ANALYSIS OF DUAL-LISTED COMPANIES IN MAINLAND AND HONG KONG

YU BAI

M.Phil

The Hong Kong Polytechnic University

2015

The Hong Kong Polytechnic University

Department of Applied Mathematics

Analysis of dual-listed companies in Mainland and Hong Kong

Yu BAI

A thesis submitted in partial fulfilment of the requirements

for the degree of Master of Philosophy

June 2015

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)

_____BAI Yu_____(Name of student)

Dedicate to my parents.

Abstract

This thesis attempts to research price differences in Chinese segmented stock markets. Companies in China can be listed as A- share trading in Shanghai (SH) and H- share trading in Hong Kong (HK). However, the large and persistent price gaps between A- and H- share prices have been observed, raising concerns on market segmentation within China and its implications on the efficiency of price discovery. In the thesis, the factors are studied, which can be used in understanding the dissimilarity between the two markets. The topics are considered to include the use of statistical tests to perform factor analysis, and dynamic time-warping algorithm to conduct a technical analysis about the pattern of both markets.

Topic 1, the some analyses cover the main economic and company-specific factors that influence the AH Premium Index, including the fundamental, technical, and market microstructure factors. Data is collected from 50 companies listing in both markets from April 2011 to June 2014, which are grouped into three clusters by the k-means clustering technique. An appropriate factor model is built for each cluster, and statistical tools are applied to test the model. The results demonstrate that different factors can be contributed to explain the price gaps in different clusters. It is noted that large price gaps are mostly related to the shares of small market capitalization. Since the small supply of A- share and information asymmetry are disadvantages for international investors, the trading strategy in relation to small cap stocks are more likely to succeed. On the contrary, the small price gaps account

for the large market capitalization. The price gaps have been narrowing in recent years, which may pave the way to convergence after issuing the Shanghai-Hong Kong Stock Connect. Relative price convergence, but not absolute price convergence, is likely to occur Obizhaeva and Wang (2013).

Topic 2 is based on the cluster results of topic 1. In line with the results of variance test, the price differences between A- and H- share markets are significantly obvious. In other words, the price differentials in the high and low premium clusters are larger than that in the non-premium cluster. The thesis applies dynamic time-warping algorithm to fit the patterns of two stocks for the same company, using the nine examples to explain this economical phenomenon. The results present that the market factors obviously have impact on the premium cluster, and the non-premium group is influenced by some fundamental factors.

Acknowledgements

The endeavor of carrying out research is a fascinatingly non-isolated activity. I am grateful to the several individuals who have supported me in various ways during the MPhil program and would like to hereby acknowledge their assistance.

First and foremost, I wish to express my deep thanks to my chief supervisor, Dr. Cedric Yiu, for his enlightening guidance, invaluable discussions and insightful ideas throughout the years. What I have benefited most from him is the rigorous and diligent attitude to scientific research.

Furthermore, at the forefront of my MPhil experience has been the guidance and kindness of other professors, who have been a constant source of inspiration and mentorship.

Meanwhile, I would like to express my special thanks to my parents and my friends for their love, encouragement and support.

Contents

List of Figures

List of Tables

List of Notations

CAP	Information asymmetry
TUR	Turnover liquidity
V	Volume liquidity
TS	Elasticity of demand difference
EPS	Investment philosophy difference
VOL	Differential risk attitudes
ER	Exchange rate
MA	A- share market condition
MH	H- share market condition

Chapter 1

Introduction

1.1 Background

The stock market in China rapidly expanded in the early 1990s with the establishment of Shanghai and Shenzhen stock exchanges. Foreign investors were forbidden to trade on Chinese stock market in past years, because of the concern over the destabilization of the marketable capital flows. This restriction was implemented by setting up separate classes of shares for domestic Chinese and foreign residents. Based on the different locations and investor origins, it classifies into the different typies. For example, Chinese listed companies can issue domestic-listed shares (A share), foreign capital stocks are listed in China (B share), and H, N, S and L stocks refer to overseas-listed foreign shares. Therefore, the Chinese stock markets can partly stand for the segmented stock market. Over the past decade, many Chinese companies have done the capital management via overseas listings, and by issuing H- share in Hong Kong Stock Exchange, (Melvin (2003)). However, the large and persistent gaps between A- and H- share prices have been observed, raising concerns over market segmentation within China and its implication on the efficiency of price discovery. As more mainland China companies choose to adopt the dual-list model, this type of stocks is likely to command an increasing share of the markets. With the rapid development of Chinese capital market, the researchers have used differ-

ent arbitrage strategies to make money and analyze financial market movements by studying the price differences in stock segmented market. Thus, the price differences between A- and H- shares of Chinese dual-listed companies have drawn considerable attention.

Studies on the price gaps in segmented markets could date back to the 1980s. When the stock market was segmented, foreign investors could trade simultaneously in domestic and foreign stock exchanges, (Errunza and Losq (1985), Hietala (1989) and Karolyi (2006)). Empirical analysis showed that risk-free arbitrage opportunities were scarce, because self-regulating in the market mechanism would restore imbalance. A number of financial theories suggested that stock prices could reflect the intrinsic value of enterprises, (Stulz and Wasserfallen (1995)). Therefore, dual-listed shares should have the same mechanism when considering future returns and dividends, (Solnik (1974), Bergström and Tang (2001)). However, the same stock might be priced differently in separate markets due to the market segmentation. For example, empirical findings in previous studies have shown that prices of foreign shares were higher than those of domestic shares, (Bailey and Jagtiani (1994), Domowitz et al. (1997)), and this phenomenon could be identified in many stock markets including Thailand, Switzerland, Norway and Singapore. In terms of Chinese dual-listed companies, several studies focused on the A- and H- share markets. Fernald and Rogers (2002a) proposed a hypothesis on differential risk by stating that, the domestic investors had lesser chances to invest in abroad to diversify risks, and then used this hypothesis to explain the Chinese discount puzzle. Various factors could influence the A-H premium market, including the exchange rate variation, information asymmetry, company scale, liquidity and expectation of A- share holders, for more details, one could refer to Fernald and Rogers (2002b) and Lin (2004).

However, few studies have synthetically analyzed the factors which may affect the existence of the price differences in the A- and H- stock markets. The reasons

for focusing on A- and H- share markets in this thesis are as follows. First, investors analyze shares using Chinese characters. In other words, they have similar investor behavior. Because A- and H- stock markets have high turnover, it seems that they are not prone to have issues on liquidity. The recently announced Shanghai-Hong Kong Stock Connect program also allows mutual market access between the mainland China and Hong Kong. This program will enable arbitrage between the two markets, allowing the current discount of China listed A- share to close the gap with their Hong Kong listed equivalents. Moreover, the program will also increase foreign access to the companies solely listed on Chinese domestic A- share markets. This thesis uses five-year data from 2009 to draw the index existed in the line movements, and explain the performance of AH Premium Index. Although the AH Premium Index rate has remained at approximately 100, an inversion phenomenon obviously exists in the entire A-H stock market in recent years. Under this condition, the questions are raised including why the price difference is narrow, who leads the premium, and what dominative factors can be traded at discount versus the H- share. The details of the index line are shown in Figure 1.1:

