBC and BEIJING ENT) are under-weighted. When we summarize the figure by industry groups, weighting in Financials is under-weighted and it may be the reason that the two equities from the top 10 holdings reported (AIA & BOC HONG KONG), which count for 7% of the portfolio, cannot be selected in our model. We find the related constituents in other industry groups where the groups are reported in the fund performance review. We have equities in Utility (BEIJING ENT, CHINA EB INT'L, CHINA WINDPOWER), Energy (CNOOC), Information technology (TENCENT), Consumer discretionary (SANDS CHINA LTD, CHUANG'S CHINA, SA SA INT'L, TSE SUI LUEN) and Industrials (YIP'S CHEMICAL, CIMC, ND PAPER).

To conclude, these examples show that our peak search algorithm can effectively identify the unknown constituents in the portfolio, and investors can benefit by applying this approach to have more understanding about the sensitivities of the fund with reference to the model selected.

# 4. Factor Model with Multiple Timeframes in Independent Variables

## 4.1 Motivation

In the analysis of economic and financial information, we usually deal with flood of factors from different sources, this may come from the production factors, market supply, demand and pricing, finance markets performance, government budgets, and monetary policies; it is important for us to identify the relations between factors to facilitate the proper model set up to explain the target variable; such as Philip curve explains the relations between unemployment rate and CPI. Business cycle becomes a popular approach in macroeconomic data analysis after Hodrick and Prescott (1981), it is an important work to explain the movements of the macroeconomic variables in terms of business cycles. His work further supports the development of modeling in relating different macroeconomic variables based on their co-movements in the business cycles. By choosing the right leading indicators, we can monitor the economic performance and forecast the possible movements. In financial market analysis, investors may not have a homogeneous view on the amount of risk prepared to take and the quality of the stock markets. This quality can be affected by varieties of factors; such as, economy, local or international political issues, investor buying behavior and company-specific news which is useful when specific stock is considered. Ross (1973, 1976) proposes Arbitrage Pricing Theory (APT) which derived from non-arbitrage opportunities existed in efficient financial markets, and security return is determined by multi-factor linear model in terms of unexpected macroeconomic information and risk premium expected. Chan et al. (1986), McElroy and Burmeister (1988) investigate the economic indicators, such as, industrial production, price index, personal consumption, inflation, credit risk premium, interest rate, term structure, and oil prices for testing the APT model in the US stock market. Later, Clare and Thomas (1994), Priestley (1996), Costa et al. (1997), Gianchandani (1997), and Türsoy et al. (2008) use exchange rate, unemployment rate, money supply and lending and deposit rate to the APT model.

However, most previous researchers use economic factors announced by monthly or lower frequency, this may restrict the model as only sampling in single frequency is allowed, and more frequent sampling data must align in sampling with the lowest frequency; thus, information loss will result in the more frequent data as not all the data is fully utilized, since economic data with different sampling frequencies cannot be captured in the model at the same time. Therefore, the solution should not depend on the time domain, but focus on the common patterns between the dependent and independent variables. In this chapter, we focus our work on the possibility of using the frequency domain method to match data under different samplings. Modelling in frequency domain can be further supported by business cycle theory applies to asset pricing modeling (Chen, 1996), since stock prices should revert to the fair value of discounted cash flows of output cycle in the long run, and discount rate which highly relates to term and default premium, correlates to the business conditions (Fama and French, 1989). By this concept, stock price or return time series and economic data time series can be decomposed into a trend (constant), cycles, and noise; while each one should not be correlated. Specific techniques can be deployed for the components as they have their own characteristic, such as the dominant cycles with different frequencies. In this situation, the spectral analysis technique can well suit in this problem by identifying and to quantify the different frequency components of a data series. Masset (2008) discusses the use of frequency domain techniques to investigate different economic time series, such as GDP and the unemployment figure. While Lacobucci (2005) applied the filtering method to decompose US inflation and US unemployment time series into typical business cycle components, and found the relation to support the Phillips curve; however, this relation cannot be shown if only uses the raw data. As a result, it is more advantageous and insightful to investigate the mechanism of target variables by cyclical components of the economic variables, we also extend our search by matching the frequencies of variables in different timeframes which still does not cover by the literature yet. In the next section, we will show the importance of using multiple timeframes in analysis.

### 4.1.1 Importance of Multiple Timeframes

Since data in low frequency sampling may provide directional information to review the effect in the dependent variable, we demonstrated this with an example by comparing two sets of dependent variables which comprised of variables in different sampling intervals. Dependent variable $F_1$ contained three factors sampled daily ( $x_1, x_2, x_3$ ); while, dependent variable $F_2$ consisted of all three factors in the dependent variable $F_1$ , and two additional factors $(x_4, x_5)$ sampled by weekly and monthly respectively. The equations for all the variables and factors are listed as below:

Daily factors:

$$x_1 = 0.1\sin(\frac{2\pi t}{20} - \frac{\pi}{3}) + \sin(\frac{2\pi t}{3600} + \frac{5\pi}{6}) + 0.1\varepsilon \text{ ,}$$

$$x_2 = 0.1\sin(\frac{2\pi t}{50} + \frac{\pi}{6}) + 0.4\sin(\frac{2\pi t}{600} - \frac{\pi}{3}) + \sin(\frac{2\pi t}{3600} - \frac{\pi}{6}) + 0.1\varepsilon \text{ ,}$$

$$x_3 = 0.2\sin(\frac{2\pi t}{100}) + \sin(\frac{2\pi t}{1200} - \frac{\pi}{6}) + 0.1\varepsilon \text{ ,}$$

Weekly factors:

$$x_4 = 0.15\sin(\frac{2\pi t}{14 \times 5} + \frac{\pi}{5}) + 0.6\sin(\frac{2\pi t}{180 \times 5} - \frac{\pi}{6}) + \sin(\frac{2\pi t}{520 \times 5} + \frac{5\pi}{6}) + 0.1\varepsilon \text{ ,}$$

Monthly factors:

$$x_5 = 0.1\sin(\frac{2\pi t}{15 \times 20} + \frac{\pi}{6}) + 0.4\sin(\frac{2\pi t}{350 \times 20} - \frac{\pi}{3}) + 0.2\varepsilon \text{ ,}$$

<u>Dependent variables:</u>

$F_1 = x_1 + x_2 + x_3$,

$F_2 = x_1 + x_2 + x_3 + x_4 + x_5$,

where $\varepsilon$ follows normal distribution with zero mean and 1 unit of standard deviation.

According to Figure 4.1, the movements of the two lines are quite consistent before 1700-th day; however, $F_1$ moves downwards, and $F_2$, which contained variables in weekly and monthly sampling, moves upwards instead. This example illustrates the value to consider factors which are sampled in low frequency in the analysis for more thorough study in predicting future price movements under the dynamic environment. Therefore, we should consider other techniques that combine variables with different timeframes into one framework. Spectral analysis is a good technique to represent factors in different frequencies even their timeframes are different, and this technique has extensive research in supporting its applications in the economic and finance fields.

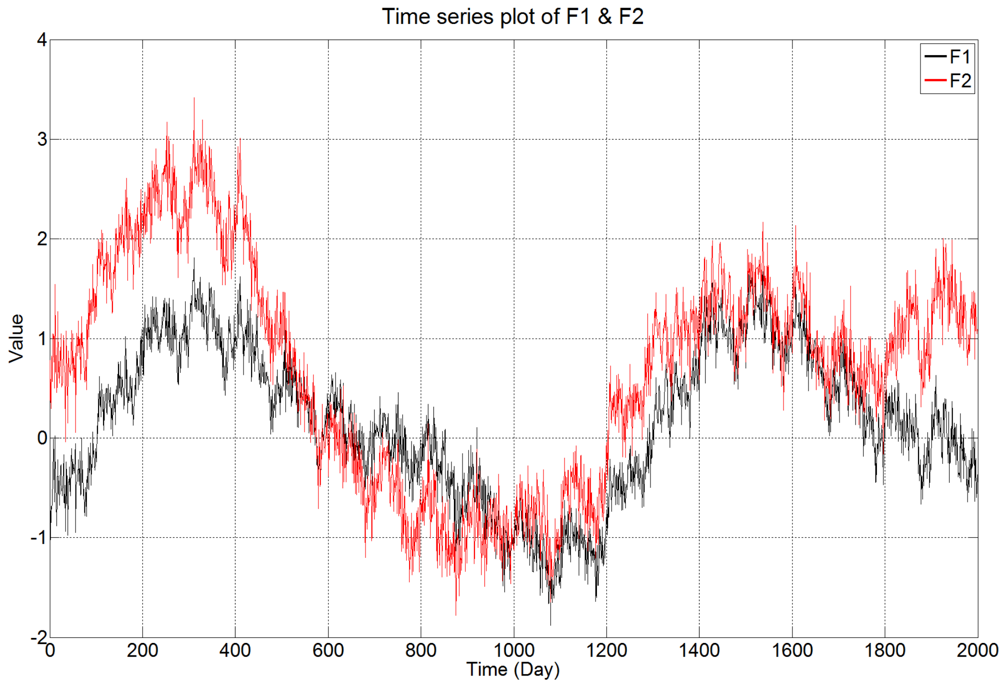

**Figure 4.1 Time series plot of IFFT outcomes for $F_1$ (red line) and $F_2$ (black line)**

## 4.2 Model Formulation

With reference to section 2.3 for the time series modelling, we explain the stock market index or particular macroeconomic information (act as dependent variable) with macroeconomic data in consideration (act as independent variable) by using multiple linear regression model, and the model in time series is denoted as,

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \varepsilon_t \, .$$

Then, all variables should be transformed in Discrete Fourier Transform defined in (2.5) as we perform regression in frequency domain environment for the peak search algorithm, and the model becomes

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + S_{\varepsilon}(\omega) \, , \tag{4.1}$$

where the notations of this regression model are as follows:

$S_y(\omega)$ is the filtered frequency spectrum for the stock market index or particular macroeconomic information ($y$) as the dependent variable

$S_{x_i}(\omega)$ is the filtered frequency spectrum for the macroeconomic data ($x_i$) as independent variable

$S_{\varepsilon}(\omega)$ is the regression residual in terms of the frequency spectrum

$\omega$ is the pre-defined frequency range in regression modelling (where in our examples, $\omega = \{\omega_n = \frac{n}{T}, n = -\frac{T-1}{2}, \cdots, -5, -4, 4, 5, \cdots, \frac{T-1}{2}, \frac{T}{2}\}$ ; where $T$ is sample size)

$b_i$ is the regression coefficient for the independent variable $x_i$

$r$ is the total number of macroeconomic factors (independent variables) included in the model

Lag variables will be included if regression residual shows autocorrelation, and the final model in the frequency domain becomes

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + \sum_{p=1}^{P} a_p S_{y,t-p}(\omega) + \sum_{q=1}^{Q} \sum_{i=1}^{r} d_{i,q} S_{x_{i,t-q}}(\omega) + S_{\varepsilon'}(\omega) ,$$

where $P$ is the largest time lag applied and $a_p$ is lag variable coefficients in dependent variable, and $Q$ is the largest time lag applied and $d_{i,q}$ is lag variable coefficients in independent variables. And the final model by time series can be written as,

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \sum_{p=1}^{P} a_p y_{t-p} + \sum_{q=1}^{Q} \sum_{i=1}^{r} d_{i,q} x_{i,t-q} + \varepsilon_t .$$

The analysis outcome went through the processes as follows in order to fulfill the requirements in regression:

- Stationary: Frequency domain filter removes non-stationary components in dependent and independent variables (see section 2.3.1). Cycles greater than 333.3 days (i.e. frequency is 3/1000) were disregarded in the regression analysis.
- Multiple Timeframes: Matching the frequency spectrums of the independent variables in different timeframes for the regression spectrums. (see section 2.3.2)
- Error Analysis: Regression residual should be free from autocorrelation by Durbin-Watson test; otherwise, lag variables will be introduced to solve this problem. (see section 2.3.3)
- Residual Normality Check: After the error analysis, we check the residual for normality by Dickey Filler test. If residual is non-normal, we use bootstrapping to ensure the model is significant in F statistics. (see section 2.3.4)

Then, the properties of the macroeconomic factors will be discussed in section 4.3. Finally, we related our two examples to the macroeconomic data in different timeframes and discuss those factors selected in section 4.4.

## 4.3 Macroeconomic Factors

Data were downloaded from the Bloomberg Terminal system, where we took the closing value for the macroeconomic data. The system returns “N/A” if data on that date is not available. We focused on US macroeconomic indicators as the independent variables to be selected for modeling. Table 4.1 summarizes the number of macro factors that we considered in different timeframes.

**Table 4.1 Number of macro factors for selection by regression analysis**

| | Timeframe | | | |
|---|---|---|---|---|
| | Daily | Weekly | Monthly | Total |
| Number of macro factors considered | 188 | 1,311 | 201 | 1,700 |

Table 4.2 summarizes the categories that we covered in the macroeconomic data and the number of macro-indicators chosen in each category. It has a brief description of each category with example and explain its importance to the economy and stock market. We covered most of the categories available in Bloomberg Terminal except the forecast data. It is because forecast data by analyst advice may be subjective, and there may be a high probability for the forecast data deviated from the actual result in the future; thus, they are not the good indicators to explain the rations to the stock market index or other macroeconomic information.

**Table 4.2 Major categories in macroeconomic factor**

| # | Category | Description | Number of factors included |
|---|---|---|---|
| 1 | National Account | It measures the output of the whole economy, and it can be in terms of income and expenditure for the households, corporations and government sectors within or over other countries' economies; for example, GDP, GNP and industrial production. | 4 |
| 2 | Consumer / Producer Price | It measures the increase or decrease in the price level over a basket of goods & services consumed by consumer / producer. Some examples are consumer price index (CPI), producer price index (PPI), and IMF US PPI / WPI (Wholesaler Price Index). | 16 |
| 3 | Labour Market | It provides indication about the supply and demand of labour, and the employment level and wages in general for the economy. US Initial / Continuing Jobless Claims SA (Seasonally Adjusted) is one of the examples. | 45 |
| 4 | Economic Activity & Business Conditions | It indicates the healthiness and business prospect in the economy and affects every industry sector in the region. It relates to the mood of investment, confidence in consumer spending, which in aggregate may consider as business climate. Interest rates, inflation, and employment are some of the parameters in business operations, which usually subject to the change in business climate. Some examples include manufacturing capacity utilization, factory / sales inventory level, manufacturing new order, PMI, and (national) activity index. | 64 |
| 5 | Housing Market | Housing market is a major sector in the economy, it equals to around five percentages of the US GDP. The prosperity of housing market may have a good impact to its related industries, like construction, demand in durable goods, and mortgage loans & financing. Indicators such as housing price & sales (Case-Shiller Composite-20 City Home Price Index, US Existing Homes Sales) and building permits (Private Housing Authorized by Building Permits by Type Total), usually regard as key information in reviewing the housing market status. | 76 |

| # | Category | Description | Number of factors included |
|---|---|---|---|
| 6 | Retail Sector & Consumer Confidence | It measures goods sold to final customers for personal consumption via different channels, this includes physical stores, sales by mail order or even online; for example, ICSC US Retail Chain Store Sales Index is an indicator to measure sales changes in chain store. In addition, surveys to consumers regarding their spending outlook reflect in consumer confidence indicators, such as Bloomberg US Weekly Consumer Comfort Index. The indicators can preview the demand and activeness in consumer sector, and this strongly correlates to the overall economy. | 68 |
| 7 | Personal Sector | It measures aggregated household income and expenditure. The income may include, but not limited to, the amounts from individual's wages and salaries, dividends, interests, and rental incomes. Wages and salaries are expected to be the major component in personal sector, and US Personal Income MoM (Month over Month) is one of the indicators which measures the month to month changes of the income. While, expenditure may include, but not limited to, the payments on goods and services, interest payments (e.g. mortgage), and to the government (e.g. tax and social services). The difference between income and expenditure should be the saving for the personal section, and Bureau of Statistics announces the percentage in disposable income as saving every month. The changes in saving percentage, income and expenditure amounts can affect the capital flows, pricing for goods and services and capital markets. | 15 |
| 8 | External Sector | It measures the interaction between the local economy and other countries' economies in terms of monetary term (e.g. dollar amount). For example, US Trade Balance of Goods and Services SA includes the trade balance (i.e. total export deducted by import) from goods and services. Balance of payment in capital flows is also the key measurement in external sector. Interaction between countries can affect the local economy when the changes to the overall production become significant. | 6 |

| # | Category | Description | Number of factors included |
|---|---|---|---|
| 9 | Government Sector | Government sector is usually a major sector in the economy, it provides goods and services that private sector cannot be or does not willing to supply, but importance to society. The goods and services may include national security, social safety, infrastructure, public transit, and social welfare (e.g. education, health care, and subsidy for the poor). Since government incomes are mainly sourced by selling public resources (e.g. lands, business licenses) and compulsory tax incomes, these facilitate the redistribution of resources from surplus sectors to deficit sectors. Thus, decisions by policy makers should induce the direct impact to the economy via government sector, for example, US Treasury Total Public Debt Outstanding indicates the debt level in the government sector as a whole; and indirect impact to the market operations via the policy changes such as controlling federal reserve rate. | 42 |
| 10 | Financial & Monetary Sector | Its indicators mainly come from the financial markets and the related institutes, for example, stock markets, bond markets, and credit markets; some examples are US Generic Govt 10 Year Yield, Mortgage Bankers Total Points 30-Yr Jumbo Effective Rate, Stock Prices 500 Common Stocks, and ICI Retail Money Market Funds Government Treasury & Repo (Repurchase Agreement) Total Net Assets. While, some indicators related to the financial intermediaries, such as banks, insurance companies, fund houses and brokers, may be important to review the status of the economy. For example, US Commercial Bank Assets Bank Credit NSA, Securities Mortgage-Backed Securities held by commercial banks. Some indicators show the effects and the flows of money to the economy, for example money supply measured by M1-3. Indicators from financial & monetary sector can review business condition and risk appetite in the markets, which provide important reference to economists and Federal Open Market Committee in forming monetary policies. | 1292 |
| 11 | Commodity Market | Commodities are the key component in personal consumption and production. Manufacturing consumes energy, metals, food, chemicals to produce goods that supporting the consumption and economic activities for different sectors. Therefore, the changes in production level and its price s may have an impact on the factor input in other sectors. For example, USDA WASDE Supply & Use US Corn Production monitors the supply of corn. | 72 |

**Commodity Price:**

Commodity prices are usually determined by the interaction of supply and demand factors; while, financial markets, such as fixed income, equity and foreign exchange, also response to the price movements in commodities' future contracts, especially some of the major commodities. Rising trend in crude oil price usually connects to the negative impact on the equity and bond markets, since this may imply accelerated inflation in the future and lead to a higher return premium. In addition, higher production cost is expected as crude oil & its related products are nearly involved in every processing; this may harm business profit if the increase cannot be shifted to the consumers. US stock market crash in 1973-74 led by oil price soar is one of the examples. Other commodities, such as agricultural price, tend to receive less attention from the markets except it shows unexpected movements significantly, it indicates the changes of production factor and market will gradually reflect this unexpected information. (Tainer, 2006; Baumohl, 2012)

**Inflation / Deflation:**

Inflation (Deflation) is higher (lower) general price level expected on goods and services compared with the past. Since the baskets for goods and services that people needed are different in each country or even region of the same country; thus, there is no acceptable one standard for measuring price level in the worldwide. The market mainly focuses on the consumer price index (CPI) and producer price index (PPI) which review the changes in price level on the two major sectors in consumer and manufacturing. Equity and fixed income have a negative impact when inflation increased sharply, especially for unexpected change by more than 20% from the market expectation, and market participants will form a new path for inflation expectation. It is because CPI / PPI is a lag indicator to the business cycle and the financial market, participants focus on unexpected changes to adjust their expectation toward future streams of business profitability. Market will first reflect the inflation result by the core rate (i.e. energy and food items), then attention will be placed on any unexpected changes in specific categories, such as housing, transportation in CPI; equipment and nondurable goods in PPI; this leads to capital flows between different sectors or even asset classes by responding to changing expectation in the financial markets. (Tainer, 2006; Baumohl, 2012)

**Financial Market:**

An efficient financial market should be able to provide a platform to transfer excess money from the surplus sectors to the deficit sectors, and facilitate in regulating the circulation of money in the market. In US, this is achieved through market operations by Federal Reserve System. The system can control the capital flows by setting discount rate and reserve requirements, or through open market operations by buying and selling securities to affect the market prices. Open market operations are the flexible ways to regulate funds reserved in the Federal Reserve Banks by changing the market demand for funds. This leads to a change in federal funds rate due to Federal Reserve System operations. In addition, the yields of corporate bonds will be affected when the supply of treasury securities (risk-free assets) increase; then this pushes up the yield for treasury securities and draws out capital from corporate bond to treasury as it is relatively cheaper if other conditions keep the same. Therefore, market force will drive the corporate bond price lower until equilibrium price attained, this is called the “crowded out” effect to private sector as treasury securities compete with the corporate issues. The borrowing cost increases when interest rate is higher, and it is less attractive for new investment to take place; this is a bad news to equity market as economic activities expected to be cool down.

Market participants always pay attention to the yield curve; this is more meaningful to analyze the interest rate by different maturity periods rather than just watching one rate. It is because yield curve can better reflect the market price for money in different terms (periods in maturity). When economic growth is moderate, we expect slight changes in interest rate over time, and current market prices should show flat yield curve with a small difference between short term and long term interest rates. A positively sloped yield curve is a usual pattern in most of the time; it is because investors need more return in order to tie up their funds for longer horizon. In addition, a positively sloped yield curve also previews an accelerating economic growth with higher than current inflation rate expected in long run, and this is a good news to equity market as accelerated economic growth means higher risk appetite and stimulates the buying force in the stock market. However, inverted yield curve is not a common form, but we faced this situation during operation twist by Federal Reverse System to inject short term liquidity after the financial crisis, in order to alleviate short term credit demand in the market. (Tainer, 2006; Baumohl, 2012)

Financial Intermediaries, such as banks, are also the key players in the financial system. Developing counties depend more on intermediaries, while developed countries, like US and Germany, rely more on financial markets to facilitate transfer of funds to sectors in needed. Banks accept credit risk and lend the funds from saving by different sectors. Banks can control their risks by assessing borrowers' operations and credit information; they can develop the close relationship with the borrowers and keep track of their status to the best interest of the banks. Therefore, changes in asset and liability portfolio can envisage the attitude and willingness for intermediaries to take risk on different sectors or industries; it is a favorable signal to asset markets when banks pro-actively lend money outside, this help to boost the economic growth as a whole; however, when banks become conservative and keep more capital, this should indicate riskier economy. Investors will request a higher risk premium for higher potential risk taken, and this drives down the prices for risky asset classes, but this is favorable to risk-free asset as more capital goes to the money market. (Franklin and Gale, 2000)

### Housing Market:

Housing market is a major sector in the economy, it equals to around five percentages of the US GDP. The prosperity of housing market may have a good impact to its related industries, like construction, demand in durable goods, and mortgage loans & financing. Housing starts and permits are important monthly indicators to identify the level of housing activities, and permits are regarded as the leading indicator for housing starts and the economy. Fix income market considers favorable figures in housing starts and home sales as bad news to the bond price, since this signifies interest rate rise in the blooming economy and suppress bond price to achieve higher yields; however, it is good news for equity market and strong dollar in foreign exchange market, where growth in the economy implies business profit growth even interest rate is expected to rise as well. Housing starts & permits show its importance when market coming out from recession; it is a confirmation signal to expect stronger economic activities soon when housing starts & permits increase, and equity market should rebound from their bottom level. Changes in new and existing homes sales and pricing can indicate the demand for housing and relate to the current level of economic activities, and they are worth for our attention during the expansion or recession phase. The sales and pricing figure should move in the same direction with the housing starts & permits when market in an increasing trend; otherwise, the economic outlook will be skeptical if the discrepancy persists. (Tainer, 2006; Baumohl, 2012)

### Economic Activity / Business Conditions:

Total business inventories, capacity utilization, new order and economic activity survey are the collection of indicators to review the activeness in the economy and business conditions. They are some effective ways to quantify the level of economic activities and relate to the phase of business cycle. During market downside, we expect drop in utilization and new order, it is unfavorable to equity markets; while bond prices increase in low interest rate environment. When businesses liquidate inventory in poor economic situation (drop in business inventories), market should expect production increase in short period and this is positive to equity market and local currency exchange rate for improvement in corporate earnings and economic momentum, but negative impact to bond market. In prosperity economy, production increases to fulfill economic growth; however, if the production capacity cannot catch up with the growth (e.g utilization maintains at extremely high level at over 88%), it will reflect in unplanned drop in inventories, and general price level will increase to increase quantity supply, and this may affect the healthy growth of the economy when growth is mainly contributed by the price level increase. Market will respond to this situation by requiring a higher rate of return on investment as it is more likely for the economic downturn. (Tainer, 2006; Baumohl, 2012)

In the next section, we first studied some empirical examples for the importance of macro factors; then, we deployed the peak search algorithm for selecting factors to explain key macro data and SPX index, and results will be summarized for further discussion.

## 4.4 Results and Discussion

### 4.4.1 Empirical Examples to Relate Macro Factors

In this section, we demonstrated that frequency domain method is effective in identifying common characteristics between variables by comparing the results from some literature as the example and the results by using our method. We based on the result by Lacobucci (2005), which indicates that a strong relation between US inflation and unemployment rate to support the theory of Phillip curve. Frequency domain method can draw the same conclusion when we investigated the co-spectrum for US inflation and unemployment rate using the same sample from 1960 – 2002. We followed the steps by Lacobucci to perform band pass for the input data with the cycle period greater than 252 months (21 years) or smaller than 12 months. Then, we smooth out the density of the co-spectrum by taking average of the three density values along the frequency axis as equation (4.1), and the density plots for Lacobucci's method (left), and our method (right) are shown accordingly in Figure 4.2.

$$\bar{w}_3(\lambda_s) = \frac{w(\lambda_{s-1}) + w(\lambda_s) + w(\lambda_{s+1})}{3} \tag{4.1}$$

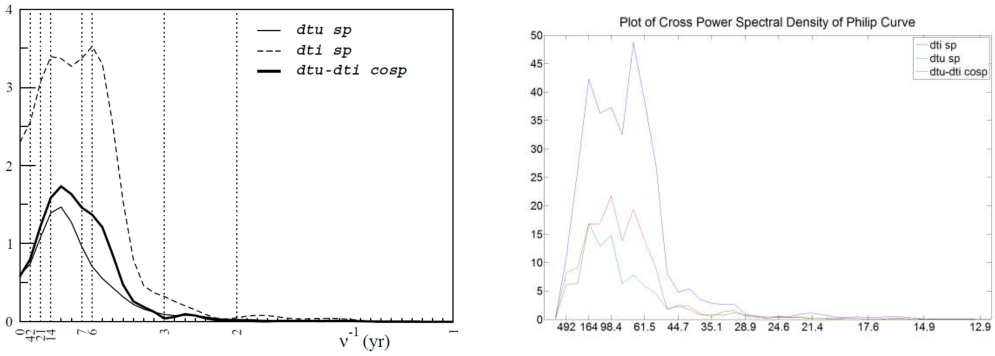

**Figure 4.2 Co-spectrum density for Phillip curve (Lacobucci's method (left), and our method (right))**

In Figure 4.2, dtu refers to the density spectrum of the US unemployment rate (dotted / blue line); dti refers to the density spectrum of the US inflation rate (blank / green line), and dtu-dti refers to the co-spectrum density for the two lines (bolded black / red line). In the co-spectrum (bolded black / red line), both graphs show the peak ranged from around 6 – 14 years which is consistent with the conclusion made by Lacobucci (2005). This demonstrated that macroeconomic information could be modeled by other macroeconomic data in linear term.

In the second example, we showed that considering variables in multiple timeframes can improve the results by relating hedge funds to major macroeconomic and finance factors in Hasanhodzic and Lo (2007) according to equation (4.2).

$$R_{it} = \alpha_i + \beta_{i1} f_1 + ... + \beta_{iN} f_N + \varepsilon_{it} \tag{4.2}$$

It is a *N*-factor risk model, where $R_{it}$ is *i*-th hedge fund return at time *t*, and $\beta_{i1}$ is the sensitivity of the first factor for *i*-th hedge fund, and $f_1$ is the first risk factor in model. We hope that similar adjusted R$^2$ result in our regression when compared with the results in the paper. We performed regression on monthly return data for the 725 hedge funds downloaded from the Bloomberg terminal with 50 monthly sample points until end of Mar 2013. There are six factors suggested; however, some factors may not be valid anymore, and the Table 4.3 indicates the factors that we employed in regression analysis:

**Table 4.3 The factor employed in Hasanhodzic and Lo (2007) and our regression**

| **Category** | **Hasanhodzic and Lo (2007)** | **Factors used in my Model** |
|---|---|---|
| Foreign Exchange | The US Dollar Index return | Spot Dollar Index return |
| Bond | The return on the Lehman Corporate AA Intermediate Bond Index | CS Emerging Market Corporate Bond AA Bucket Total Return |
| Credit | The spread between the Lehman BAA Corporate Bond Index and the Lehman Treasury Index | The spread between the Moody's Bond Indices Corporate BAA and Bloomberg US Generic Govt 30 Year Yield |
| Stock market | The S&P 500 total return | The S&P 500 total return |
| Commodity | The Goldman Sachs Commodity Index (GSCI) total return | Goldman Sachs Commodity Index Total Return Chicago Mercantile Exchange |
| Market Volatility | The first-difference of the end-of-month value of the CBOE Volatility Index (VIX) | The first-difference of the Chicago Board Options Exchange SPX Volatility Index |

* Since factors related to Lehman Brother are not available anymore, we chose factors from other vendors which tracking the similar information.

**Table 4.4 Adjusted $R^2$ of the regression by hedge fund strategy groups**

| Strategy Group | Average Adjusted $R^2$ (My Data) Sample: Jan 2010 – Mar 2013 | | | Average Adjusted $R^2$ (Hasanhodzic and Lo (2007)) Sample: Feb 1986 – Sep 2005 | | |
|---|---|---|---|---|---|---|
| | Number of Hedge Fund | Average | Max | Number of Hedge Fund | Average | Max |
| CTA/Managed Futures | 124 | 16.11% | 73.36% | 114 | 15.3% | 70.0% |
| Equity Hedge | 353 | 38.95% | 91.57% | 520 | 21.6% | 90.2% |
| Event Driven | 22 | 29.03% | 57.39% | 169 | 19.5% | 68.5% |
| Fixed Income Arbitrage | 63 | 31.44% | 68.15% | 62 | 14.9% | 78.9% |
| Macro | 72 | 36.96% | 83.35% | 54 | 14.8% | 74.0% |
| Multi-Strategy | 91 | 33.27% | 82.23% | 59 | 12.9% | 51.7% |

According to the regression model in equation (4.2), we summarize adjusted R$^2$ of the regression results grouped by hedge fund strategies in Table 4.4. We found that the adjusted R$^2$ is similar when comparing the maximum adjusted R$^2$ of the groups from the paper; thus, an equity portfolio could be related to some macroeconomic and financial data in a linear model with good explanation power. Next, we will base on our peak search algorithm and include our data pool of macroeconomic variables to perform the linear regression again for a hedge fund with high adjusted R$^2$ in monthly return. In this regression analysis, we use the daily return for comparison, so that variables in different timeframes were considered. We had significant improvement in R$^2$ for our model compared with the factor model proposed by the paper. We can demonstrate the improvement by including factors with different timeframes, and the results are summarized as shown in Table 4.5.

**Table 4.5 Hedge fund example: GLOBAL DYNAMIC OPPORT-FDF 86**

Adjusted R-squared: 41.94% with the 6 factors from the paper (VIF: 3.92)

76.82% with 11 factors from our database (VIF: 1.65)

| # | **Macro Factors** | **Sampling** | **Coefficient Value** |
|---|---|---|---|
| 1 | S&P 500 Total Return Index | daily | $1.53\ x10^{-4}$ |
| 2 | US Total Public Debt Outstanding TIPS | monthly | $4.38x10^{-5}$ |
| 3 | US Labor Force Participation Rate Total SA | monthly | -2.54117 |
| 4 | Foreign Related Institutions Total Assets SA | weekly | $5\ x10^{-4}$ |
| 5 | S&P/Case-Shiller Composite-20 City Home Price Index SA MOM % Change | monthly | 0.615327 |
| 6 | Conference Board US Leading Index Leading Credit Index | monthly | 0.466064 |
| 7 | US Foreign Net Transactions | monthly | $-4.49x10^{-3}$ |
| 8 | Monetary Base Total YoY NSA | weekly | $1.6696x10^{-2}$ |
| 9 | DOE Crude Oil Output Implied Demand Data | weekly | $-6.44x10^{-5}$ |
| 10 | Johnson Redbook Index Same Store Sales Monthly YoY | weekly | $4.2253x10^{-2}$ |
| 11 | Bloomberg US Weekly Consumer Comfort Index for Incomes $40K To $49.9K | weekly | $5.961x10^{-3}$ |

In this section, we investigated some key macroeconomic data that can be explained by other macroeconomic factors, and modelling by variables in multiple timeframes can improve the result significantly. In the next section, we used our methodology to explain the Cleveland Federal Reserve Bank Financial Stress Index and S & P 500 Index (SPX) by macroeconomic factors in multiple timeframes. Most of the crucial factors should be selected to our model with high $R^2$ in the two examples. Then we performed the residual analysis, and included the variables in different lag periods to remove the autocorrelation in regression residual.