Figure 1.1: AH Premium Index

Figure 1.2 illustrates that almost ten industries account for the AH Premium

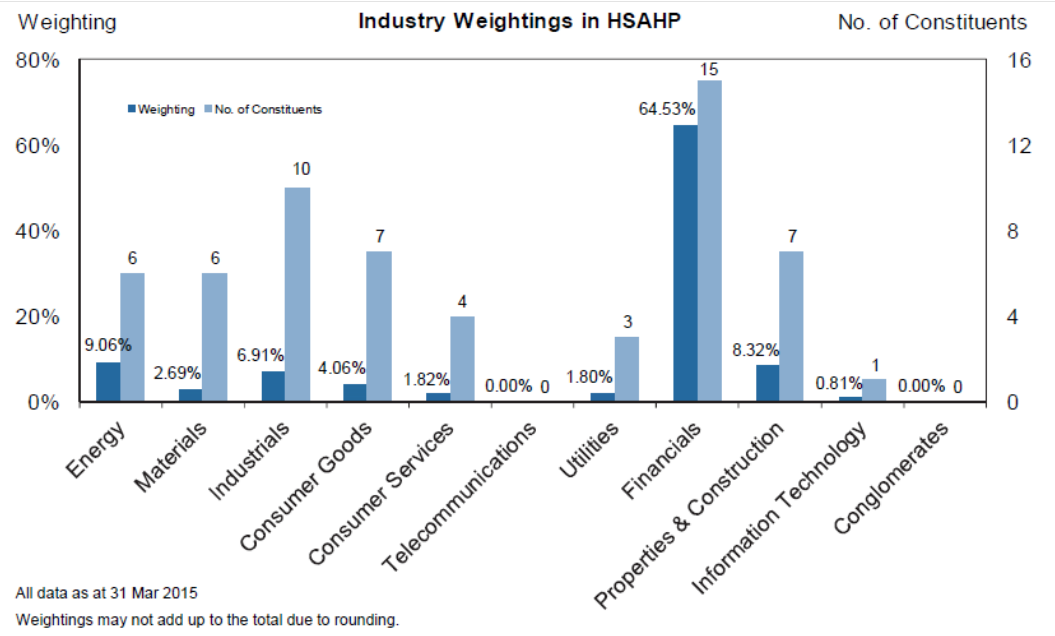

Figure 1.2: Industry weights in AH Premium Index

Index, and different industries have different weightings. The financial field has the largest weightings, whereas conglomerates and telecommunications are the least constituents, with their values being 64.53% and 0% respectively. The weighting percentage of energy, properties and construction, and industrial companies is approximately 7%. Other industries, such as materials and consumer goods, have the relatively lower weightings in the entire AH Premium Index.

This thesis aims to fill this gap. The impact of price gaps on dynamics of the corresponding A- and H- share prices is examined, by using a panel dataset of daily closing prices of A- and H- shares of 50 dual-listed companies from April 2011 to June 2014. Based on the AH Premium Index, which is calculated and provided by Bloomberg, it seems that we can divide the entire AH Premium Index into three clusters as high, low and non-premium by the k-means clustering technique. According to the traditional hypotheses on possible influential economic and company-specific factors, we conduct a regression analysis to build factor models. Then, after comparing the factor model of the entire AH Premium Index, we find a better adjusted goodness-of-fit

value when building a separate model for each cluster. Generally, information asymmetry, liquidity differences, elasticity of demands differences, investment philosophy, differential risk, exchange rate, and market performance explain price differences in both stock markets. However, we find that the market condition mainly affects the stocks in the low premium cluster, whereas, the information asymmetry and trading liquidity prevail in all three clusters. The level of price differences in A- and H- share markets is also obviously different under ANOVA test. In addition, large price gaps are noted to be associated mostly with shares of small market capitalization, and vice versa. Using the dynamic time-warping algorithm to fit the pattern, the results show that the level of premium cluster is much larger than that of the non-premium cluster, particularly in 2014.

1.2 Literature review

1.2.1 Study of market segmentation

The market segmentation exists in different countries for many years. When referring to the market segmentation, it is a marketing strategy which involves cutting a broad target market into subsets of consumers, businesses, or countries, (Solnik (1974), Sun and Tong (2000)). Many researchers have been interested in this topic for a long time. For example, Errunza and Losq (1985) examined whether the international capital market was integrated or segmented. Another research found an asset pricing theory for the segmented Chinese stock markets, focusing on A- and B- share markets, (Chen et al. (2001)). According to previous literatures, market segmentation could make the domestic shares to be traded at a discount to the foreign ones. Domowitz et al. (1997) found the differences between foreign and domestic traders in trading behavior, and demand elasticity. Yang (2003) also examined the relationship between the market segmentation and the information asymmetry in Chinese stock markets. The finding

of the paper claimed that foreign investors were more informed in emerging markets, (Amihud and Mendelson (1986), Amihud and Mendelson (1989), and Chen (1998)). However, some studies would like to research some behaviors in the A- and H- share markets, which also belonged to the segmented market. Han (2006) implied that share liquidity and corporation scale were the main factors influencing the price gaps between A- and H- share markets.

1.2.2 Study of A- and H- share markets

The Chinese economy has openned the door to oversea investors, with rapid growth. The huge capital for restructuring need to be injected by the state enterprises and private companies in China. In addition, they also need to integrate with international corporations. Because of the capital control, the stock market has been still segmented from the global stock market in the mainland China. Hong Kong has been the preferred overseas stock market for Chinese listing firms, see Chang et al. (2007). A- shares in China were traded in Shanghai and Shenzhen stock exchanges, and the foreign individual investors could not access it directly, (Mok and Hui (1998)).

Studying the price differences between A- and H- share markets is conducive to understand the economic phenomenon. According to the previous studies, it seems that information asymmetry, diversification effects, clientele bias, risk-free return differentials, and exchange rate risks have been significant factors in explaining discounts on shares, which could only be owned by foreign investors, (Su and Fleisher (1999) and Bergström and Tang (2001)). Besides, Fernald and Rogers (2002b) also researched the cross-company differences in relative prices paid by foreigners, and tried to solve the puzzles in Chinese stock market. In addition, one research indicated H- shares were trading at discount to A- shares, and the reason of that was highly correlated with market condition and illiquidity, referring to Wang and Jiang (2004). At the same time, another researcher conducted the quantitative study, which drew a

conclusion that information asymmetric was the reason of H- share discount. Finally, many studies considered the price gaps as the key point to research. For example, Li et al. (2006) stated that the price gaps were crucially affected by the risk premiums associated with the segmented stock markets. Arquette et al. (2008) examined the stock price differentials between New York, Hong Kong and Chinese securities.

1.2.3 Study of AH Premium Index

Since the most of the A- share investors are individual investors who do not have other ways to invest, the A- shares have been historically traded at significant premium to H- shares. These Chinese retail investors tend to chase speculative performance in stock market, which may lead to wild swings like the bubble of 2007. Mainland retail investors have been constrained to invest in H- shares. In the mainland, it is also not permitted to do the short-selling, there also has been no way to arbitrage by shorting the A- shares and going long the equivalent H- shares. However, the implement of the Qualified Domestic Institutional Investor (QDII) scheme prodives a chance to open the investment door.

However, the Hang Seng China AH Premium Index is a convenient way to get a handle on the divergence between the prices of A- and H- shares. The Hang Seng China AH Premium Index is a member of the Hang Seng China AH Index Series launched on 9 July 2007. It measures the absolute price divergences between the A- and H- share for mainland China companies, (Seasholes and Liu (2011)). Because of the differences of market characteristic between the mainland and Hong Kong stock markets, the dual-listed companies have been bearing the price gaps, (Han (2006) and Peng et al. (2008)). The idea of designing the AH Premium Index is to give the market with a brief, and easy-to-follow barometer to measure such price differences. It aims to calculate the weighted average premium or non-premium of the A- share prices versus the H- share prices, the stocks belong to the dual-listed companies

between the mainland and Hong Kong.