### 4.4.2 The Cleveland Federal Reserve Bank Financial Stress Index

This variable tracks the distress of US financial system in the economy. This is a coincident indicator announced daily to the potential distress developed in different markets. They include bond / credit markets, equity markets, foreign exchange markets, funding markets (inter-bank markets), and real estate markets. Higher value in the variable stands for higher risk in distress, please refer to reference #1 for details. The regression analysis result is listed in Table 4.6, and residual analysis is shown in Figure 4.3.

First, we determined to take 1000 daily sample points for analysis, where we downloaded the data until 26 Mar 2013. The starting date for most of the daily factors was around Mid March 2009, which ranged from 6 Jan 2009 to 16 Mar 2009. According to section 2.3.2 in matching the frequencies in the timeframes, we used 200 sample points for weekly factors, and 46 sample points for monthly factors. The corresponding starting date for weekly factors was mid of May for most of the cases, which ranged from 23 Jan 2009 to 22 Jun 2009; while most of the monthly factors started at the beginning of May, and they ranged from 2 Mar 2009 to 12 May 2009. We used $c = 1.6$ in initialization for the peak search algorithm. Generated optimal models were deployed as the base for factor adding and replacement to obtain the search result for models with more factors, and the table below shows the $R^2$, AIC and BIC value for the optimal models with 4 – 9 factors.

| Number of factor in model | $R^2$ | AIC value | BIC value |
|---:|---:|---:|---:|
| 4 | 77.78% | 10662.22 | 3813.598 |
| 5 | 81.33% | 10490.99 | 3647.279 |
| 6 | 83.33% | 10380.30 | 3541.488 |
| 7 | 84.12% | 10333.69 | 3499.784 |
| 8 | 86.53% | 10172.12 | 3343.115 |
| 9 | 86.97% | 10140.73 | 3316.623 |

We found that AIC and BIC value tends to include more factors by taking the minimum value, but the increase in $R^2$ is quite small indeed; for example, there is only 0.44% improvement in $R^2$ for a 9-factor model. Finally, we chose a 6-factor model with $R^2$ increase by more than 1% for further analysis.

**Table 4.6 Regression analysis of financial stress index**

R-squared / Adjusted R-squared: 83.33% / 83.23%

VIF: 3.0043 (6 factors)

Sample standard deviation: 0.448712

| # | Macro Factors | Sampling | Coefficient Value | Standardized Coefficient | Sample Standard Deviation |
|---|---|---|---|---|---|
| 1 | Conference Board US Leading Index Stock Prices 500 Common Stocks | Monthly | -0.17111 | -0.68846 | 1.805415 |
| 2 | US Generic Govt 10 Year Yield | Daily | -1.01318 | -0.48444 | 0.214549 |
| 3 | Commercial Paper Outstanding for Financial Cos Foreign Issuers Others | Weekly | 0.06948 | 0.22138 | 1.429711 |
| 4 | S&P/Case-Shiller Composite-20 City Home Price Index YoY | Monthly | -8.34367 | -0.45253 | 0.024336 |
| 5 | Fed Rsrv Rate Paid by Fixed Rate Payer Int Rate Swap with Maturity of Ten Year | Daily | -0.55581 | -0.31473 | 0.254086 |
| 6 | IMF US PPI/WPI | Monthly | 3.275963 | 0.330605 | 0.045283 |

According to the residual analysis in Figure 4.3, residual shows significant autocorrelation in ACF and PACF plots. In addition, Durbin-Watson test also rejects the null hypothesis of no autocorrelation in the time series at over 99% significant.

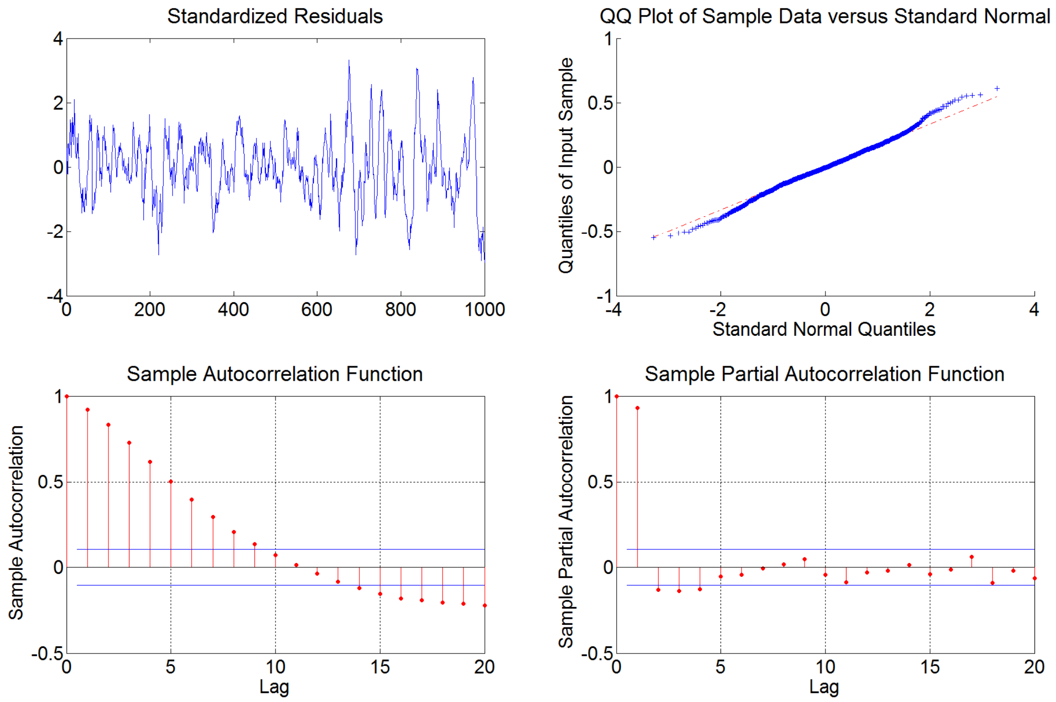

**Figure 4.3 Residual analysis for regression on the financial stress index**

Thus, lag variables were included on top of the model in Table 4.6 to remove autocorrelation property of the residual; the final model should use minimal lag variables to achieve the lowest autocorrelation in the residual, and the following algorithm indicates the model selected to explain financial stress index:

$$Y_t = \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \hat{b}_3 X_{3t} + \hat{b}_4 X_{4t} + \hat{b}_5 X_{5t} + \hat{b}_6 X_{6t} + \hat{b}_7 Y_{t-1} + \hat{b}_8 Y_{t-2} + \hat{b}_9 Y_{t-5} + \hat{\varepsilon}_t$$

In Figure 4.4, residual analysis shows no autocorrelation in the model with lag variables, and Durbin-Watson test cannot reject the null hypothesis of no autocorrelation in the time series at 94% significant. However, the residual shows non-normal property for value larger than two standard deviations in the distribution, especially with flat tail and extreme value at the two sides. As residual is non-normal, significant test with normal assumption cannot directly apply in the regression; we used bootstrapping by case resampling to test the consistency of the model by F-test. Case resampling means randomly sample with replacement from the same pair of dependent and independent variables in the original database. The F-test result is

847.6 which is the one percentile of the bootstrap distribution, and the critical bound is $F(9,991) \approx 2.43$ at 99% significant, which means the result is greater than the critical value at 99%. We summarize the regression result in Table 4.7, which consists of six factors and three lag variables.

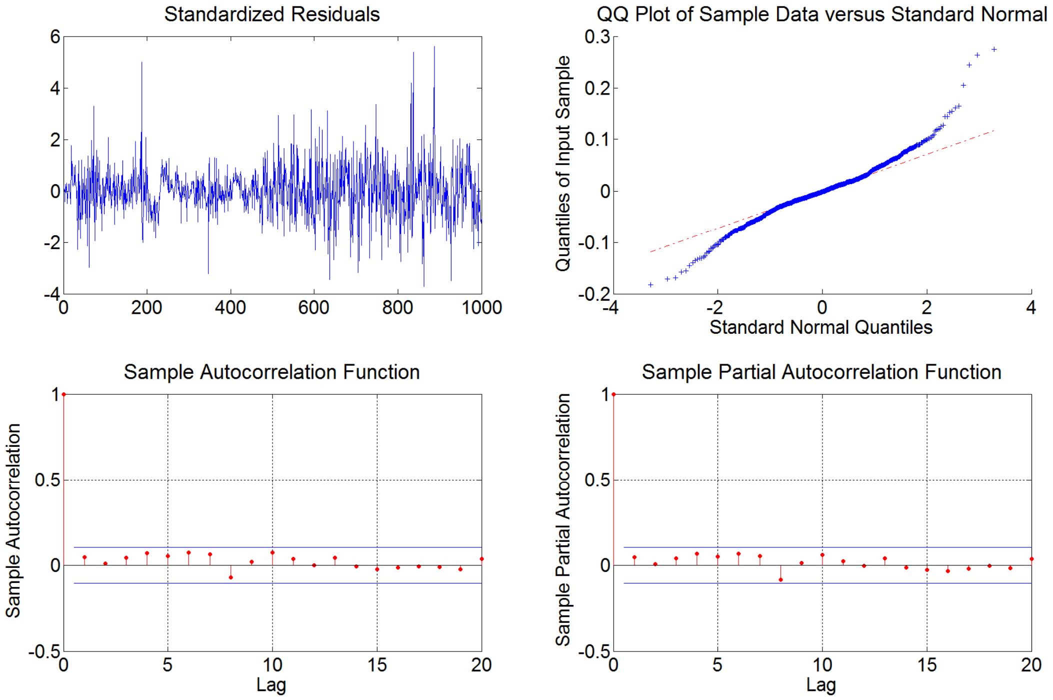

**Figure 4.4 Residual analysis for regression on the financial stress index with lag variables**

**Table 4.7 Regression analysis of the financial stress index (with lag variables)**

R-squared / Adjusted R-squared: 98.81% / 98.80%

VIF: 110.32

Sample standard deviation: 0.448712

| # | Macro Factors | Sampling | Coefficient Value | Standardized Coefficient | Sample Standard Deviation |
|---|---|---|---|---|---|
| 1 | Conference Board US Leading Index Stock Prices 500 Common Stocks | Monthly | -0.01474 | -0.05929 | 1.805415 |
| 2 | US Generic Govt 10 Year Yield | Daily | -0.16896 | -0.08079 | 0.214549 |
| 3 | Commercial Paper Outstanding for Financial Cos Foreign Issuers Others | Weekly | 0.005468 | 0.017421 | 1.429711 |
| 4 | S&P/Case-Shiller Composite-20 City Home Price Index YoY | Monthly | -0.91247 | -0.04949 | 0.024336 |
| 5 | Fed Rsrv Rate Paid by Fixed Rate Payer Int Rate Swap with Maturity of Ten Year | Daily | -0.02606 | -0.01476 | 0.254086 |
| 6 | IMF US PPI/WPI | Monthly | 0.391957 | 0.039556 | 0.045283 |
| 7 | Cleveland Federal Reserve Bank Financial Stress Index (1-day lag) | Daily | 1.099573 | 1.104494 | 0.450721 |
| 8 | Cleveland Federal Reserve Bank Financial Stress Index (2-day lag) | Daily | -0.07509 | -0.07578 | 0.4528 |
| 9 | Cleveland Federal Reserve Bank Financial Stress Index (5-day lag) | Daily | -0.13307 | -0.13621 | 0.459302 |

In Table 4.7, the model includes 3 daily factors, 1 weekly factor and 2 monthly factors to explain the movements of the dependent factor; while, the model also includes 3 lag variables (1-day, 3-day, and 5-day lags) from dependent factor to remove the autocorrelation in the residual. This model is significant by F-test using bootstrapping resampling. We discuss each factor that explains the dependent factor into four groups; they are capital market, interest rate, real estate market and price level of commodity. Then, we go through the standardized regression coefficients for the 6 factors which show the relative movements to the dependent variable in terms of standard deviation.

Capital market: The movements of Standard & Poors 500 stock index (SPX) reflects the inventors' sentiments and risk premium of the capital market; S&P 500 has a broad selection of common stocks in different industries, and it is a good benchmark of the general economic activities. The index can indicate the potential distress of the economy, a rising market generally has lower distress risk, and a negative coefficient in the model indicates the effect.

In our resulting model, increasing commercial paper outstanding by non-domestic issuers may lead to higher chance of distress of US economy; it is probable as US economy may be more vulnerable to outside factors when more outstanding loan is issued by institutes which are outside the US region. The following table shows the current structure of commercial paper outstanding, and the outstanding loan by non-domestic issuers is marked by *.

| **Current structure of commercial paper outstanding:** | |
|---|---|
| Nonfinancial | • Domestic<br>• Foreign<br>• Other |
| Financial | • Domestic: It further classifies into US owned, Foreign bank parent, Foreign nonbank parent, and other<br>• Foreign: It classifies into bank and other *<br>• Other |
| Asset-backed | |
| Other | |

Interest rate: US Generic Govt 10 Year Yield can be regarded as mid-term interest rate, and it is an important factor to indicate cost of money without bearing any risk. In recent monetary operations by the US and rest of major economies, more money is circulated in the markets which leads to lower cost of money, and this cost is reaching below the inflation rate until now, and further decrease in interest rate may indicate the economic and credit problem in long run, and this correlates to higher chance of distress of the economy. Another factor is the rate paid by the fixed rate payer for an interest rate swap with maturity of ten years, negative coefficient in the model reflects lower distress level when for a potential increase of mid-term float rate in the market. Mild increasing trend of interest rate may consider US economy to be recovered in a healthy situation, and this lower the chance of distress level.

Real-estate market: Year over Year home price index reflects general price trend of real estate market, mild increasing trend of home price may consider US economy recovered in a healthy situation, and this lower the chance of economic distress level.

Price Level of products: US producer price index (PPI) or wholesale price index (WPI) reflects the movements of cost price for a country's products. An increase of the price index will lead to negative distress impact to US economy as the result of our model; this is reasonable because increasing production cost may harm business profit and demand from the market as purchasing power reduces due to production cost increased.

When analyzed the standardized regression coefficients, the financial stress index is more sensitive (by more than 36%) to the movement of US Generic Govt 10 Year Yield than the US stock market. The home price index and price levels of product costs take the third and fourth position with around 20% and 40% less than the magnitude of the US stock market factor. The last two factors, which are outstanding loan by non-domestic issuers and interest rate swap between fix and floating rate, show less significant effect, they are less than a one-fourth of the effect compared with US Generic Govt 10 Year Yield.

### 4.4.3 S & P 500 Index

The Standard & Poor's 500 (S & P 500 Index) is one of the most representative market indices for American stock market, and its movements are based on the market capitalization of the 500 largest companies listed on the NYSE or NASDAQ. S & P 500 Index is one of the most followed market indices to review the market performance and regarded as a leading indicator of the economy. We used $c = 1.6$ in initialization for the peak search algorithm. Generated optimal models were deployed as the base for factor adding and replacement to obtain the search result for models with more factors, and the table below shows the $R^2$, AIC and BIC value for the optimal models with 4 – 9 factors.

| Number of factor in model | $R^2$ | AIC value | BIC value |
|---:|---:|---:|---:|
| 4 | 69.84% | 20155.02 | 13306.40 |
| 5 | 72.58% | 20062.38 | 13218.67 |
| 6 | 75.31% | 19959.91 | 13121.10 |
| 7 | 76.19% | 19925.61 | 13091.70 |
| 8 | 78.28% | 19836.36 | 13007.36 |
| 9 | 78.68% | 19820.04 | 12995.93 |

We found that AIC and BIC value tends to include more factors by taking the minimum value, but the increase in $R^2$ is quite small indeed; for example, there is only 0.4% improvement in $R^2$ for a 9-factor model. Finally, we chose a 6-factor model with $R^2$ increase by more than 1% for further analysis. The regression analysis result is listed in Table 4.8, and residual analysis is shown in Figure 4.5.

**Table 4.8 Regression analysis of S & P 500 index**

R-squared / Adjusted R-squared: 75.31% / 75.16%

VIF: 1.9343 (6 factors)

Sample standard deviation: 45.43177

| # | Macro Factors | Sampling | Coefficient Value | Standardized Coefficient | Sample Standard Deviation |
|---|---|---|---|---|---|
| 1 | ICI Retail Money Market Funds Government Treasury & Repo Total Net Assets | Weekly | -0.05023 | -0.56984 | 515.4244 |
| 2 | Mortgage Bankers Total Points 30-Yr Jumbo Effective Rate | Weekly | 261.3679 | 0.452892 | 0.078723 |
| 3 | MFG+TRD INV/SALES RATIO MANUFACT | Monthly | -25336.6 | -0.21488 | 0.000385 |
| 4 | USDA WASDE Supply & Use US Corn Production/Million Bushels | Monthly | -0.34094 | -0.28493 | 37.96834 |
| 5 | Small Dom Chartered Comm Banks Other Securities Mortgage-Backed Securities NSA | Weekly | 113.3819 | 0.195858 | 0.07848 |
| 6 | Private Housing Authorized by Bldg Permits by Type Total SAAR | Monthly | -5.73284 | -0.17805 | 1.411022 |

According to the residual analysis in Figure 4.5, residual shows significant autocorrelation and slightly non-normal property. This is further supported by Kolmogorov-Smirnov test rejects normal distribution of the time series at over 99% significant, and Durbin-Watson test rejects the null hypothesis of no autocorrelation in the time series at over 99% significant.

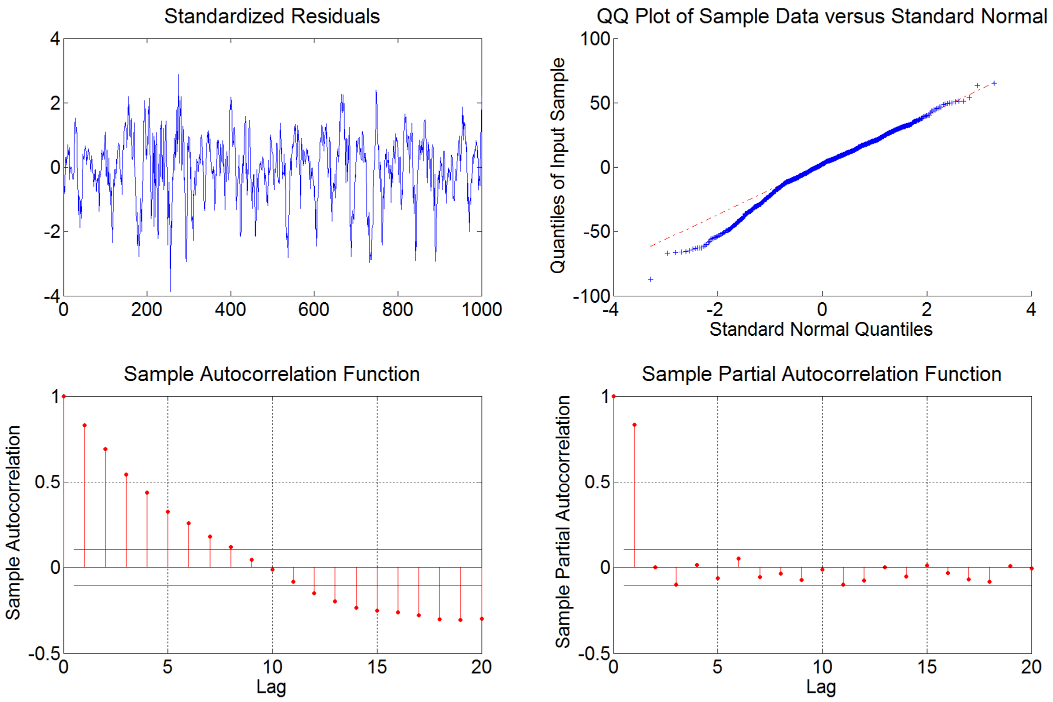

**Figure 4.5 Residual analysis for regression on S & P 500 index**

Then, we include the lag variables in our model to remove autocorrelation property of the residual, and the final best model is the following algorithm:

$$Y_t = \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \hat{b}_3 X_{3t} + \hat{b}_4 X_{4t} + \hat{b}_5 X_{5t} + \hat{b}_6 X_{6t} + \hat{b}_7 Y_{t-1} + \hat{\varepsilon}_t$$

According to the residual analysis for SPX model with lag variables in Figure 4.6, residual shows no autocorrelation, and Durbin-Watson test cannot reject the null hypothesis of no autocorrelation in the time series at 90% significant; the residual still maintains the non-normal property with more volatile movements at 200-300 day in the diagram, and this period is within the Euro credit crisis started in 2010 Jan. In addition, we showed that this model is highly significant by F-test in bootstrapping resampling. The F-test result is 1712.5 which is the one percentile of the bootstrap distribution, and the critical bound is F(7,993) $\approx$ 2.66 at 99% significant; thus, this result is greater than the critical value at 99%. We summarize the regression result in Table 4.9, and magnitudes of the coefficients are the same even with lag variables included.

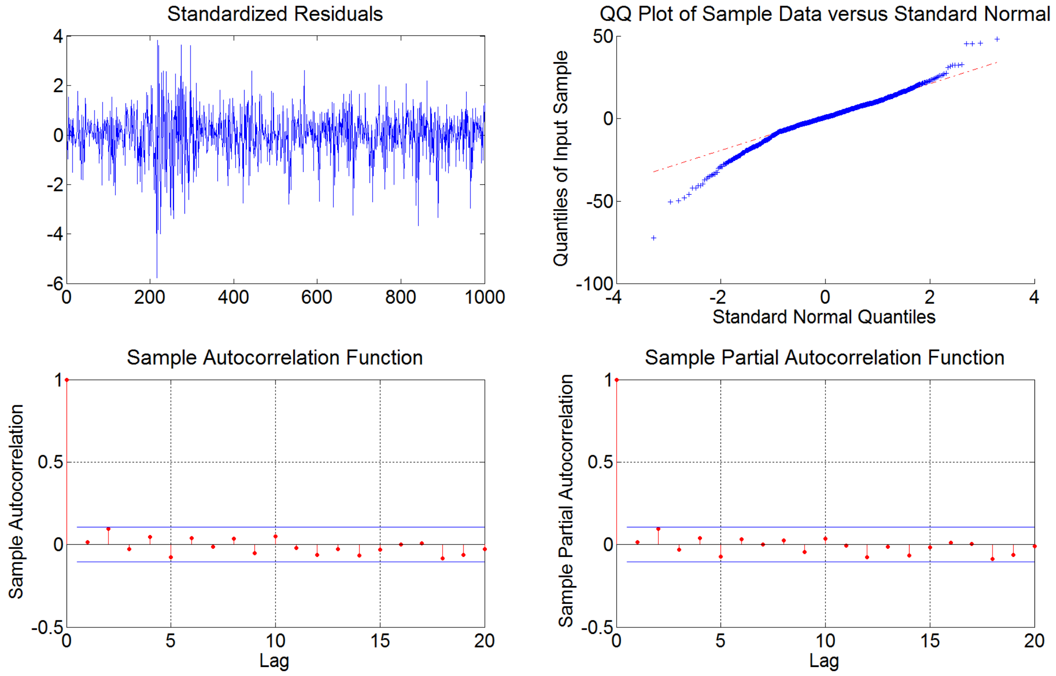

**Figure 4.6 Residual analysis for regression on S & P 500 Index with lag variable**

**Table 4.9 Regression analysis of S & P 500 index with lag variables**

R-squared / Adjusted R-squared: 92.37% / 92.32%

VIF: 4.0187 (6 factors + 1 lag factors)

Sample standard deviation: 45.43177

| # | Macro Factors | Sampling | Coefficient Value | Standardized Coefficient | Sample Standard Deviation |
|---|---|---|---|---|---|
| 1 | ICI Retail Money Market Funds Government Treasury & Repo Total Net Assets | weekly | -0.0087 | -0.09873 | 515.4244 |
| 2 | Mortgage Bankers Total Points 30-Yr Jumbo Effective Rate | weekly | 45.19413 | 0.078311 | 0.078723 |
| 3 | MFG+TRD INV/SALES RATIO MANUFACT | monthly | -4268.16 | -0.0362 | 0.000385 |
| 4 | USDA WASDE Supply & Use US Corn Production/Million Bushels | monthly | -0.06858 | -0.05732 | 37.96834 |
| 5 | Small Dom Chartered Comm Banks Other Securities Mortgage-Backed Securities NSA | weekly | 21.9474 | 0.037912 | 0.07848 |
| 6 | Private Housing Authorized by Bldg Permits by Type Total SAAR | monthly | -0.82944 | -0.02576 | 1.411022 |
| 7 | S & P 500 Index (1-day lag) | Daily | 0.827888 | 0.828721 | 45.47749 |

In Table 4.9, the model includes 3 weekly factors and 3 monthly factors to explain the movements of the dependent factor which is sampled in daily; while, the model also includes 1-day lag variable from dependent factor to remove the autocorrelation in the residual. In addition, this model is significant by F-test using bootstrapping resampling. Daily factor was not selected in the model by our search algorithm, this may be related to the high frequency components of the S & P 500 index, since they cannot correlate with the corresponding frequencies of the macroeconomic data in daily sampling. We discuss each factor that explains the dependent factor into three groups; they are interest rate & capital market, real estate market and production factor. Then, we go through the standardized regression coefficients for the 6 factors which show the relative movements to the dependent variable in terms of standard deviation.

Interest Rate & Capital Flow: The increase of 30 years effective rate of jumbo loan is regarded as the factor driving the rise of the SPX. This is favorable to equity market with increasing trend of interest rate for the risky asset, and this indicates a mild increase of the economy as a whole.

Our model shows a negative coefficient in ICI Retail Money Market Funds Government Treasury & Repo (Repurchase Agreement) Total Net Assets. When more investment goes to the money market funds, then a possible capital outflow from equity market and this gives the negative impact to the equity prices.

It is also favorable to the equity market when commercial banks hold more mortgage-backed securities, this indicates investors are more acceptable to risky assets and willing to take more risk for better return, and this can probably drive a better performance in the stock market according to the result in our model.

Real Estate Market: Building permits shows a negative coefficient in the model, that means the SPX tends to drop when more building permits is granted. This is a reasonable explanation to the dependent variable, since more buildings supply in future will have an impact on the return of rental income and properties value. This might also affect the stock prices in certain industries, such as REITs, and real estate development. Thus, SPX should be affected as these industries are some of the key constituents in SPX, and an index should reflect the price movements on wide ranges of industries.

Production Factor: Increase in inventory to sales ratio implies business inefficiency and lock up of working capital. Investors will expect profit growth to be hindrance, and riskier to the businesses as inventory cannot be transformed to cash quickly; and investors may expect higher discount factor (required rate of return) on their investment, and this leads to equity price decrease. In addition, increase in US corn production will have a negative impact to SPX index in our model, this is because more corn supply may induce lower corn price in the market, and corn is the key component to produce corn syrup in food industry, feed stock, and energy use by transforming corn to ethanol, this will affect production cost and profit in wide ranges of industries.

When analyzed the standardized regression coefficients, SPX is more sensitive (by more than 26%) to the movements of capital flow than the changes in long term interest rate. The sum of the production factors has a comparable influence against the effect of the capital flow to the SPX movements. In addition, risk appetite from financial institutes to hold riskier assets (e.g. Mortgage-Backed Securities) indicates a significant impact to SPX movements; this is around 40% of the movements by capital flow factor (net assets of money market funds). The last factor on total building permits shows less influence among the other factors in the model.

# 5. Influence of Technical Indicators on Market Trading Activities

## 5.1 Motivation

Technical analysis, it is an analysis method based on the price movements in the market to predict the future price movements, and advocates will make their investment from the results of technical analysis. Traditionally, academic field does not support technical analysis is an effective way to predict market movements or can gain an abnormal return over other strategies, such as buy-and-hold; however, recent literature by Lo and Hasanhodzic (2010) and Blume et. al. (1994) show that technical analysis can review "information" that cannot be found from the raw data in a quantitative way, as market statistics can only extract part of the information but not necessary the full picture. Lo and Wang (2009) stresses that the value of technical analysis help the investors to learn the underlying uncertainties in the economy. Investment decisions triggered by technical analysis should move transaction volume as well when investors embed this mindset in their decision process. There are some articles showing the significant relation between price movements and volume changes; Smirlock and Starks (1988) suggests a causal relationship at firm level between absolute price changes and trading volume from his empirical study on S&P 500 equities in 49 trading days. Lo and Wang (2009) concludes that volume and prices should be analyzed together in order to entail the workings of the market, as volume and prices are considered being affected by some fundamental factors, and one of the key factors may be the economic force.

According to the literature as shown, technical analysis, which extracts information from the price movements, can be value-added in building trading strategies. Moreover, trading volume also suggests taking part in the analysis, since it affects the benefits of trading strategies when selecting technical indicators. These are the motivation in this chapter to gain more specific knowledge about the individual stock characteristic by relating the technical indicators usually adopted by the investors with trading volume together. It is possible to connect these decisions to the volume transacted, as some of the investors make their buy and sell decisions by using different technical indicators, and leads to increase in trading volume. These phenomena are supported by Lo and Hasanhodzic (2000) where increasing volume tends to have a higher chance of occurrence for some repetitive patterns by using technical indicators in the price series.

In this section, we tried to find out the relations between technical patterns and transacted volume. Moreover, it is valuable to test the consistency of these relations over the same share in different markets. It is because investors in different markets may use different approaches in making investment decisions even for the same company. We applied the peak search model to relate different common technical indicators (independent factors) to transacted volume data (dependent factor), and company listed in both China (A share), and Hong Kong (H share) were included in this analysis.

## 5.2 Model Formulation

We explain the stock transacted volume (act as dependent variable) with the value of the technical indicators in consideration (act as independent variable) by using multiple linear regression model, and the model in time series is denoted as,

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \varepsilon_t \, .$$

Then, all variables should be transformed in Discrete Fourier Transform defined in (2.5) as we perform regression in frequency domain environment for peak the search algorithm, and the model becomes

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + S_\varepsilon(\omega) \, , \tag{5.1}$$

where the notations of this regression model are as follows:

$S_y(\omega)$ is the frequency spectrum for the transacted volume of stock ($y$) as the dependent variable

$S_{x_i}(\omega)$ is the frequency spectrum for the value of technical indicator ($x_i$) for stock ($y$) as independent variable

$S_\varepsilon(\omega)$ is the regression residual in terms of the frequency spectrum

$\omega$ is the pre-defined frequency range in regression modelling (where in our examples, $\omega = \{\omega_n = \frac{n}{T}, n = -\frac{T-1}{2}, \cdots, -5, -4, 4, 5, \cdots, \frac{T-1}{2}, \frac{T}{2}\}$ ; where $T$ is sample size)

$b_i$ is the regression coefficient for the independent variable $x_i$

$r$ is the total number of technical indicators (independent variables) included in the model

Lag variables will be included if regression residual shows autocorrelation, and the final model in the frequency domain becomes

$$S_y(\omega) = \sum_{i=1}^{r} b_i S_{x_i}(\omega) + \sum_{p=1}^{P} a_p S_{y,t-p}(\omega) + \sum_{q=1}^{Q} \sum_{i=1}^{r} d_{i,q} S_{x_{i,t-q}}(\omega) + S_{\varepsilon'}(\omega),$$

where $P$ is the largest time lag applied and $a_p$ is lag variable coefficients in dependent variable, and $Q$ is the largest time lag applied and $d_{i,q}$ is lag variable coefficients in independent variables. And the final model by time series can be written as,

$$y_t = \sum_{i=1}^{r} b_i x_{it} + \sum_{p=1}^{P} a_p y_{t-p} + \sum_{q=1}^{Q} \sum_{i=1}^{r} d_{i,q} x_{i,t-q} + \varepsilon_t \,.$$

Both dependent and independent variables were in daily timeframe, and we ensured the stationary of the transacted volume data in the dependent variables; thus, the analysis outcome went through the two processes as follows in order to fulfill the requirements in regression:

- Multiple Timeframes: Matching the frequency spectrums of the independent variables in different timeframes for the regression spectrums. (see section 2.3.2)
- Error Analysis: Regression residual should be free from autocorrelation by Durbin-Watson test; otherwise, lag variables will be introduced to solve this problem. (see section 2.3.3)
- Residual Normality Check: After the error analysis, we check the residual for normality by Dickey Filler test. If residual is non-normal, we use bootstrapping to ensure the model is significant in F statistics. (see section 2.3.4)

In next section, we will provide two examples to demonstrate the types of the technical indicators which have a significant impact to the transacted volume of the pair of stocks we investigated. The pair of stocks in our example is the same company but listed in different markets. This application can identify the technical indicators that affect the transacted volume of the financial assets, and understand the types of technical indicators applied by investors for the same stock traded in different markets.