1.2.4 Study of cluster analysis

In some studies, cluster analysis is used to capture the natural structure of the data, and divide the data into sorts that are meaningful and useful. To the end, they found a new clustering method named CLARANS, based on randomized search. According to the previous researches, it had different ways to classify a set of objects into different clusters, (V.Estivill-Castro (2002)). After that, another study proposed the algorithm called hierarchical correlation ordering. The algorithm determines the cluster hierarchy, and visualizes it using correlation diagrams, referring to Achtert et al. (2006). Meanwhile, Frey and Dueck (2007) devised another method called affinity propagation, which was considered as a measurement of similarity between pairs of data points. Recently, relative clustering methods have been still applied by some researchers from different academic fields, for example, the study has stated some important cluster methods, such as k-means or expectation-maximization, which were suitable for finding ellipsoid-shaped clusters, (Kriegel et al. (2011)).

1.2.5 Study of dynamic time warping

There are a large variety of approaches to define pattern similarity measurements. Pattern similarity measurements are an essential ingredient in pattern matching. Matching has been approached in a bulk of ways, including tree pruning, (Umeyama (1993)), the generalized hough transform, (Ballard (1981)) or pose clustering, (Huttenlocher and Ullman (1987)), and geometric hashing, (Wolfson and Rigoutsos (1997)). The alignment method was referred to Huttenlocher and Ullman (1987), and the wavelet transform, see Jacobs et al. (1995). In order to overcome the inconveniences of rigid distances such as Euclidean Distance, many similarity measurements are specifically designed for the time series data. Among them, the most popular method

is probably Dynamic Time Warping (DTW), one may refer to Berndt and Clifford (1994). This distance is able to deal with transformations such as the local warping and shifting. Furthermore, it allows the comparison between series of different lengths. The dynamic time-warping algorithm is popular to be used in different industries. Sakoe and Chiba (1978) examined the spoken word recognition using DTW. Bahlmann and Burkhardt (2004) also stated that how to use DTW to complete hand writing recognition system. Another researcher was likely to use dynamic time warping to measure the similarity between the curves, referring to Efrat et al. (2007). However, in recent several years, some studies apply the dynamic time warping algorithm to measure the accuracy of compression, specially about the problem of ECG data compression, (Shorten and Burke (2011)).

1.3 Summary of contributions of the thesis

The original contributions of this thesis are considered as follows:

- This thesis emphasizes on the price differences in Chinese segmented stock markets. We analyze the main economic and company-specific factors that influence the AH Premium Index, including the fundamental, technical, and market microstructure factors. An appropriate factor model is built for it, which can be useful in understanding the dissimilarity between the two markets.

- The thesis divides the entire AH Premium Index market into different clusters, which can be solved by k-means clustering. The results show that different factors are responsible for explaining the price gaps in different clusters. For instance, the information asymmetry, trading liquidity, and market conditions are three prominent factors for different clusters.

- In addition, an observation of the stock prices between A- and H- share mar-

kets indicates the existence of obvious price differences. It seems that large price gaps are mostly associated with shares of small market capitalization, and vice versa. The thesis uses dynamic time-warping algorithm to fit the patterns of two stocks for the same company, followed by analyzing nine companies to explain similarity and dissimilarity of the patterns. This study further demonstrates that the cluster analysis of AH Premium Index is useful for understanding the characteristics of different industries from a technical angle.

1.4 Organization of the thesis

The thesis is structured as follows.

- Chap. 2 describes the model and variables. The chapter firstly shows the explained variable named A-H premium. All the data in the thesis are collected from Bloomberg, and from April 2011 to June 2014. After then, it explains the independent variables, including almost nine economical factors that can impact on the A-H premium from mathematical aspects, companying with analyzing traditional economical hypothesis. Finally, the chapter chooses the model based on the panel data, it responses the variables on different samples during multiple observations.

- Chap. 3 studies on stating k-means clustering method and dynamic time-warping algorithm. First of all, the definition of k-means clustering methodology is proposed. In addition, with the brief introduction of dynamic time-warping algorithm, a general programme can be stated and ran, which is used to show this algorithm. Finally, in order to demonstrate dynamic time-warping algorithm can be applied in matching the pattern of the time series, the thesis uses it to fit the patterns between the ADRs and H- share price by three companies.

- Chap. 4 firstly provides the result analysis about the k-means clustering, including the unit root test, model fitting, residual analysis, and ANOVA test. After that, the chapter states the results of dynamic time-warping, based on three clusters in A- and H- share markets. According to different performances about the pattern match in different clusters, the chapter analyzes the reason about it.

- Chap. 5 concludes the whole thesis and plans for the future work.

Chapter 2

Model and Variables

The purpose of this chapter is to discuss the model and variables. Firstly, we set the explained variable called A-H premium. After then the chapter states the factors which influence the price gaps in both stock markets from the financial and mathematical points. Final subsection lists the model for this economical phenomenon.

2.1 The explained variable

The explained variable in the model is A-H premium, denoted by $Y_{i,t}$. All the data in the thesis have been collected from Bloomberg, and the sample period is from April 2011 to June 2014. The study uses the same real-time intra-day data for A- and H- shares, H- share prices are converted into RMB at the daily exchange rate. Because the panel dataset is imbalanced as dates of the initial public offering of these companies are different. The selection of April 2011 as the starting point of the sample period is based on the following considerations. Firstly, the exchange rate and some policy impaction were stable during the sample period. Furthermore, some large financial firms have been listed only since 2009, and the imbalance of panel can be a greater concern if the starting point is selected to be a much earlier date. Thus, the data can be valid and reliable from 2011 excluding certain non-trading days and some companies without complete data sets, about 50 companies with 759 time series

have been collected in the sample. The details of all testing companies are shown in Tables 2.1 and 2.2.

Table 2.1: The details about testing company-1

No.	Company Name	A	H	Industry
1	CITIC Sec	600030	6030	Finance
2	Minsheng Banking	600016	1988	Finance
3	ABC	601288	1288	Finance
4	Ping An	601318	2318	Finance
5	Bankcomm	601328	3328	Finance
6	NCI	601336	1336	Finance
7	Haitong Securities	600837	6837	Finance
8	ICBC	601398	1398	Finance
9	CCB	601939	939	Finance
10	Bank of China	601988	3988	Finance
11	China Life	601628	2628	Finance
12	CITIC Bank	601998	998	Finance
13	CPIC	601601	2601	Finance
14	China Rail Cons	601186	1186	Real estate and building industry
15	China Railway	601390	390	Real estate and building industry
16	Beijing North Star	601588	588	Real estate and building industry
17	China Comm Cons	601800	1800	Real estate and building industry
18	Sinopec Corp	600028	386	Energy industry
19	Yanzhou Coal	600188	1171	Energy industry
20	China Oilfield	601808	2883	Energy industry
21	China Coal	601898	1898	Energy industry
22	China Shenhua	601088	1088	Energy industry
23	Anhui Expressway	600012	995	Service industry
24	China Shipping Development	600026	1138	Service industry
25	China South Air	600029	1055	Service industry

As a result, there are altogether 37950 trading day observations. We define the premium rate of A-share relative to H-share, one can refer to Wang and Jiang (2004), the equation is given as:

$$Y_{i,t} = \frac{P^A_{i,t} - P^H_{i,t} * ER_t}{P^H_{i,t} * ER_t}, \tag{2.1}$$

where i ($i = 1, 2, 3, .., 50$) denotes the i th company and t ($t = 1, 2, 3, .., 759$) denotes the t th day. Beside, $P^A_{i,t}$ and $P^H_{i,t}$ are the closing price of A- and H- shares at time t for company i.