## 5.3 Results & Discussion

We used raw data (without the further adjustment) with 1,000 daily sample points for dependent and independent factors, since the dependent variable is stationary in Dickey-Fuller test. We downloaded the data until 29 Aug, 2014 from the Bloomberg Terminal system. In order to avoid common peaks among independent variables in low frequency ranges, we only considered spectrum with cycles not greater than 333.33 days (i.e. starting from the fourth position of the spectrum) in our regression analysis. We used $c = -1$ in initialization for the peak search algorithm and set the parameter ($r$) ranged from 5 to 9 factors in order to choose the best model by BIC. The table below shows the $R^2$, AIC and BIC value for the optimal models with 5 – 9 factors for Tianjin Captial in A share market (600874).

| Number of factor in model | $R^2$ | AIC value | BIC value |
|---:|---:|---:|---:|
| 5 | 84.67% | 44309.06 | 37465.34 |
| 6 | 86.66% | 44172.76 | 37333.95 |
| 7 | 87.50% | 44110.05 | 37276.14 |
| 8 | 88.22% | 44052.73 | 37223.73 |
| 9 | 88.58% | 44024.13 | 37200.02 |

We found that AIC and BIC value tends to include more factors by taking the minimum value, but the increase in $R^2$ is quite small indeed; for example, there is only 0.36% improvement in $R^2$ for a 9-factor model. Finally, we chose a 6-factor model with $R^2$ increase by more than 2% for further analysis, and we chose other models based on the same criterion accordingly.

In the analysis, we took Tianjin Captial and Northeast Electric Development as the examples to compare the technical indicators selected between A shares (600874 / 000585 in Table 5.1 – 2 / 5.3 – 4 respectively) and H shares (1065 / 42 in Table 5.5 – 6 / 5.7 – 8 respectively). In the next section, we introduced the independent variables that we considered for searching, and brief description of some common technical indicators were given. Finally, we will discuss the models for each share between A and H share markets, and conclude the difference in A & H share markets based on the observation found in the models accordingly.

### 5.3.1 Technical Indicators as Independent Variables

Technical analysis focuses on recognizing the consistent patterns from historical prices in order to predict the future price movements, as practitioners have strong faith for the market prices to repeat themselves. Practitioners deploy different methods, such as statistical techniques and rule based computing to transform the price movements into the buy and sell signals for trading. These technical indicators are usually classified by the characteristic captured in the price data, they are spread (range of movements), trend and oscillation types. Technical indicators support traders to develop their strategies to exploit the repetitive patterns in the prices under different market situations; in ideal case, trader should be able to earn abnormal profit consistently after risk adjusted. (Azzopardi, 2010)

In academic field, we can interpret “market movements repeat themselves” as the existence of cycles in the price series. This supports by the Fourier transform as any time series can be represented in the combination of cycles in different frequencies. When price series is dominated by certain frequencies in the period, it should be profitable to trade by that frequency when it is significant enough.

In Table 5.1, it lists the technical indicators that our model can select into three main groups (spread, trend, and oscillation); in order to have the best model of technical indicators that influence the transacted volume, some variations in the technical indicators were included as follows:

- Set the parameters in calculating the technical indicators: It includes changes in the number of period considered in the formula, or sensitivity of the technical indicators.
- Adjustment in the output of technical indicator: the literatures show that transacted volume relates to the price movements; thus, we also include the magnitude of the movements after adjusting the value of the technical indicators by the absolute value of the first difference ($\left| f(t) - f(t-1) \right|$), and the squared value of the first difference ($(f(t) - f(t-1))^2$) respectively. Where $f(t)$ is the output value of the technical indicator at time *t*.
- Sample the technical indicators in different time frames: It includes technical indicators sampled in daily, weekly and monthly.

**Table 5.1 Technical indicators for model selection**

| **Group** | **Indicator** | **Description (Extracted from Bloomberg)** | **Parameter Setting** |
|---|---|---|---|
| Spread | Average True Range | The concept of the true range was developed by J. Welles Wilder. High / low range in a bar is extended to the previous close when it is greater than the current high, or the previous close is lower than the current low. Then a moving average of the true range is calculated. | Set the number of period from 5 to 100, and increase value by 5 each time |
| Oscillation | Bollinger Bands | The Bollinger bands gauge a security's trading activity by varying with the security's volatility rather than staying at fixed percentage intervals. | Set the number of period from 5 to 100, and increase value by 5 each time |

| **Group** | **Indicator** | **Description (Extracted from Bloomberg)** | **Parameter Setting** |
|---|---|---|---|
| Oscillation | Channel | The Channel displays an upper and lower band at the extreme prices over a specific period as well as a retracement line between the bands. | Set the number of period from 5 to 100, and increase value by 5 each time<br>Retracement % is 50% (Bloomberg default value) |
| Oscillation | Commodity Channel Indicator | The Commodity Channel Index developed by Donald Lambert measures the variation of a security's price from its statistical mean. | Set the number of period from 5 to 100, and increase value by 5 each time |
| Oscillation | Fear / Greed Index | Fear / Greed index is an oscillator based on the averaging true range that measures the ratio of buying strength to the sell strength. It indicates whether the Bulls or the Bears are in control. | For sensitivity from 2 to 20, and increase value by 1 each time |
| Oscillation | Momentum | Momentum is the difference between a current value and a value which precedes the current by the specified number of periods. It expresses the strength of the value change | • Set the number of period from 5 to 100, and increase value by 5 each time<br>• Two sets of momentum smoothing period by 5 and 10 respectively |
| Oscillation | Moving Average Convergence Divergence | The Moving Average Convergence / Divergence indicator developed by Gerald Appel determines the turning points in a trend by the difference of two exponential moving averages in specific periods. | Use default setting only: the three smoothing period as 16, 26, and 9 |
| Oscillation | Moving Average Envelopes | Moving average envelopes is the simple moving average of the prices for the specified number of periods. In addition, two banding values are calculated using the specified factors | Set the number of period from 5 to 100, and increase value by 5 each time |

| **Group** | **Indicator** | **Description (Extracted from Bloomberg)** | **Parameter Setting** |
|---|---|---|---|
| Oscillation | Moving Average Oscillator | The moving average oscillator is used to determine trend turning points. The calculation type and the moving average type in addition to the periods can be specified | Use default setting only: the three smoothing periods as 6, 36, and 9 |
| Oscillation | Parabolic Time / Price System | The parabolic time / price system developed by J. Welles Wilder follows a security's price movements in the form of a parabolic shaped pattern which provides trailing stop and reverse (SAR) prices | Use default setting only: the three parameters as 0.02, 0.02, and maximum value as 0.2 |
| Oscillation | Rate of Change | Rate of change expresses the relation between the most recent price relative to a price in the past determined by the ROC Period | Set the number of period from 5 to 100, and increase value by 5 each time |
| Oscillation | Relative Strength Index | It measures the velocity of directional price movements indicating overbought and oversold conditions. | Set the number of period from 5 to 100, and increase value by 5 each time |
| Oscillation | Stochastic | The stochastic oscillator, a momentum indicator introduced by George Lane in the 1950s, compares the closing price to its price range over a given time span. | Use default setting only:<br>TAS_K = 20<br>TAS_D = 5<br>TAS_DS = 5<br>TAS_DSS = 3 |
| Oscillation | Williams' %R | Williams %R developed by Larry Williams calculates the difference between a security's most recent closing price to its highest price in a given time period | Set the number of period from 5 to 100, and increase value by 5 each time |
| Trend | Directional Movement Indicator | The Directional Movement Indicator (DMI) developed by J. Welles Wilder quantifies the strength of a price trend. | Set the number of period from 5 to 100, and increase value by 5 each time |
| Trend | Exponential Moving Average | An exponential moving average is calculated by applying a percentage of today's closing price to yesterday's moving average value | Set the number of period from 5 to 100, and increase value by 5 each time |

| **Group** | **Indicator** | **Description (Extracted from Bloomberg)** | **Parameter Setting** |
|---|---|---|---|
| Trend | Hurst Exponent | The Hurst Exponent sometimes referred to as an index of dependence which measures the tendency of a time series to either return to the mean or in a trending direction | Set the number of period from 5 to 100, and increase value by 5 each time |
| Trend | Simple Moving Average | A simple, or arithmetic, moving average is calculated by summing the closing price of the security over a period then dividing by length of the period. | Set the number of period from 5 to 100, and increase value by 5 each time |
| Trend | Triangular Moving Average | A triangular moving average places the majority of weight on the middle portion of the price series. It is a doubly smoothed simple moving average. | Set the number of period from 5 to 100, and increase value by 5 each time |
| Trend | Variable Moving Average | A variable moving average is an exponential moving average that automatically adjusts the smoothing percentage based on the volatility of the data series. | Set the number of period from 5 to 100, and increase value by 5 each time |
| Trend | Weighted Moving Average | A weighted moving average is designed to put more weight on recent data and less weight on historical data. The weight of each price is its sequence number in the specified period. | Set the number of period from 5 to 100, and increase value by 5 each time |

As a result, there were more than 5,700 factors for selection in total. We explain the characteristic of the three main categories of technical indicators as below; this classification may help the investors to understand the nature of movements for the technical indicators in the stock market.

Spread: (Lo and Hasanhodzic, 2011)

This type of technical indicators measures the volatility of the price movements and provides indication to the movement strength. Traders can set the trading range for entry in and exit from the market, as volatility should increase when the market goes into a bullish or bearish direction. This increase shows the expectation of traders to move up or bid down the price in continuous action, while decreasing range reflects lack of interest to the assets. Average True Range is the average of the true ranges, and it is one of the spread indicators. The true range considers the closed price from current and previous trading day according to the formula as below.

$$I_t = \max((P_t^H - P_t^L), \left|P_t^H - P_{t-1}\right|, \left|P_t^L - P_t^L\right|)$$

$I_t$ is output of the true range at time *t*, where $P_t^H$, $P_t^L$, $P_t$ are the highest, lowest and closed price at time *t*. The Average True Range at time *t* is calculated by the exponential moving average of the *n* days.

$$f(n,t) = \frac{(n-1)\,f(n,t-1) + I_t}{n}, \quad \text{where} \quad f(n,0) = \frac{\sum_{t=1}^{n} I_t}{n}$$

Trending: (Lo and Hasanhodzic, 2011)

Trade on price trend is one of the common strategies deployed by traders, and trending indicators can single out the price trend by a suitable period and sampling timeframe. This technique can smooth out randomness and irregularity in the price movements, so that traders can focus on the trend component in making trading decisions. We will mention some technical indicators below for trend strategies.

Moving average is a technique to calculate a general mean level over time by using observations within a defined period, and the moving parameters highly depend on its application; this determines the cutoff point between trend and cycle, and set the weightings for the data points accordingly. The simplest form of moving average is simple moving average, and the formula is $f(n,t) = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i}$ , where $P_t$ is the current closed price and $n$ is the number of period for averaging in the time series in consideration. In simple moving average, fixed weighting applies for all data points, exponential moving average uses $\frac{2}{n+1}$ as the weighting for the latest data input, and the weighting of variable moving average includes a user defined volatility team multiplied to the weightings for exponential moving average, and the formula is as to follow.

Exponential Moving Average: $f(n,t)=\frac{2}{n+1}(P_{t-i}-f(n,t-1))+f(n,t-1)$

Variable Moving Average: $f(n,t)=\frac{2\sigma_t}{n+1}(P_{t-i}-f(n,t-1))+f(n,t-1)$

Directional Movement Indicator ( $f(n,t)$ ) is a lag indicator (trend must first established) to show the strength of the trend with value between 0 and 100; however, it does not relate to the trend momentum or direction that much. We interpret trend weakness when the value below 20, and it indicates good strength for the value at 40 or above. This indicator is a moving average of the combination by two directional indicators developed by Wilder, where one indicator is positive ( $D_t^+$ ) and another one is negative ( $D_t^-$ ). It can be represented in mathematics by,

$$\text{If} \quad (P_t^H - P_{t-1}^H) > (P_{t-1}^L - P_t^L), \quad D_t^+ = \max((P_t^H - P_{t-1}^H), 0), \quad else \quad D_t^+ = 0$$

$$\text{If} \quad (P_{t-1}^L - P_t^L) > (P_t^H - P_{t-1}^H), \quad D_t^- = \max((P_{t-1}^L - P_t^L), 0), \quad else \quad D_t^- = 0$$

$$f(n,t)=\frac{\sum_{i=0}^{n-1}(D_{t-i}^+ - D_{t-1}^-)}{\sum_{i=0}^{n-1}(D_{t-i}^+ + D_{t-1}^-)}.$$

Oscillations: (Siegel, et al., 2014; Lo and Hasanhodzic, 2011)

Oscillator is a type of technical indicators which likely to be fluctuated up or down within a mean value which changes from time to time. When oscillator reaches an extreme level, this usually indicates overbought or oversold during that period, and price reversal is expected. In practice, traders use centered oscillators for price momentum, for example, channel, commodity channel indicator and moving average convergence divergence; and use banded oscillators to review the overbought and oversold levels, such as Bollinger Bands, and Relative Strength Index.

Channel: Price channels display in three lines, where upper channel is a line drawn by the highest price over the *n* period, and lower channel is a line drawn from the lowest price over the *n* period. In between the upper and lower channel, it is a centerline of the indicator. This indicator can locate overbought or oversold levels in the trending phase when price approaching upper / lower channel. The formulas of the three lines are as follows,

Upper Channel Line: $I^{+}(n,t) = \max(P^{H}_{t-(n-1)}, P^{H}_{t-(n-2)}, \ldots, P^{H}_{t-2}, P^{H}_{t-1}, P^{H}_{t})$

Lower Channel Line: $I^{-}(n,t) = \max(P^{L}_{t-(n-1)}, P^{L}_{t-(n-2)}, \ldots, P^{L}_{t-2}, P^{L}_{t-1}, P^{L}_{t})$

Centerline: $f(n,t) = \dfrac{I^{+}_{t} + I^{-}_{t}}{2}$

Bollinger Bands: It relates price evolution to the volatility, and uses three lines to indicate this development by providing a volatility band above and below the mean prices. A centerline, which regards as the mean price over time, calculates by using a moving average. The band is constructed from rolling standard deviation of the price series multiplied by a pre-defined scalar; thus, the band will widen when market movements become vigorous, and narrower if otherwise. Trader can estimate the chance of overbought and oversold from the Bollinger Bands and the price movements. The mathematical representation is as follows, Standard Deviation: $\sigma(n,t)=\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(P_{t-i}-\overline{P}_t)^2}$, where $\overline{P}_t=\frac{1}{n}\sum_{j=0}^{n-1}P_{t-j}$

Middle Band: $\overline{P}_t=\frac{1}{n}\sum_{j=0}^{n-1}P_{t-j}$

Upper Band: $f^+(n,k,t)=\overline{P}_t+k\sigma(n,t)$

Lower Band: $f^-(n,k,t)=\overline{P}_t-k\sigma(n,t)$ .

Moving Average Convergence Divergence (MACD): It is a momentum oscillator by subtracting two moving averages with different smoothing periods. A default value setting is usually 12, 26 and 9, but traders can fine-tune this setting based on the development of the trend. Traders can identify the potential turning point from the trend when MACD histogram crossovers the zero line. The formula can be summarized as,

Exponential Moving Average: $f(n,t)=\frac{2}{n+1}(P_{t-i}-f(n,t-1))+f(n,t-1)$

MACD Line: $g(n_S,n_L,t)=f(n_S,t)-f(n_L,t)$

Signal Line: $h(n_S,n_L,n_E,t)=\frac{2}{n_E+1}(g(n_S,n_L,t)-g(n_S,n_L,t-1))+g(n_S,n_L,t-1)$

MACD Histogram: $\Delta_t=g(n_S,n_L,t)-h(n_S,n_L,n_E,t)$

Relative Strength Index (RSI: $f(n,t)$): It is a momentum indicator to calculate the ratio between the average price with positive change and the average price with negative change over a pre-defined period. This indicator is useful for confirming the overbought and oversold conditions, and the following is the formula for RSI,

$$P^U := \{P_{t-i} \mid P_{t-i} - P_{t-(i+1)} \geq 0 \quad for \quad i = 0,\ldots,n-1\}$$

$\bar{P}^U = \frac{\sum_{i=1}^{n_U} P_i^U}{n_U}$, where $n_U$ is number of elements in $P^U$, and $\bar{P}^D = \frac{\sum_{i=0}^{n-1} P_{t-i} - \sum_{i=1}^{n_U} P_i^U}{n - n_U}$.

Then, $f(n,t) = 100 - \frac{100}{1 + \frac{\bar{P}^U}{\bar{P}^D}}$.

Commodity Channel Indicator (CCI: $f(n,t)$ ): It measures the spread for price deviation between the average daily price and the average price level. Traders can use this indicator to determine overbought (oversold) region when CCI goes above (below) the average during price increase (decrease). CCI is calculated by the formula as below,

Typical Price ($I_t$): $I_t = \frac{P_t^H + P_t^L + P_t}{3}$

Standard (Mean) Deviation: $\sigma(n,t) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(P_{t-i} - \bar{P}_t)^2}$, where $\quad \bar{P}_t = \frac{1}{n}\sum_{j=0}^{n-1} P_{t-j}$

$$f(n,t) = \frac{(I_t - \frac{1}{n}\sum_{i=0}^{n-1} I_{t-i})}{0.015\sigma(n,t)}$$

Traders usually take 0.015 as the multiplier to the mean deviation, so that CCI values would be ranged between -100 and +100 for most of the time (approximately 70 - 80%).

### 5.3.2 Regression Analysis of Tianjin Capital – H share (Code: 1065)

**Table 5.2 Regression analysis of Tianjin Capital – H share (1065)**

R-squared / Adjusted R-squared: 85.05 % / 85.01%

VIF: 2.508 (4 factors)

Sample standard deviation: 4,564,890

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---|---|
| Variable Moving Average by 25-day period (Daily) in squared value of the first difference | 0.26836 | 0.001149 |
| Average True Range by 5-day period (Daily) | 0.25301 | 0.048354 |
| Average True Range by 100-day period (Daily) in absolute value of the first difference | 0.29370 | 0.001097 |
| Channel (Period Max) by 40-day period (Daily) in absolute value of the first difference | 0.35596 | 0.038381 |

According to the residual analysis in Figure 5.2, residual shows slightly autocorrelation in the first three lags of the PACF plot, and extreme value in some samples with large volume transaction. This is further supported by Kolmogorov-Smirnov test rejects normal distribution of the time series at over 99% significant, and Durbin-Watson test rejects the null hypothesis of no autocorrelation in the time series at over 99% significant.

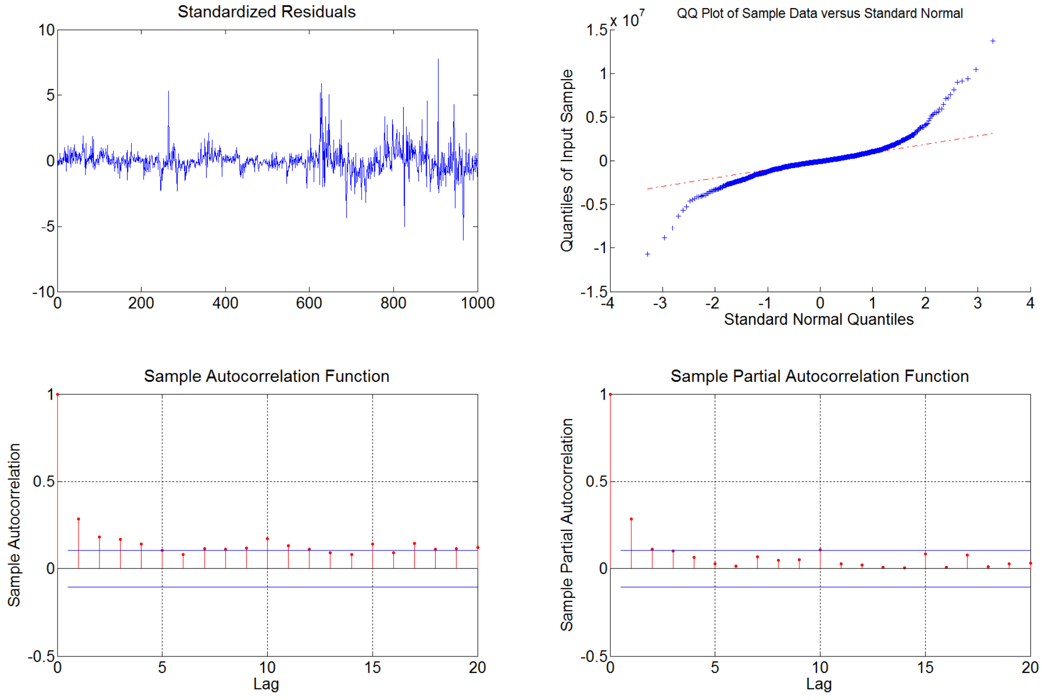

**Figure 5.1 Residual analysis for Tianjin Capital – H share (1065)**

Then, we include the lag variables in our model to remove autocorrelation property of the residual, and the final best model is the following algorithm:

$$Y_t = \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \hat{b}_3 X_{3t} + \hat{b}_4 X_{4t} + \hat{b}_5 X_{2(t-1)} + \hat{b}_6 Y_{t-1} + \hat{b}_7 Y_{t-4} + \hat{b}_8 Y_{t-6} + \hat{b}_9 Y_{t-10} + \hat{\varepsilon}_t$$

In Figure 5.2, there is no autocorrelation in the tests for residual analysis in H share model when lag variables added, and Durbin-Watson test cannot reject the null hypothesis of no autocorrelation in the time series at over 97% significant; the residual still contains non-normal property due to extreme values by the large volume transactions in some period. In order to test the consistency of the model, we use bootstrapping by case resampling to run the F-test. The F-test result is 522.8 which is the one percentile of the bootstrap distribution, and the critical bound is F(9,991)≈ 2.43 at 99% significant, which means the result is greater than the critical value at 99%. We summarize the regression result in Table 5.3, and magnitudes of the coefficients are the same even with lag variables included.

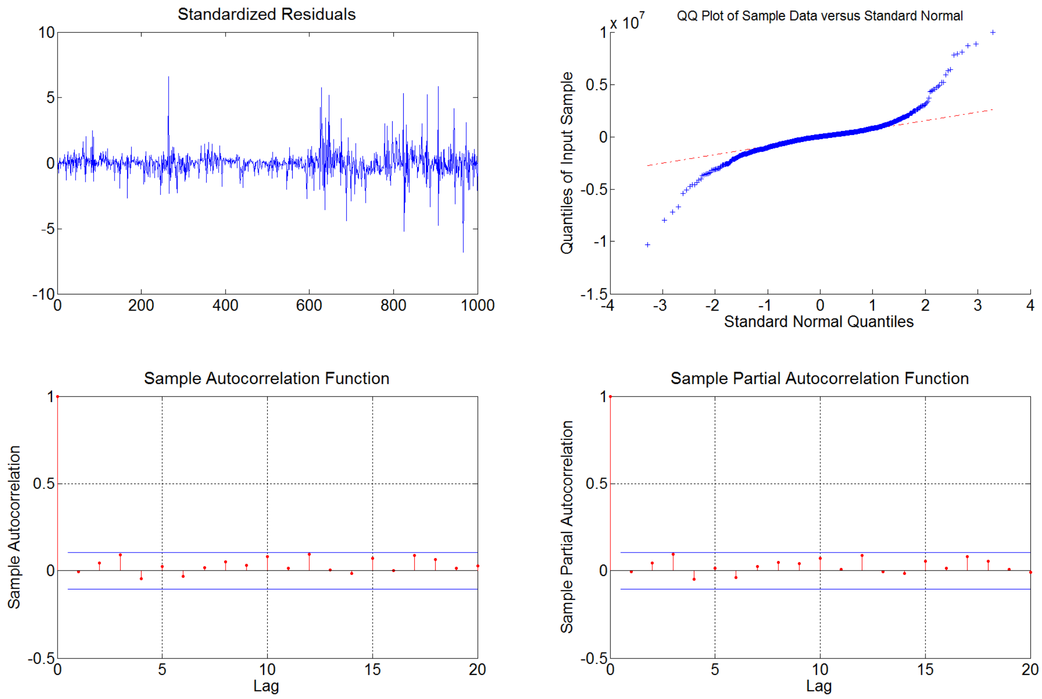

**Figure 5.2 Residual analysis with lag variables for Tianjin Capital – H share (1065)**

**Table 5.3 Regression analysis with lag variables of Tianjin Capital – H share (1065)**

R-squared / Adjusted R-squared: 89.02 % / 88.93%

VIF: 14.6907 (4 factors + 5 lag factors)

Sample standard deviation: 4,564,890

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---|---|
| Variable Moving Average by 25-day period (Daily) in squared value of the first difference | 0.13266 | 0.001149 |
| Average True Range by 5-day period (Daily) | 0.59734 | 0.048354 |
| Average True Range by 100-day period (Daily) in absolute value of the first difference | 0.19752 | 0.001097 |
| Channel (Period Max) by 40-day period (Daily) in absolute value of the first difference | 0.32759 | 0.038381 |
| Average True Range by 5-day period (Daily) – 1-day lag | -0.52665 | 0.04833 |
| Trading Volume – 1-day lag | 0.30683 | 4,565,845 |
| Trading Volume – 4-day lag | 0.07751 | 4,567,327 |
| Trading Volume – 6-day lag | 0.02475 | 4,568,584 |
| Trading Volume – 10-day lag | 0.04902 | 4,570,761 |

In Table 5.3, the model includes four factors in daily timeframe to explain the movements of the dependent factor with $R^2$ over 80%; meanwhile, five lag variables (1-day, 4-day, 6-day and 10-day lags) from dependent factor, and 1-day lag variable from one of the independent variables were included to remove the autocorrelation in the residual. This model is significant by F-test using bootstrapping resampling. All factors in the model have positive coefficients; average true range shows the highest impact in the model, then Channel with period in maximum by its absolute value of the first difference also has a significant impact to the volume transacted. Except the Average True Range uses raw data in the model, other factors have adjustment in their data by the squared or the absolute value of their first difference.

### 5.3.3 Regression Analysis of Tianjin Capital – A share (Code: 600874)

**Table 5.4 Regression analysis of Tianjin Capital – A share (600874)**

R-squared / Adjusted R-squared: 86.66 % / 86.59 %

VIF: 4.9886 (6 factors)

Sample standard deviation: 11,891,316

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---|---|
| Relative Strength Index by 100-day period (Daily) | 0.65045 | 3.010126 |
| Exponential Moving Average by 5-day period (Daily) in absolute value of the first difference | 0.51527 | 0.056001 |
| Fear / Greed Index by 2-day period (Daily) | 0.21277 | 0.150592 |
| Weighted Moving Average by 20-day period (Weekly) | -0.26600 | 0.097775 |
| Average True Range by 5-day period (Daily) | 0.25064 | 0.104146 |
| Moving Average Convergence Divergence by 12-day & 26-day periods (Daily) in absolute value of the first difference | -0.27065 | 0.015865 |

According to the residual analysis in Figure 5.3, residual shows slightly autocorrelation in the first and third lags of the ACF and the PACF plots, and extreme values for the large transacted volume. This is further supported by Kolmogorov-Smirnov test rejects normal distribution of the time series at over 99% significant, and Durbin-Watson test rejects the null hypothesis of no autocorrelation in the time series at over 99% significant.

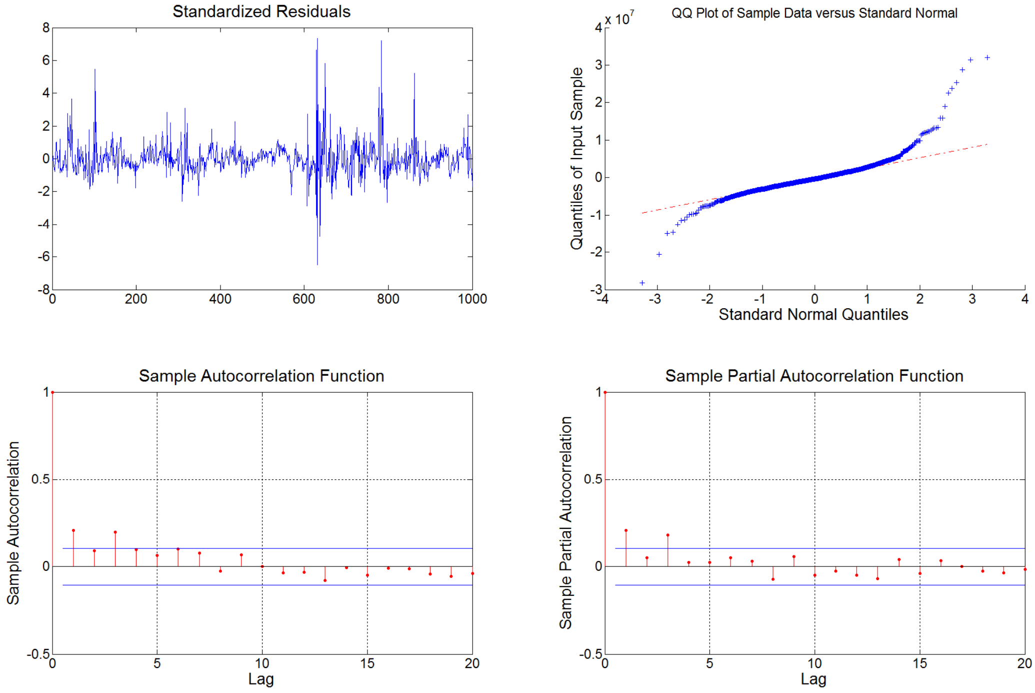

**Figure 5.3 Residual analysis for Tianjin Capital – A share (600874)**

Then, we include the lag variables in our model to remove autocorrelation property of the residual, and the final best model is the following algorithm:

$$Y_t = \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \hat{b}_3 X_{3t} + \hat{b}_4 X_{4t} + \hat{b}_5 X_{5t} + \hat{b}_6 X_{6t} + \hat{b}_7 X_{5(t-1)} + \hat{b}_8 X_{3(t-2)} + \hat{b}_9 Y_{t-1} + \hat{b}_{10} Y_{t-3} + \hat{\varepsilon}_t$$

As shown in Figure 5.4, residual analysis shows no autocorrelation for the A share model with lag variables, and Durbin-Watson test cannot reject the null hypothesis of no autocorrelation in the time series, as significant level to reject the null hypothesis is lower than 90%; the residual still contains non-normal property due to extreme values by the large volume transactions in some period. In order to test the consistency of the model, we use bootstrapping by case resampling to run the F-test. The F-test result is 624.8 which is the one percentile of the bootstrap distribution, and the critical bound is F(10,990)≈2.34 at 99% significant; thus, this result is greater than the critical value at 99%. We summarize the regression result in Table 5.5, and magnitudes of the coefficients are the same even with lag variables included.

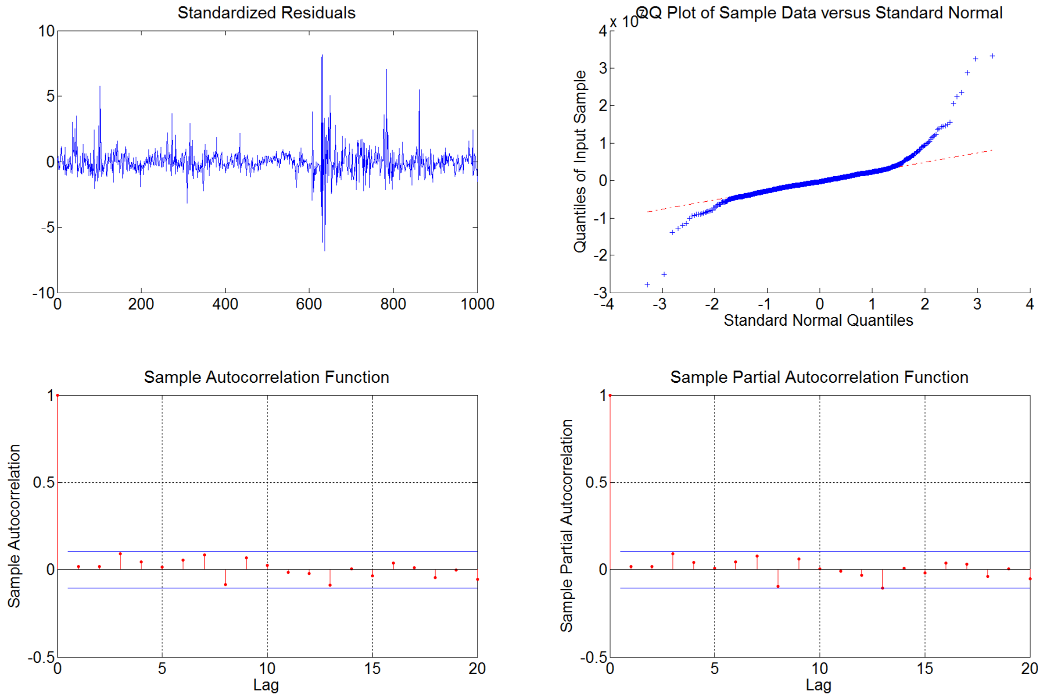

**Figure 5.4 Residual analysis with lag variables for Tianjin Capital – A share (600874)**

**Table 5.5 Regression analysis with lag variables of Tianjin Capital – A share (600874)**

R-squared / Adjusted R-squared: 88.24 % / 88.13 %

VIF: 12.4698 (6 factors + 4 lag variables)

Sample standard deviation: 11,891,316

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---|---|
| Relative Strength Index by 100-day period (Daily) | 0.49561 | 3.010126 |
| Exponential Moving Average by 5-day period (Daily) in absolute value of the first difference | 0.37509 | 0.056001 |
| Fear / Greed Index by 2-day period (Daily) | 0.21855 | 0.150592 |
| Weighted Moving Average by 20-day period (Weekly) | -0.19726 | 0.097775 |
| Average True Range by 5-day period (Daily) | 0.53971 | 0.104146 |
| Moving Average Convergence Divergence by 12-day & 26-day periods (Daily) in absolute value of the first difference | -0.20173 | 0.015865 |
| Average True Range by 5-day period (Daily) – 1-day lag | -0.39509 | 0.104151 |
| Fear / Greed Index by 2-day period (Daily) – 2-day lag | -0.05289 | 0.150588 |
| Trading Volume – 1-day lag | 0.17837 | 11,891,087 |
| Trading Volume – 3-day lag | 0.10071 | 11,891,475 |

In Table 5.5, the model includes six factors to explain the movements of the dependent factor with $R^2$ over 88%, most of the factors are daily timeframe and only one factor in weekly timeframe; two lag variables (1-day and 3-day lags) from the dependent variable and two lag variables (1-day and 2-day lags respectively) from the independent variables were included to remove the autocorrelation in the residual. This model is significant by the F-test using bootstrapping resampling. Except for Exponential Moving Average and MACD indicators which use the absolute values of their first differences, the other factors use raw data directly in the model. The Average True Range and Relative Strength Index show the greatest effect to the transacted volume with positive coefficients, which means that the volume increases when there is a larger price spread or increasing prices. The Exponential Moving Average in the absolute value of the first difference also shows significant effect in positive correlation to the changes in the transacted volume. However, the model has two indicators with negative coefficients, namely the Weighted Moving Average and MACD by midrange smoothing; these indicators imply that divergence from the mean or upward trend in midrange should reflect less volume transacted.