Table 2.2: The details about testing company-2

No.	Company Name	A	H	Industry
26	Jiangsu Expressway	600377	177	Service industry
27	China East Air	600115	670	Service industry
28	Air China	601111	753	Service industry
29	Guangshen Railway	601333	525	Service industry
30	CSCL	601866	2866	Service industry
31	Dalian Port (PDA)	601880	2880	Service industry
32	Huaneng Power	600011	902	Public utilities
33	Huadian Power	600027	1071	Public utilities
34	Tianjin Capital Environmental Protection Group	600874	1065	Public utilities
35	Datang Power	601991	991	Public utilities
36	Tsingtao Brew	600600	168	Consumer goods
37	Shanghai Pharma	601607	2607	Consumer goods
38	GAC Group	601238	2238	Consumer goods
39	GreatWall Motor	601633	2333	Consumer goods
40	Shenji Group Kunming Machine Tool	600806	300	Industrials
41	Dongfang Elec	600875	1072	Industrials
42	Luoyang Glass	600876	1108	Industrials
43	Zhengzhou Coal Mining Machinery Group	601717	564	Industrials
44	CSR	601766	1766	Industrials
45	Jiangxi Copper	600362	358	Raw materials industry
46	Maanshan Iron & Steel	600808	323	Raw materials industry
47	CHALCO	601600	2600	Raw materials industry
48	MCC	601618	1618	Raw materials industry
49	Zijin Mining	601899	2899	Raw materials industry
50	Nanjing Panda Electronics	600775	553	Information technology

2.2 The independent variables

Some research efforts exist in understanding the factors, which contribute to the price gaps. Meanwhile, the market segmentation leads to a variety of factors that can impact on the A-H premium, (Wang and Jiang (2004)). According to the previous research about the price difference phenomenon between domestic and foreign shares, the research just used some factors to explain the price gap in other segmented stock markets, (Bergström and Tang (2001), Fernald and Rogers (2002a), Yang (2003) and Han (2006)). However, we analyze multiple potential factors here based on the previous work, improving it and analyzing the price gap in the A-H stock markets. In addition, the sample scope and time period definitely keep the price difference between A- and H- share markets. For some factors that may not be observed

directly, the proxy indicators will be applied based on the traditional analysis. They are discussed in the following subsections.

2.2.1 Information asymmetry

In the market, the larger companies attract more attention with more financial data and research reports being published, while smaller companies may be appealing to professional investors, who are more capable of finding the relevant information. However, it is argued that the domestic investors are better informed about the business prospects of the relatively small dual-listed mainland companies. It seems that the larger the company becomes, the higher stock price may grow in the market with richer capital. The total capital of the company is used to measure the scale of company operations.

To demonstrate it better, we use the concept from Chan (1993). Assuming a company issues two shares, the stocks are traded over T periods, and stock value in period t is given by:

$$V_{i,t} = V_0 + \sum_{\tau=1}^{t} \Delta V_{i,\tau}, \tag{2.2}$$

where V_0 is the constant for different stocks in the same company, and $\Delta V_{i,\tau}$ represents for the change in the value at period τ. The perturbation can be written as:

$$\Delta V_{i,\tau} = W_\tau + S_{i,\tau}, \tag{2.3}$$

where W_τ is the common information component between companies and market, following $\mathcal{N}(0, \sigma_w^2)$, whereas, $S_{i,\tau}$ is a stock specific information component and follows $\mathcal{N}(0, \sigma_s^2)$. Let $E(S_{i,\tau}, S_{j,\tau}) = 0$, $i \neq j$, W_τ and $S_{i,\tau}$ are independent. At period $t-1$, the market maker in stock i does not observe $\Delta V_{i,t-1}$ directly, but instead observes the signal $\theta_{i,t-1} = \Delta V_{i,t-1} + \varepsilon_{i,t-1}$. In the first, market makers observe the true value of $\Delta V_{i,t-1}$ at period t. The scenario is similar to the model in Admati

and Pfleiderer (1988), where signals are assumed useful for only one period. It requires that market makers observe $\Delta V_{i,t-1}$ perfectly at period t, whereas the second scenario requires only that market makers observe $\Delta V_{i,t-1}$ more precisely at period t.

Assume that at period t, the market maker observes the true values of $\Delta V_{i,\tau}$, $\tau = 1, 2, ...t-1$, and a current signal $\theta_{i,t}$.

$$\begin{aligned} P_{i,t} &= E(V_{i,t} \mid \theta_{i,t}) \\ &= V_0 + \sum_{\tau=1}^{t-1} \Delta V_{i,\tau} + E(\Delta V_{i,t} \mid \theta_{i,t}) \\ &= V_0 + \sum_{\tau=1}^{t-1} \Delta V_{i,\tau} + \frac{Cov(V_{i,t}, \theta_{i,t})}{Var(\theta_{i,t})} * \theta_{i,t} \\ &= V_0 + \sum_{\tau=1}^{t-1} \Delta V_{i,\tau} + \frac{\sigma_w^2 + \sigma_s^2}{\sigma_w^2 + \sigma_s^2 + \sigma_{i,\varepsilon}^2} * \theta_{i,t}. \end{aligned} \tag{2.4}$$

The equation is derived based on the conditional expectation of the joint normal distribution. In summary, if company i has a higher total capital than company j, then it may receive more attention, hence more information can be obtained. Therefore, the relative investment risk may be lower, i.e. $\sigma_{i,\varepsilon}^2 < \sigma_{j,\varepsilon}^2$, in turn it implies $P_{i,t} > P_{j,t}$. In other words, we can use the logarithm of a company′s total capital to indicate the information asymmetry, namely $CAP_{i,t} = \log(CAP_{i,t})$, in which $CAP_{i,t}$ is total capital of the company. This argument of the information asymmetry suggests that the trading strategy has a higher chance of success in the shares of relatively small capitalizations. The signaling effect of the A share price movements to H share price is stronger as foreign investors have the limited independent information, because of the small companies.

2.2.2 Trading liquidity

The trading liquidity is often found to be the main reason explaining the price divergence, (Silber (1991) and Amihud (2002)). If the market is with the lower liquidity, the investors need to pay for the higher transaction cost, Chen and Xiong (2001) stated this point. Due to the bid-ask spread, it is natural for investors to expect a lower price in order to compensate for the higher transaction costs. In fact, a stock is considered to be active in the market whenever there is a higher turnover or volume. Therefore, we choose two indicators to measure liquidity, namely the ratio of A- and H- share turnovers and the ratio of trading volumes, as proxies of the liquidity. In other words, the factor $TUR_{i,t} = \frac{T^A_{i,t}}{T^H_{i,t}}$ is applied, where $T^A_{i,t}$, $T^H_{i,t}$ are the turnovers of the dual-listed company, together with the factor $V_{i,t} = \frac{V^A_{i,t}}{V^H_{i,t}}$, with $V^A_{i,t}$, $V^H_{i,t}$ being the trading volumes of the A- and H- stocks for the same company.

2.2.3 Elasticity of demand difference

The different demands by investors for the stocks may lead to differentiation in prices. If demand growth rate is greater than the volatility of its prices, it can be explained by demand elasticity. The supply and demand of shares in the segmented markets can be researched by this hypothesis. The thesis employs the model of Stulz and Wasserfallen (1995) here, which demonstrates the relationship between the demand curve and stock price. Assuming:

$$I = Q_A * P_A + Q_H * P_H, \tag{2.5}$$

where I stands for the income, $P_A = P_A(Q_A)$ is the price of A stock, Q_A is the demand of A stock, while S_A is supply, satisfying $Q_A = S_A$; similarly for $Q_H = S_H$.