In the comparison between A and H share models, we have obtained high $R^2$ (over 80%) for both models, where the A share model without lag variable shows a slightly better result. Both models mainly consist of factors in daily timeframe and only one weekly factor in the A share model. The H share model has three out of the four factors in the absolute value or the squared value, while A share only has two out of the six factors in the absolute value or the squared value. The Average True Range is the key factor in both A and H shares with positive coefficients, it is a spread strategy which explains over 60% of the total movements in the H share, and 40% for the A share; however, the two markets show different oscillation strategies with different factors. For example, Channel is a key factor in the H share with around 26% impact in the total movements; it is 40% impact in the total movements to A share for Relative Strength Index in the long period. In addition, volume transacted also reacts to the trend strategy in the A share, like Exponential Moving Average by the relatively short period. Thus, both markets have one common factor which is sensitive to the price spread in explaining the volume transacted, but volume in the A share tends to respond by the Relative Strength Index in the longer period and moving average in a relatively shorter period, while H share volume responds to the Channel Indicator.

### 5.3.4 Regression Analysis of Northeast Electric Development– H share (Code: 42)

**Table 5.6 Regression analysis of Northeast Electric Development – H share (42)**

R-squared / Adjusted R-squared: 82.89 % / 82.84%

VIF: 9.275 (4 factors)

Sample standard deviation: 5,722,741

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---|---|
| Average True Range by 100-day period (Daily) in squared value of the first difference | 1.11565 | $4.32 \times 10^{-6}$ |
| Directional Movement Indicator (Moving Average) by 85-day period (Daily) in squared value of the first difference | 0.36379 | 0.023761 |
| Commodity Channel Indicator by 20-day period (Daily) in squared value of the first difference | 0.32452 | 12,821.07 |
| Commodity Channel Indicator by 90-day period (Daily) in squared value of the first difference | -0.70535 | 12,845.17 |

According to Figure 5.5, residual shows slightly autocorrelation in the first two lags of the PACF plot, and extreme values at some samples for the large volume transaction in the residual analysis. This is further supported by Kolmogorov-Smirnov test rejects normal distribution of the time series at over 99% significant, and Durbin-Watson test rejects the null hypothesis of no autocorrelation in the time series at over 99% significant.

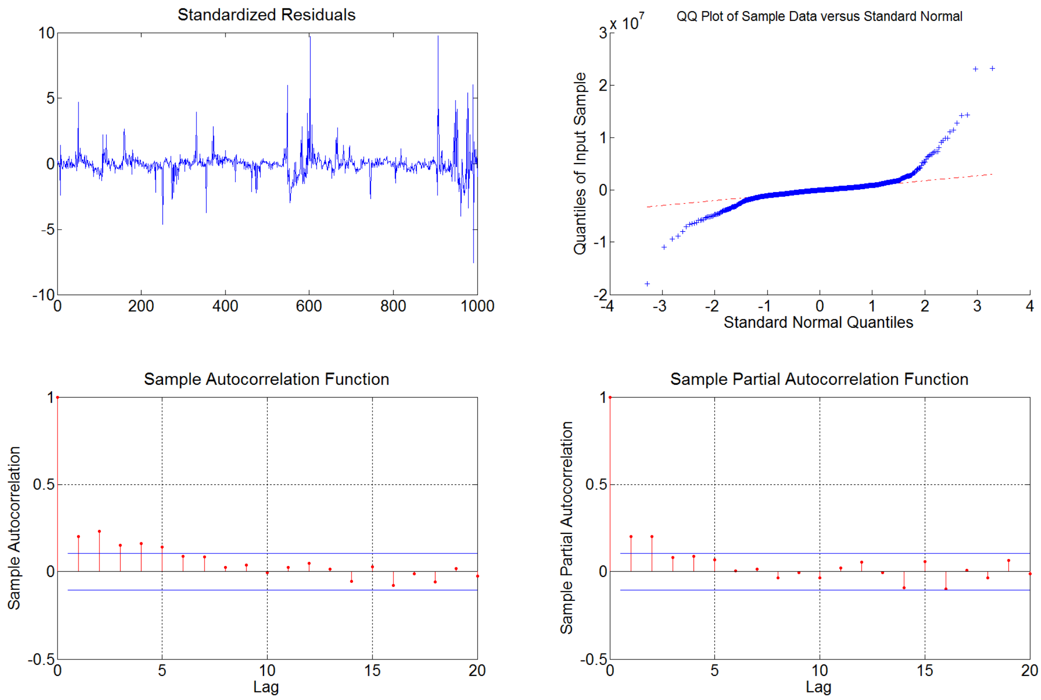

**Figure 5.5 Residual analysis for Northeast Electric Development – H share (42)**

Then, we include the lag variables in our model to remove autocorrelation property of the residual, and the final best model is the following algorithm:

$$Y_t = \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \hat{b}_3 X_{3t} + \hat{b}_4 X_{4t} + \hat{b}_5 X_{3(t-2)} + \hat{b}_6 Y_{t-1} + \hat{b}_7 Y_{t-2} + \hat{\varepsilon}_t$$

As shown in Figure 5.6, residual analysis shows no autocorrelation for the H share model with lag variables added, and Durbin-Watson test cannot reject the null hypothesis of no autocorrelation in the time series at over 97% significant; the residual still contains non-normal property due to extreme values by the large volume transactions in some period. In order to test the consistency of the model, we use bootstrapping by case resampling to run the F-test. The F-test result is 432.2 which is the one percentile of the bootstrap distribution, and the critical bound is F(7,993)≈ 2.66 at 99% significant; thus, this result is greater than the critical value at 99%. We summarize the regression result in Table 5.7, and magnitudes of the coefficients are the same even with lag variables included.

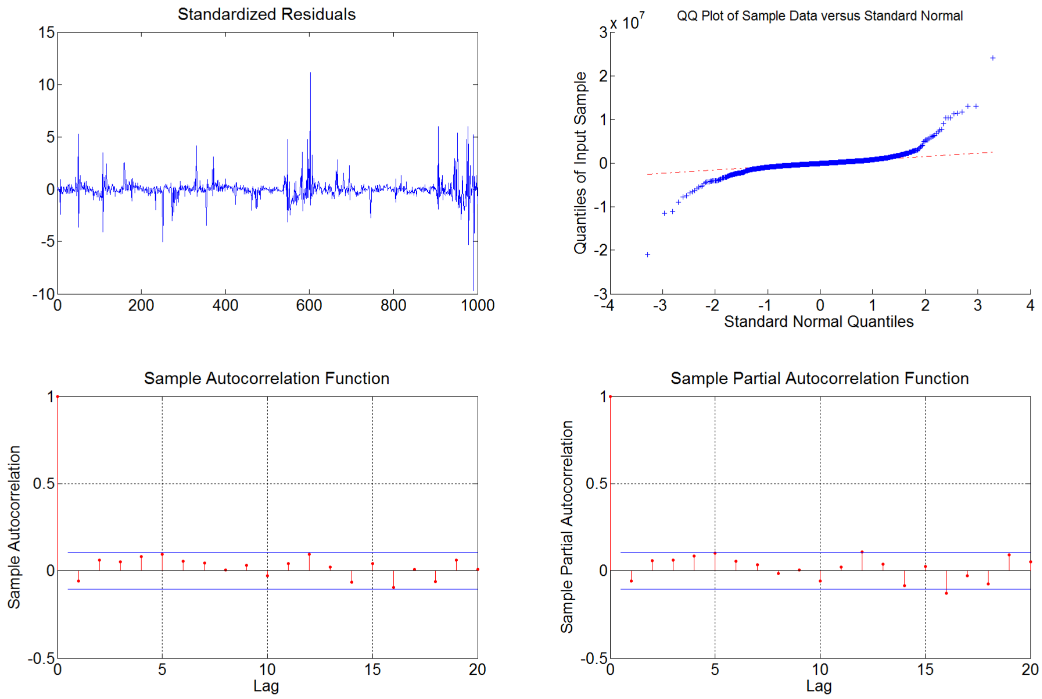

**Figure 5.6 Residual analysis with lag variables for Northeast Electric Development – H share (42)**

**Table 5.7 Regression analysis with lag variables of Northeast Electric Development – H share (42)**

R-squared / Adjusted R-squared: 85.66 % / 85.57%

VIF: 9.4277 (4 factors + 3 lag factors)

Sample standard deviation: 5,722,741

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---|---|
| Average True Range by 100-day period (Daily) in squared value of the first difference | 1.12797 | 4.32 x10$^{-6}$ |
| Directional Movement Indicator (Moving Average) by 85-day period (Daily) in squared value of the first difference | 0.22335 | 0.023761 |
| Commodity Channel Indicator by 20-day period (Daily) in squared value of the first difference | 0.28332 | 12,821.07 |
| Commodity Channel Indicator by 90-day period (Daily) in squared value of the first difference | -0.67220 | 12,845.17 |
| Commodity Channel Indicator by 20-day period (Daily) in squared value of the first difference – 2-day lag | -0.03409 | 12,821.37 |
| Trading Volume – 1-day lag | 0.18877 | 5,722,982 |
| Trading Volume – 2-day lag | 0.06485 | 5,723,845 |

In Table 5.7, the model includes 4 daily factors with adjustment by the squared value of the first difference to explain the movements of the dependent factor with R$^2$ over 85%; and 3 lag variables (1-day and 2-day lags) from the dependent factor, and 1-day lag variable from one of the independent variables were included to remove the autocorrelation in the residual. This model is significant by F-test using bootstrapping resampling.

The Average True Range shows the highest effect to the transacted volume with positive coefficient value, which means volume increase when larger price spread which regardless the direction of the movements as the indicator is the squared value of the first difference. Net effect in coefficients of the two Commodity Channel Indicators is negative, since the magnitude of the longer period is much greater than the short period one, it shows that continuous price divergence from the average price in the long period will reduce the volume transaction, and the reduction scale is much greater when the divergence becoming larger. Directional Movement Indicator shows a positive coefficient in the model, that means larger the trend strength over time (persistence of the one-side movements) should boost the volume transacted.

### 5.3.5 Regression Analysis of Northeast Electric Development – A share (Code: 000585)

**Table 5.8 Regression analysis of Northeast Electric Development – A share (000585)**

R-squared / Adjusted R-squared: 84.93 % / 84.87 %

VIF: 1.8957 (5 factors)

Sample standard deviation: 9,771,527

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---|---|
| Average True Range by 5-day period (Daily) | 0.33579 | 0.043884 |
| Directional Movement Indicator (Moving Average) by 65-day period (Daily) in squared value of the first difference | 0.33809 | 0.050256 |
| Directional Movement Indicator (Upward Trend Strength) by 100-day period (Daily) in squared value of the first difference | 0.26793 | 0.7063 |
| Bollinger Bands (Percentage - Standard deviation) by 20-day period (Daily) | 0.18438 | 0.321615 |
| Channel (Period Max) by 10-day period (Daily) in squared value of the first difference | 0.16757 | 0.012911 |

According to the residual analysis in Figure 5.7, residual shows slightly autocorrelation in the first four lags, and extreme values for the large transacted volume. This is further supported by Kolmogorov-Smirnov test rejects normal distribution of the time series at over 99% significant, and Durbin-Watson test rejects the null hypothesis of no autocorrelation in the time series at over 99% significant.

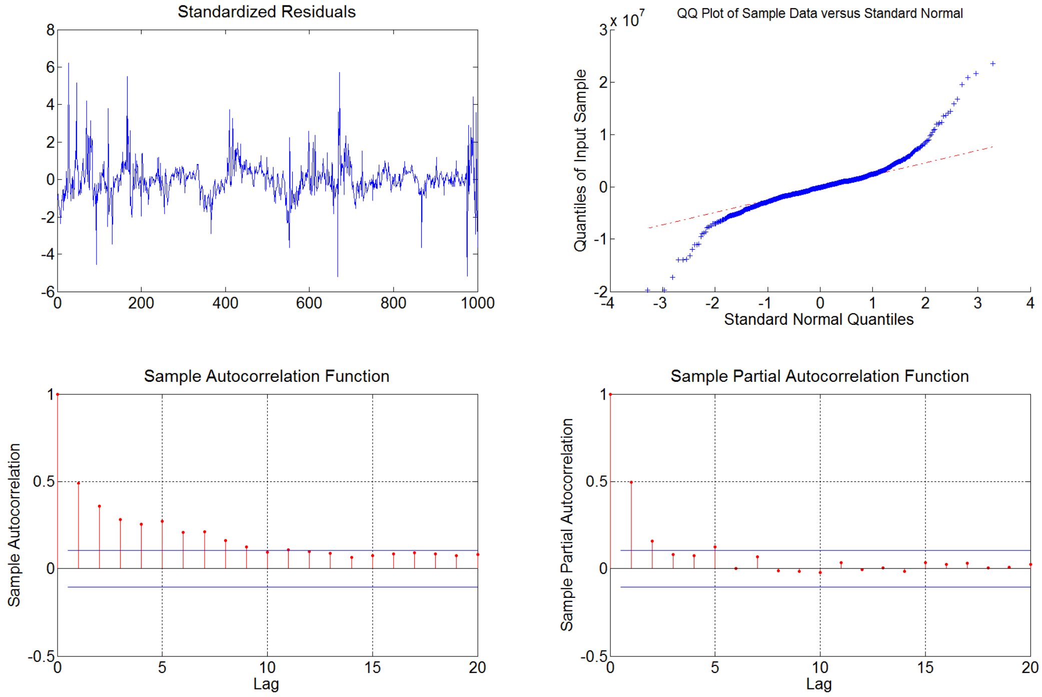

**Figure 5.7 Residual analysis for Northeast Electric Development – A share (000585)**

Then, we include the lag variables in our model to remove autocorrelation property of the residual, and the final best model is the following algorithm:

$$Y_t = \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \hat{b}_3 X_{3t} + \hat{b}_4 X_{4t} + \hat{b}_5 X_{5t} + \hat{b}_6 X_{1(t-1)} + \hat{b}_7 X_{4(t-2)} + \hat{b}_8 Y_{t-1} + \hat{b}_9 Y_{t-2} + \hat{b}_{10} Y_{t-5} + \hat{\varepsilon}_t$$

According to the residual analysis for the A share model with lag variables in Figure 5.8, residual shows no autocorrelation, and Durbin-Watson test cannot reject the null hypothesis of no autocorrelation in the time series, and its significant level is lower than 90%; the residual still contains non-normal property due to extreme values by the large volume transactions in some period. In order to test the consistency of the model, we use bootstrapping by case resampling to run the F-test. The F-test result is 626.1 which is the one percentile of the bootstrap distribution, and the critical bound is F(10,990) ≈ 2.34 at 99% significant; thus, this result is greater than the critical value at 99%. We summarize the regression result in Table 5.9, and magnitudes of the coefficients are the same even with lag variables included.

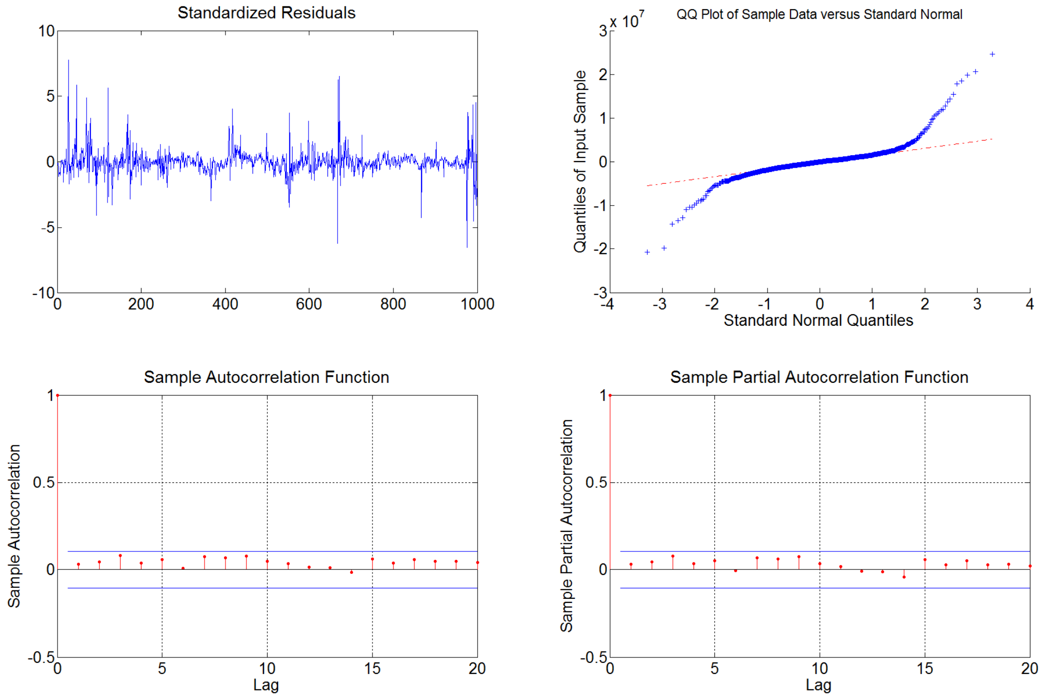

**Figure 5.8 Residual analysis with lag variables for Northeast Electric Development – A share (000585)**

**Table 5.9 Regression analysis with lag variables of Northeast Electric Development – A share (000585)**

R-squared / Adjusted R-squared: 89.55 % / 89.46 %

VIF: 8.1983 (5 factors + 5 lag variables)

Sample standard deviation: 9,771,527

| **Technical indicators** | **Standardized Coefficient** | **Sample Standard Deviation** |
|---|---:|---:|
| Average True Range by 5-day period (Daily) | 0.47275 | 0.043884 |
| Directional Movement Indicator (Moving Average) by 65-day period (Daily) in squared value of the first difference | 0.13954 | 0.050256 |
| Directional Movement Indicator (Upward Trend Strength) by 100-day period (Daily) in squared value of the first difference | 0.25043 | 0.7063 |
| Bollinger Bands (Percentage - Standard deviation) by 20-day period (Daily) | 0.10638 | 0.321615 |
| Channel (Period Max) by 10-day period (Daily) in squared value of the first difference | 0.13994 | 0.012911 |
| Average True Range by 5-day period (Daily) – lag 1 day | -0.38104 | 0.043758 |
| Bollinger Bands (Percentage - Standard deviation) by 20-day period (Daily) – 2-day lag | -0.06515 | 0.706432 |
| Trading Volume – 1-day lag | 0.38077 | 9,776,027 |
| Trading Volume – 2-day lag | 0.0952 | 9,774,327 |
| Trading Volume – 5-day lag | 0.06305 | 9,621,398 |

In Table 5.9, the model includes 5 daily factors with positive coefficients to explain the movement of the dependent factor with $R^2$ over 84%. It also includes 3 lag variables (1-day, 2-day and 5-day lags) from the dependent factor, and 2 lag variables (1-day and 2-day lags respectively) from the independent variables in order to remove the autocorrelation in the residual. This model is significant by the F-test using bootstrapping resampling. Except for the Average True Range and Bollinger Bands indicators which use raw data, the other three factors use the squared values of their first differences.

Average True Range and Directional Movement Indicators for upward trend strength show the highest effect to the transacted volume, this means the volume increases when there is a larger price spread or increasing prices. Channel by period maximum price and Bollinger Bands indicators are also included in the model, but their magnitudes are much smaller compared to other key factors. This shows that upward price movement is more significant to the increase in the transacted volume.

In the comparison between A and H share models, we obtain high $R^2$ (over 80%) for both markets, but the A share shows slightly better results for $R^2$ for models without lag variable. All factors are in the daily timeframe. Most of the factors use the squared values of their first differences, and only two factors in the A share models use raw data without any adjustment. The Average true range is the key factor in the two models with positive coefficients, and this factor uses a longer period and shows a higher magnitude (over 60% of the total movements) in the H share model when compared to the A share model (40% of the total movements). In addition, both markets show similarity in trend strategy by using the directional movement indicators. The H share model considers trend strength over a longer period that has an impact on 23% of the total movements, while the A share model considers the strength in midrange period with 35% to the total movements affected. In using the oscillation strategy, both models show that divergence from the mean price has a significant impact to explain the transacted volume, and in particular the A share model uses Bollinger Bands and Channel with positive value which affected 22% of the total movements. This implies divergence in terms of the price deviation from the mean price have a direct impact to the transacted volume. On the other hand, the H share model uses two Commodity Channel Indicators, the longer period indicator shows a negative coefficient, and the net effect is around -0.4 standard deviation to the volume transacted by 1 standard deviation of the movements to the two indicators; this means continuous price divergence from the average price in the longer period will reduce the volume transaction.

Discussion for the findings in the two markets

We found that factors in the H shares are usually with adjustment by the squared and the absolute value of the first differences, while factors in the A shares use more raw data in the models. Since the transacted volume data (dependent variable) is in the daily timeframe, this explains that the factors selected are almost from the daily timeframe for both markets. In terms of the number of factors in the models, the A share models tend to use more factors and have higher $R^2$ for models without lag variable compared to the H share models. The price spread usually shows a positive relation to the volume transacted, since Average True Range is the common factor with a positive coefficient in the models. Regarding the period covered in the indicators, we found that factors in the A share models may have a shorter period (lower than 30 days), while factors in the H share models use more midrange and longer period (longer than 50 days) for their indicators. Finally, the two markets show significant high impact by relating the price spread to the volume transacted, and the H share models show a higher impact to spread strategy with over 60% impact to the total movements in the two cases, versa the A share models have 40% impact; However, the strategies vary in different shares in terms of the indicators selected for trend and oscillation strategies. In H shares, oscillation strategies show more impact to the volume transacted in the two cases; while, the A shares do not have majority type of the strategies used besides the spread strategy.

# 6. Conclusion

This thesis examined a new search algorithm for financial time series based on frequency peak patterns of factors. Our methodology facilitates multi-factor modelling of the explained variable in matching the key patterns with the explaining variables. One major advantage of the new method is the capability of incorporating variables with different timeframes in the same model.

Using the proposed method, we carried on to study three different types of applications in finance. First, we can quantitatively relate the price series to a pool of tradable assets (factors) based on their fingerprints (key patterns). As illustrative examples, we have identified HSI and its sub-indices successfully; for mutual funds, we have matched most of the top 10 constituents disclosed in the regular reports. Second, we have developed a multi-factor model based on macroeconomic data with different timeframes to explain the economic and financial market indices. As examples, we have considered the fitted models for the US financial stress index and the SPX index and discussed the characteristics of the selected macroeconomic factors. The impact of individual factor is revealed from the signs and magnitudes of the regression coefficients. In the last application, we have demonstrated the relation between the transacted volume and technical analysis indicators of a particular share. Using A and H shares traded in Hong Kong and Mainland stock exchange markets, we picked up the influential indicators that investors might be using in their trading decisions as reflected in the transacted volume and found that the spread indicator is a significant factor in both markets. Our results support the fact that technical indicators might be using in making buy and sell decisions.

We believe the present research has made a contribution to analyze economic and financial data via a frequency domain environment. As a future extension, it is of interest to further extend the approach to analyze high frequency data, and to consider nonlinearity in the selection models.

## 7. Reference

[1] Anonymous "Cleveland Financial Stress Index [CFSI], retrieved from FRED," *Federal Reserve Bank of St. Louis*, 2016. https://fred.stlouisfed.org/series/CFSI

[2] F. Allen and D. Gale, *Comparing Financial Systems.* MIT press, 2000.

[3] P. V. Azzopardi, *Behavioural Technical Analysis.* Harriman House Limited, 2010.

[4] B. Bäth, *Spectral Analysis in Geophysics.* Elsevier, 2012.

[5] M. S. Bartlett, "Periodogram analysis and continuous spectra," *Biometrika*, vol. 37, pp. 1-16, 1950.

[6] B. Baumohl, *The Secrets of Economic Indicators: Hidden Clues to Future Economic Trends and Investment Opportunities.* FT Press, 2012.

[7] L. Blume, D. Easley and M. O'hara, "Market statistics and technical analysis: The role of volume," *The Journal of Finance*, vol. 49, pp. 153-181, 1994.

[8] D. R. Brillinger, "The digital rainbow: some history and applications of numerical spectrum analysis," *Can. J. Stat.*, vol. 21, pp. 1-19, 1993.

[9] P. Cano, E. Batle, T. Kalker and J. Haitsma, "A review of algorithms for audio fingerprinting," in *Multimedia Signal Processing, 2002 IEEE Workshop on*, 2002, pp. 169-173.

[10] N. Chen, R. Roll and S. A. Ross, "Economic forces and the stock market," *Journal of Business*, pp. 383-403, 1986.

[11] P. Chen, "Trends, shocks, persistent cycles in evolving economy: business cycle measurement in time-frequency representation," *Non-Linear Dynamics and Economics, Cambridge University Press, Cambridge*, pp. 307-331, 1996.

[12] A. D. Clare and S. H. Thomas, "MACROECONOMIC FACTORS, THE APT AND THE UK STOCKMARKET," *Journal of Business Finance & Accounting*, vol. 21, pp. 309-330, 1994.

[13] D. Corbae and S. Ouliaris, "Extracting cycles from nonstationary data," *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, pp. 167-177, 2006.

[14] M. Costa, A. Gardini and P. Paruolo, "A Reduced Rank Regression Approach to Tests of Asset Pricing," *Oxford Bull. Econ. Stat.*, vol. 59, pp. 163-181, 1997.

[15] W. L. Crum, "Cycles of rates on commercial paper," *The Review of Economic Statistics*, pp. 17-29, 1923.

[16] A. Einstein, "Methode pour la determination de valeurs statistiques d'observations concernant des grandeurs soumises a des fluctuations irregulieres," *Archives Des Sciences*, vol. 37, pp. 254-256, 1914.

[17] R. Faff and H. Chan, "A multifactor model of gold industry stock returns: evidence from the Australian equity market," *Appl. Financ. Econ.*, vol. 8, pp. 21-28, 1998.

[18] E. F. Fama and K. R. French, "Business conditions and expected returns on stocks and bonds," *J. Financ. Econ.*, vol. 25, pp. 23-49, 1989.

[19] J. Fourier, *Theorie Analytique De La Chaleur, Par M. Fourier.* Chez Firmin Didot, p{\`e}re et fils, 1822.

[20] Ghil, Michael and Allen, MR and Dettinger, MD and Ide, K and Kondrashov, D and Mann, ME and Robertson, Andrew W and Saunders, A and Tian, Y and Varadi, F and others, "Advanced spectral methods for climatic time series," *Rev. Geophys.*, vol. 40, 2002.

[21] K. Gianchandani, "A Test of Arbitrage Pricing Theory in Indian Capital Market," *Finance India*, vol. 11, pp. 353-362, 1997.

[22] C. W. Granger, "The typical spectral shape of an economic variable," *Econometrica: Journal of the Econometric Society*, pp. 150-161, 1966.

[23] B. Greenstein, "Periodogram analysis with special application to business failures in the United States, 1867-1932," *Econometrica: Journal of the Econometric Society*, pp. 170-198, 1935.

[24] M. Gresty and D. Buckwell, "Spectral analysis of tremor: understanding the results." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 53, pp. 976-981, 1990.

[25] J. Hasanhodzica and A. W. Lob, "CAN HEDGE-FUND RETURNS BE REPLICATED?: THE LINEAR CASE⋆," *Journal of Investment Management*, vol. 5, pp. 5-45, 2007.

[26] T. Hesterberg, D. S. Moore, S. Monaghan, A. Clipson and R. Epstein, "Bootstrap methods and permutation tests," *Introduction to the Practice of Statistics*, vol. 5, pp. 1-70, 2005.

[27] Hodrick, Robert J and Prescott, Edward and others, "Post-war US business cycles: An empirical investigation," 1981.

[28] A. Iacobucci, "Spectral analysis for economic time series," in *New Tools of Economic Dynamics*Anonymous Springer, 2005, pp. 203-219.

[29] C. M. Judd, G. H. McClelland and C. S. Ryan, *Data Analysis: A Model Comparison Approach.* Routledge, 2011.

[30] H. Kimura, "Sun-Spots and Faculæ On the harmonic analysis of sun-spot relative numbers," *Monthly Notices of the Royal Astronomical Society*, vol. 73, pp. 543, 1913.

[31] A. W. Lo and J. Hasanhodzic, *The Evolution of Technical Analysis: Financial Prediction from Babylonian Tablets to Bloomberg Terminals.* John Wiley & Sons, 2011.

[32] A. W. Lo and J. Hasanhodzic, *The Heretics of Finance: Conversations with Leading Practitioners of Technical Analysis.* John Wiley and Sons, 2010.

[33] A. W. Lo, H. Mamaysky and J. Wang, "Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation," *The Journal of Finance,* vol. 55, pp. 1705-1770, 2000.

[34] A. W. Lo and J. Wang, "Stock market trading volume," 2009.

[35] C. Mailhes, N. Martin, K. Sahli and G. Lejeune, "Condition monitoring using automatic spectral analysis," in *Structural Health Monitoring,* 2006, pp. 1316-1323.

[36] P. Masset, "Analysis of Financial Time-Series Using Fourier and Wavelet Methods," *Available at SSRN 1289420,* 2008.

[37] M. B. McElroy and E. Burmeister, "Arbitrage Pricing Theory as a Restricted Nonlinear Multivariate Regression Model Iterated Nonlinear Seemingly Unrelated Regression Estimates," *Journal of Business & Economic Statistics,* vol. 6, pp. 29-42, 1988.

[38] A. A. Michelson, "Determination of periodicities by the harmonic analyzer with an application to the sun-spot cycle," *Astrophys. J.*, vol. 38, pp. 268, 1913.

[39] M. Nerlove, "Spectral analysis of seasonal adjustment procedures," *Econometrica: Journal of the Econometric Society,* pp. 241-286, 1964.

[40] P. D. Praetz, "Testing for a flat spectrum on efficient market price data," *The Journal of Finance,* vol. 34, pp. 645-658, 1979.

[41] R. Priestley, "The arbitrage pricing theory, macroeconomic and financial factors, and expectations generating processes," *Journal of Banking & Finance,* vol. 20, pp. 869, 1996.

[42] S. A. Ross, "The arbitrage theory of capital asset pricing," *J. Econ. Theory,* vol. 13, pp. 341-360, 1976.

[43] S. A. Ross, "Return, risk and arbitrage," Wharton School Rodney L. White Center for Financial Research, 1973.

[44] J. G. Siegel, J. K. Shim, A. A. Qureshi and J. Brauchler, *International Encyclopedia of Technical Analysis.* Routledge, 2014.

[45] M. Smirlock and L. Starks, "An empirical analysis of the stock price-volume relationship," *Journal of Banking & Finance,* vol. 12, pp. 31-41, 1988.

[46] E. M. Tainer, *Using Economic Indicators to Improve Investment Analysis.* John Wiley & Sons, 2006.

[47] J. W. Tukey, "Use of numerical spectrum analysis in geophysics," *Bull.Internat.Inst.Statist*, vol. 41, pp. 267-307, 1966.

[48] J. W. Tukey, "The sampling theory of power spectrum estimates," in *Symposium on Applications of Autocorrelation Analysis to Physical Problems,* 1949, pp. 13-14.

[49] J. W. Tukey and R. W. Hamming, *Measuring Noise Color 1.* Bell Telephone Laboratories, 1949.

[50] G. Turhan-Sayan and S. Sayan, "Use of Time-Frequency representations in the analysis of stock market data," *Computational Methods in Decision-Making, Economics and Finance, Kluwer Applied Optimization Series,* 2002.

[51] T. Tursoy, N. Gunsel and H. Rjoub, "Macroeconomic factors, the APT and the Istanbul Stock Market," *International Research Journal of Finance and Economics ISSN,* pp. 1450-2887, 2008.

[52] Williams, Dudley H and Fleming, Ian and others, "Spectroscopic methods in organic chemistry," 1995.

[53] E. B. Wilson, "The periodogram of American business activity," *The Quarterly Journal of Economics*, pp. 375-417, 1934.

[54] P. Wilson and J. Okunev, "Spectral analysis of real estate and financial assets markets," *Journal of Property Investment & Finance*, vol. 17, pp. 61-74, 1999.