To maximize the income I, it can be formulated as:

$$\begin{aligned} \max_{Q_A, Q_H} \quad & I = Q_A * P_A + Q_H * P_H \\ s.t. \quad & Q_A = S_A, Q_H = S_H \\ & S = S_A + S_H, S_A > 0, S_H > 0 \end{aligned} \tag{2.6}$$

The necessary conditions are given as follows:

$$\begin{aligned} \frac{\delta I}{\delta Q_A} &= \frac{\delta P_A(Q_A) * Q_A}{\delta Q_A} + P_A(Q_A) = 0 \\ \frac{\delta I}{\delta Q_H} &= \frac{\delta P_H(Q_H) * Q_H}{\delta Q_H} + P_H(Q_H) = 0. \end{aligned} \tag{2.7}$$

Comparing the above two equations, we obtain:

$$\frac{\delta P_A(Q_A) * Q_A}{\delta Q_A} + P_A(Q_A) = \frac{\delta P_H(Q_H) * Q_H}{\delta Q_H} + P_H(Q_H). \tag{2.8}$$

It clarifies that the marginal revenues of both shares are the same, which implies:

$$MR^A = MR^H. \tag{2.9}$$

where $MR^i = \frac{\delta P_i(Q_i) * Q_i}{\delta Q_i} + P_i(Q_i)$. Assume the elasticity of demand for stock i is:

$$\varepsilon_i = -\frac{\delta Q_i}{\delta P_i(Q_i)} * \frac{P_i(Q_i)}{Q_i} \tag{2.10}$$

It is easily seen that $MR^i = P_i(1 - \frac{1}{\varepsilon_i})$. Thus based on the equation $MR^A = MR^H$, we have:

$$P_A(1 - \frac{1}{\varepsilon_A}) = P_H(1 - \frac{1}{\varepsilon_H}) \tag{2.11}$$

which implies:

$$\frac{P_A}{P_H} = \frac{1 - \frac{1}{\varepsilon_H}}{1 - \frac{1}{\varepsilon_A}} \tag{2.12}$$

Thus, the result shows that the differential demand and the scarcity of those stocks lead to pricing gaps in the segmented markets. Here, we use the number of issued shares as indicators. The amount of issued shares can be all or part of the total amount of authorized shares of a corporation. It will move the demand curve to the right and lead to a rise in share prices, when increasing the demand for shares. Generally speaking, if the amount of issued shares will be larger, trading activities may be more active and stable. The factor is $TS_{i,t} = \frac{TS^A_{i,t}}{TS^H_{i,t}}$, where $TS^A_{i,t}$ and $TS^H_{i,t}$ are the numbers of issued shares of dual-listed companies.

2.2.4 Investment philosophy difference

The philosophy of investors is quite different if they are separated geographically. It can react to the fundamental information differently, and the investors are willing to pay different levels of premium to the fair prices. Because of the different views, it can easily lead to the price differences between the A- and H- shares. Here, we use earning per share as an indicator to account for the investment philosophy difference, (Aharony et al. (2000)). This indicator serves to reflect the profitability of the company as a whole, and to rank companies within and across the categories. The factor is $EPS_{i,t} = \frac{EPS^A_{i,t}}{EPS^H_{i,t}}$, where $EPS^A_{i,t}$ and $EPS^H_{i,t}$ are earning per shares of the dual-listed companies. Thus, it is important to build the linkages between the two markets by improving the access of investors.

2.2.5 Differential risk attitudes

Simplifying risk attitudes stand for the investors and companies can receive the degree of risks, in order to arrive their goal. Thus, the different investors have different investment styles and the levels of risk aversion, which can lead to the price difference, (Fama and French (1993)). Sun and Tong (2000) said that the domestic

investors have been more likely to invest for short-term benefit. Nevertheless, the development of the A stock market is slower than that of H share market, the later is more mature. Thus, it also has some professional investors in the mature stock market, companying with H stock market, much more non-rationality investors exist in A stock market. Here, we choose the ratio of volatility as an indicator to measure risk attitude differences with the factor $VOL_{i,t} = \frac{VOL^A_{i,t}}{VOL^H_{i,t}}$.

2.2.6 Exchange rate

The volatility of exchange rate can lead to differential prices for dust-list company. H shares are trading with Hong Kong dollars which is pegged against the U.S. dollars, but the dividends are principally denominated in RMB. The pricing mechanism may need to account for this fact, and investors will require a reasonable risk spreads to cater for potential exchange rate risk. If the RMB will be increased in valuation, it can lead to the positive return by exchanging for the investors in H-share market. Meanwhile, it will be worthy to invest H-share markets. From another perspective, if the RMB continues to increase the value, which can lead to the development of the A-share demands, the price gap will be larger. Basically, the expected exchange rate will influence the price differences between the A- and H- share markets. Fernald and Rogers (2002b) demonstrated the volatility of the exchange rate influenced the H-share market. As a result, this is clearly a potential factor to cause the price differences.

2.2.7 Market conditions

As some have argued, the segmentation of the markets has been exploited by some traders, with an effect of exacerbating market condition. Meanwhile, the different locations of transaction can influence the stock price movements by systemic risk and investor sentiment. Chan et al. (2008) found that it was a sensitive relationship

between the price of stocks and market environment. In order to measure the degree of influence about diversification benefits, we calculate sensitivity about the return of A-share on CSI300 Index, and H-share on Hang Seng Index respectively. CSI300 and HSI index are comprehensive reflection about the overall condition of Chinese stock market and Hong Kong stock market respectively. We can define them as follows:

$$\begin{aligned} MA_{i,t} &= \frac{Cov(R_{CSI300,t}, R^A_{i,t})}{VAR(R_{CSI300,t})} \\ MH_{i,t} &= \frac{Cov(R_{HSI,t}, R^H_{i,t})}{VAR(R_{HSI,t})}, \end{aligned} \tag{2.13}$$

where $MA_{i,t}$ is the risk arising from exposure to CSI300 index movement of A-share, and $R_{CSI300,t}$ represents the return of CSI300 index at time t, $R^A_{i,t}$ is the return of the i th company in the A- share market at time t. The similar description for the H- share has.

2.3 Model

We have to choose models based on the panel data, becasue the data used in the thesis are panel data, (Jianping and Minken (2007)). The multiple factors are used by a financial model to explain financial market performance. Comparing the factors to analyze relationships between variables and the resulting performance is the next step. It also responses the variables on the different samples during multiple observations. The thesis also uses the regression model to analyze different phenomenon about dual-listed companies in A-H share markets. Generally speaking, it is written as:

$$Y_{i,t} = b_0 + \sum_{k=1}^{n} b_k f_{i,t} + \varepsilon_{i,t}, \tag{2.14}$$

where i stands for the company, t is the time for observation, k represents the coefficients, $f_{i,t}$ is the factor at time t for company i. The model in the thesis is:

$$\begin{aligned} Y_{i,t} = b_0 + b_1 CAP_{i,t} + b_2 TUR_{i,t} + b_3 V_{i,t} + b_4 TS_{i,t} + b_5 EPS_{i,t} + b_6 VOL_{i,t} + \\ b_7 ER_t + b_8 MA_{i,t} + b_9 MH_{i,t} + \varepsilon_{i,t}, i = 1, 2, ..., 50, t = 1, 2, ..., 759, \end{aligned} \tag{2.15}$$

where, $Y_{i,t}$ is premium or non-premium rate. $CAP_{i,t}$ is the information asymmetry; $TUR_{i,t}$ and $V_{i,t}$ stand for the liquidity; the elasticity of demand is $TS_{i,t}$; $EPS_{i,t}$ is the investment philosophy; $VOL_{i,t}$ represents the differential risks; ER_t is the exchange rate for daily; $MA_{i,t}$ and $MH_{i,t}$ are indexes for the market condition about A- and H- shares.

Chapter 3

Methodology

The main purpose of this section is to present an in-depth investigation of clustering analysis and dynamic time-warping algorithm. The traditional approach in data mining, which focuses on k-means clustering, is adopted. This section first introduce the adopted method and other descriptions on clustering analysis. And then, we discuss the dynamic time-warping algorithm, specifically, how it is applied in measuring similarity of the pattern in time series field.

3.1 Clustering

Definition 3.1 (Clustering (V.Estivill-Castro (2002), Jianping and Minken (2007))). *The aim of clustering is to group a number of objects in such manner that objects in the same cluster are more similar to each other than to those in other clusters. This technique is a statistical data analysis technique, used mainly to mine the big data.*

Clustering analysis has been extensively studied in many areas, including statistics, machine learning, pattern recognition, and image processing, (Fisher (1987), Huth et al. (2008) and Michalski and Stepp (1983)). Recent efforts in data mining have fasten on the methods for effective cluster analysis in the large databases, see Ester et al. (1996). Typical cluster models are included as follows:

1. Connectivity models, such as hierarchical clustering, in which models are built based on distance connectivity, (Achtert et al. (2006)).

2. Centroid models, such as the k-means algorithm, which represents each cluster by a single mean vector, (Vattani (2011)).

3. Distribution models, where clusters are modeled using statistical distributions, such as multivariate normal distributions, which are used by the expectation-maximization algorithm, (Anandkumar et al. (2014)).

4. Density models, such as DBSCAN and OPTICS, which define clusters as connected dense regions in the data space, one can refer to Kriegel et al. (2011).

5. Subspace models, such as bi-clustering (also known as co-clustering or two-mode-clustering), in which clusters are modeled with both cluster members and relevant attributes, (Rai and Daume (2010)).

6. Group models, of which some algorithms do not provide a refined model for their results and only provide grouping information, and;

7. Graph-based models: a clique, for example, a subset of nodes in a graph such that every two nodes in the subset connected by an edge can be considered as a prototypical form of cluster, (Koller and Friedman (2009)).

The cluster analysis is useful in understanding the dissimilarities in the entire A-H premium stock market, the different performances have been found in different clusters.

3.2 K-means clustering

Bonzo and Hermosilla (2002) used the probability link function to improve calculation of cluster analysis for wider domain, whereas Mei and Chen (2010) proposed five

methods to calculate the distance between classes in a cluster analysis. Before choosing k-means clustering method, we have compared with other cluster models, such as connectivity models, density models. However, the results of them can not give us the clear clustering performance to the research. Because the feature of connectivity models is based on the connecting point with the certain distance. In addition, the density model often is used for mining the non-convex, shape-based clustering. Thus, the aim of the present study is to analyze price differences between A- and H- share markets according to the panel data. We choose the k-means clustering method, and use the R program to run it. The procedure of k-means clustering follows a simple and easy means of classifying a given data set, through a certain number of clusters fixed a priori. The selection of the number of k is based on the cluster index, we choose the diameter as the cluster index. After running the relative program, it suggests us to set the k as equal to 3.

The main idea is to define k-centroid points, one point for each cluster. Since the different locations yield different results, these centroid points can be settled in a calculated manner. Thus, the best choice is to place them as far away from each other as possible. To take each point belonging to a given data set and associate it with the nearest centroid will be the next step. The first step is considered as completed and an early group obtained when no points are pending. At this point, the new k-centroid points need to be recalculated as barycenters of the clusters obtained in the previous step. After the new k-centroid points are obtained, a new binding has to be conducted between the same data set points and the nearest new centroid. The loop results in a step by step change in the k-centroid points are observed, when they are no more changes .

The k-means clustering is described from a mathematical perspective, because it is a method of vector quantization in data mining. We set the observations as $Y_1, Y_2, Y_3, .., Y_n$, where each observation is a real vector. The aim of k-means cluster-

ing is to partition n observations into k clusters, here it has $S = \{S_1, S_2, S_3, .., S_k\}$, $k \leqslant n$, and minimize the within-cluster sum of squares. The minimization can be calculated as follows:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{y_j \epsilon S_i} \parallel y_j - \mu_i \parallel^2, \tag{3.1}$$

where μ_i is the mean of points in S_i. Accordingly, 50 dual-listed companies are classified into three big clusters for A- and H- stock markets. Ultimately, we assume $k = 3$ based on the financial theory and cluster index. According to the above traditional hypothesis, we find a common factor called total capital of company, which has a significant impact on every stock and every company, Yang (2003), Lin (2004) and Chan et al. (2008) stated it clearly. The capital also shows the scale of a company or an industry, and it can be traded as a proxy of the company. The result of the cluster analysis shows that blue-chip shares dominate the non-premium cluster, while the low premium cluster focuses on small cap shares, and the other shares account for the high premium cluster. The levels of premium and non-premium clusters are different. Obviously, the level of the premium cluster is much higher than that of the non-premium cluster.

3.3 Dynamic time-warping algorithm

Although the definitions of similarity vary from one clustering model to another, the concept of similarity in most of these models is based on the distances, for instance, Euclidean distance or cosine distance, (Hjaltason and Samet (2003)). A number of similarity measures have been designed specifically for the time series data to overcome the inconveniences of rigid distances such as Euclidean distance. The section explores a more general type to utilize the pattern similarity to measure the distance

between two objects, i.e., dynamic time-warping algorithm, (Nandyala and Kumar (2010) and Shorten and Burke (2011)). The dynamic time-warping algorithm, which is used for detecting similar shapes with different phases, obtained its popularity by being extremely elastic in its measurement of the time series, as it minimizes the effects of shifting and distorting in time by allowing elastic transformation of the time series. Nevertheless, the representation, similarity and accuracy of the time series data mining, are quite important for improving the efficiency of the time series data mining, Leigh et al. (2002), Salvador and Chan (2007) and Tappert et al. (1990) explained it. The algorithm enables the distance to handle transformations such as local warping and shifting, as well as is allowed for the comparison of series of different lengths. The path trace is shown in Figure 3.1.

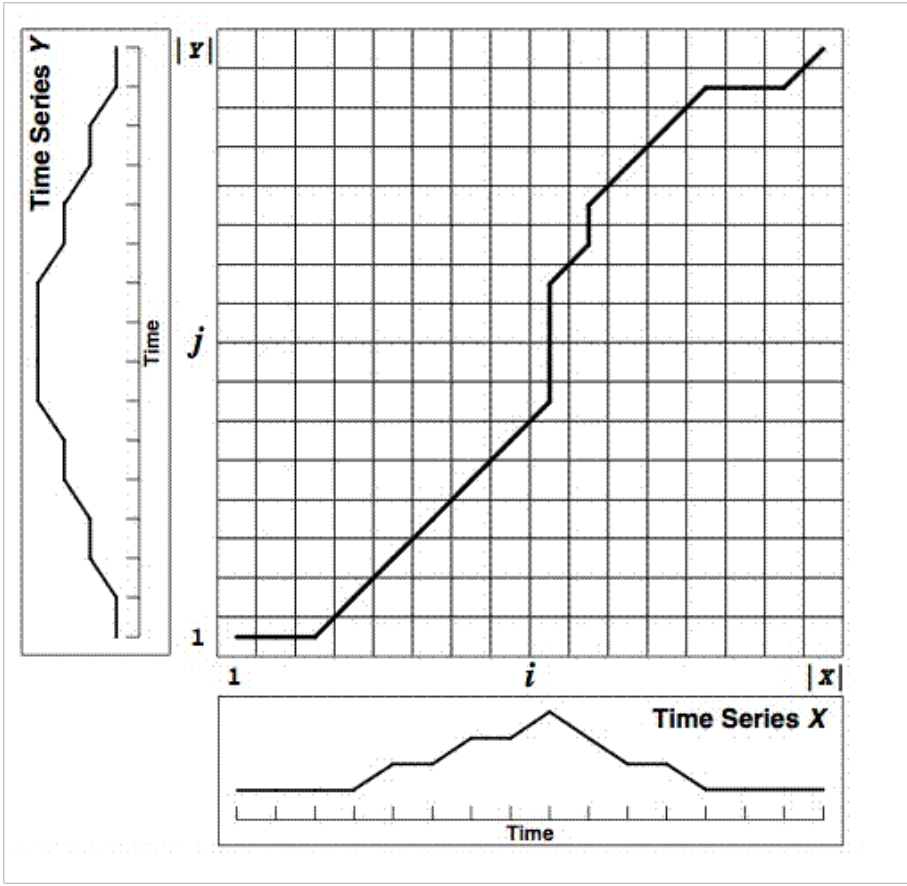

Figure 3.1: DTW-A cost matrix with the minimum distance warp path

As shown in Figure 3.1, the objective of this distance is to find the optimal alignment between two series $X = \{x_0, x_1, x_2, ..., x_{N-1}\}$ and $Y = \{y_0, y_1, y_2, ..., y_{M-1}\}$, by searching for the minimal path in a distance matrix $[D]$, which is defined a mapping between them. Each entry of the matrix $[D]$ is defined by the Euclidean distance between a pair of points (x_i, y_j). This optimization problem is subject to three restrictions, Efrat et al. (2007) stated it. The boundary condition studies the path to start in position $D(0,0)$ and to end in $D(N-1, M-1)$. The continuity condition restricte the step size, forcing the path to continue through one of the adjacent cells. Finally, the monotonicity conditions forbid the path to move backwards in the positions of the matrix. Based on this, the problem is reduced to solve the following recurrence:

$$DTW(X,Y) = C_{P*}(X,Y) = min\{C_P(X,Y), p \in P^{N*M}\} \tag{3.2}$$

where P^{N*M} is the set of all possible warping paths, building the accumulated cost matrix or global cost matrix D, which is defined as follows:

1. First row:

$$D(1,j) = \sum_{k=1}^{j} C(x_1, y_k), j \in [1, M]. \tag{3.3}$$

2. First column:

$$D(i,1) = \sum_{k=1}^{i} C(x_k, y_1), i \in [1, N]. \tag{3.4}$$

3. All other elements:

$$D(i,j) = \min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\} + C(x_i, y_j), i \in [1, N], j \in [1, M]. \tag{3.5}$$

The time cost of building this matrix is $O(NM)$, which equals the cost of the following algorithm, where X and Y are the input time series, and C is the local cost matrix representing all the pair wise distance between X and Y, see Figure 3.2.

```
1: n ← |X|
2: m ← |Y|
3: dtw[] ← new [n × m]
4: dtw(0, 0) ← 0
5: for i = 1; i ≤ n; j++ do
6:     dtw(i, 1) ← dtw(i − 1, 1) + c(i, 1)
7: end for
8: for j = 1; j ≤ m; j++ do
9:     dtw(1, j) ← dtw(1, j − 1) + c(1, j)
10: end for
11: for i = 1; i ≤ n; j++ do
12:     for j = 1; j ≤ m; j++ do
13:         dtw(i, j) ← c(i, j)+min {dtw(i − 1, j); dtw(i, j − 1); dtw(i − 1, j − 1)}
14:     end for
15: end for
16: return dtw
```

Figure 3.2: DTW-The accumulated cost matrix

However, once the accumulated cost matrix builds the warping path, which can be found by the simple backtracking from the point $P_{end} = (M, N)$ to the $P_{start} = (1, 1)$, following the greedy strategy as described by the algorithm. The research lists a simple example to run the dynamic time-warping programming, the sine and cosine functions play the two time series, see them in Figure 3.3.

After then, it is quick to use dynamic time-warping function to find the best match, Figures 3.4 and 3.5 show the result.

After applying dynamic time-warping programming, we obtain the pattern similarities in the example, (Figure 3.6).

In this section, we show the proposed measurement to explain how dynamic time warping in time series is applied. Dynamic time-warping algorithm can measure the distance between two time series, such as the prices of both the A- and H-

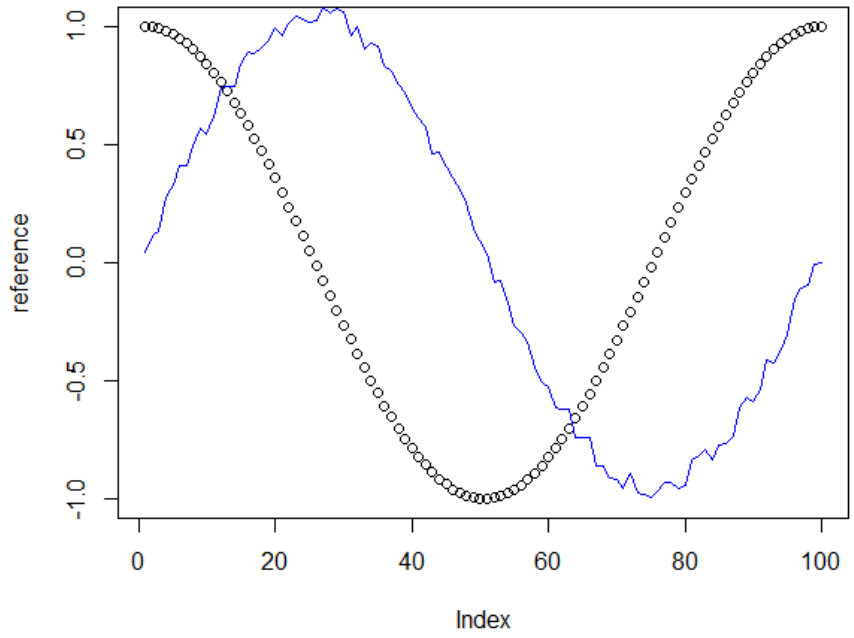

Figure 3.3: DTW-Sine and cosine

share for the same company. For the real-time trading, the distance result shows the similarities from the technical analysis. The experimental results show that the approach can effectively match the pattern of the time series, compare and analyze the dissimilarities.

In order to demonstrate the dynamic time-warping algorithm is effectively useful to fit the pattern about the time series later, the study measures the distance between the ADRs and H- share price for the same company. Because the patterns of the same company should be similar based on the empirical analysis, (Kim et al. (2000)). In addition, the globalization is not solution of barriers to trade, the customs and values can be integrated by tendency of the world's businesses. It is relatively simple to invest money in the local stock market. On the contrary, to invest a company listed on a foreign exchange is much more difficult. However, there is an easy way around

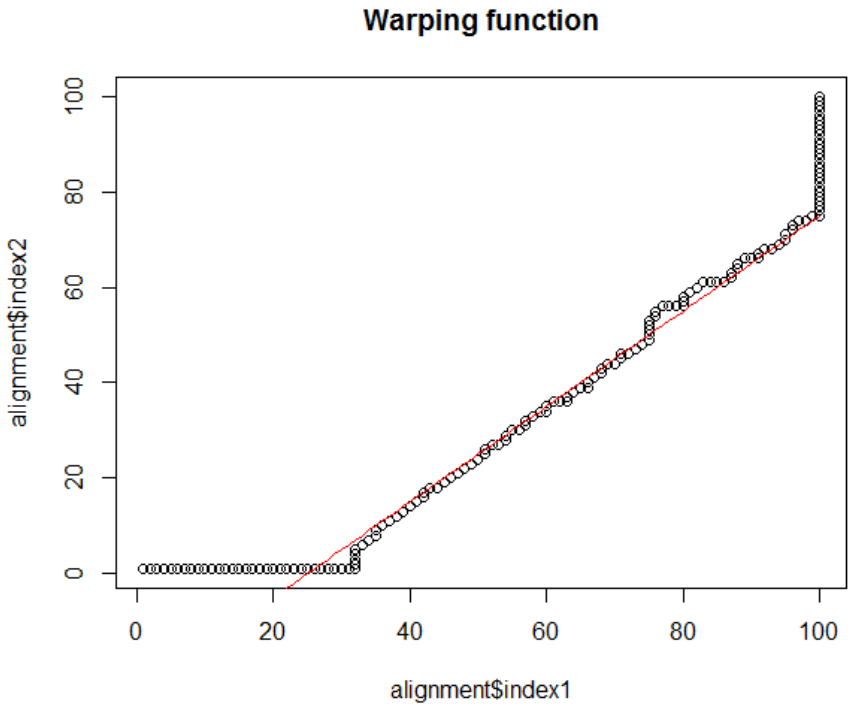

Figure 3.4: DTW-Warping function

it through American depositary receipts (ADRs). When we refer to it, theoretically, there should be no such difference. ADRs represent a certain number of shares of a foreign company. The investors can trade the ADR certificates that stand for the actual shares. ADRs are designed to be a convenient way for the United States investors to access foreign companies, and for those firms to have access to U.S. capital. The price of it should be the same as the underlying share. Thus, these descriptions are the reasons to consider ADRs as example, comparing the H- share prices.

Therefore, the thesis chooses three companies to state it validly, they have the prices of both the ADRs and H- share. The first one is the world's third largest bank by assets, named HSBC. As such, the company refers to both the United Kingdom and Hong Kong as its "home markets". HSBC has a dual primary listing on the

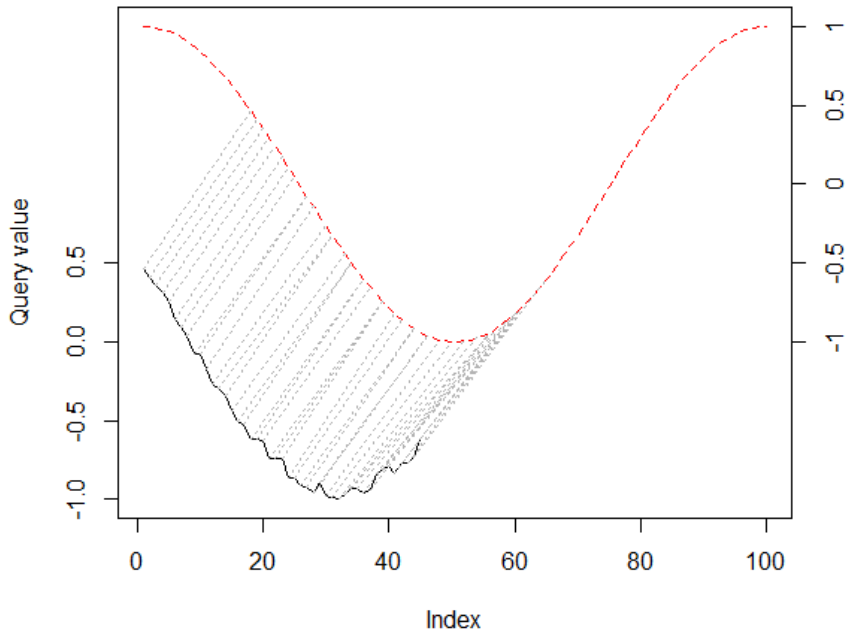

Figure 3.5: DTW-Partial alignments

Hong Kong Stock Exchange and London Stock Exchange. Besides, HSBC is a good and stable company to be considered as example, see the Figure 3.7.

However, Figure 3.7 shows us that the patterns of them are fitting well, although some points are unfitting normally. There also is often a difference between the two prices. Under some situations, essentially, an ADR may trade at a premium or discount to an underlying share. The dynamic time warping can shift the two time series to get the optimal match, it states that the prices of both ADRs and H-share are nearly similar.

Another example is about China Life Insurance Company Limited, which belongs to the financial field. This company is the first and largest state-owned insurance corporation operating in Hong Kong and Macau. The pattern fitting of it is considered as Figure 3.8.

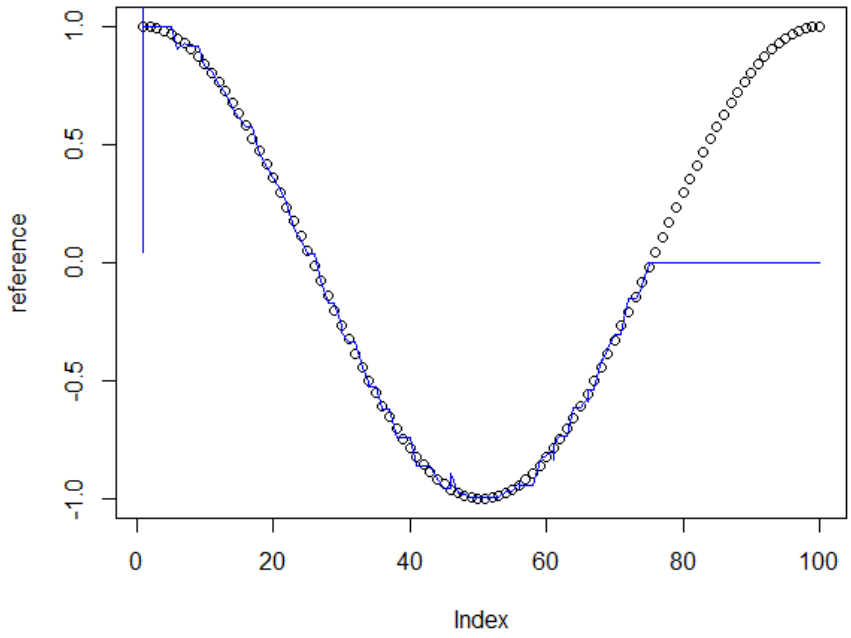

Figure 3.6: DTW-The warped sine along with cosine

Figure 3.8 shows that it has same situation with the first example, of which is nearly most of all patterns are fitting, just missing some points normally. The last example is about China Mobile Communications Corporation, it is listed on both the NYSE and the Hong Kong stock exchange. The company also is a Chinese state-owned telecommunication company that provides mobile voice and multimedia services through its nationwide mobile telecommunications network. Finally, the result of the pattern fitting by dynamic time warping is in Figure 3.9.

Because the price of ADRs is close to that of H-share, the pattern of them should be similar. This company is big scale and stable, the patterns also fit well, see Figure 3.9. These examples show the dynamic time warping is effective to apply in mining time series, and useful to understand the dissimilarity of different patterns.

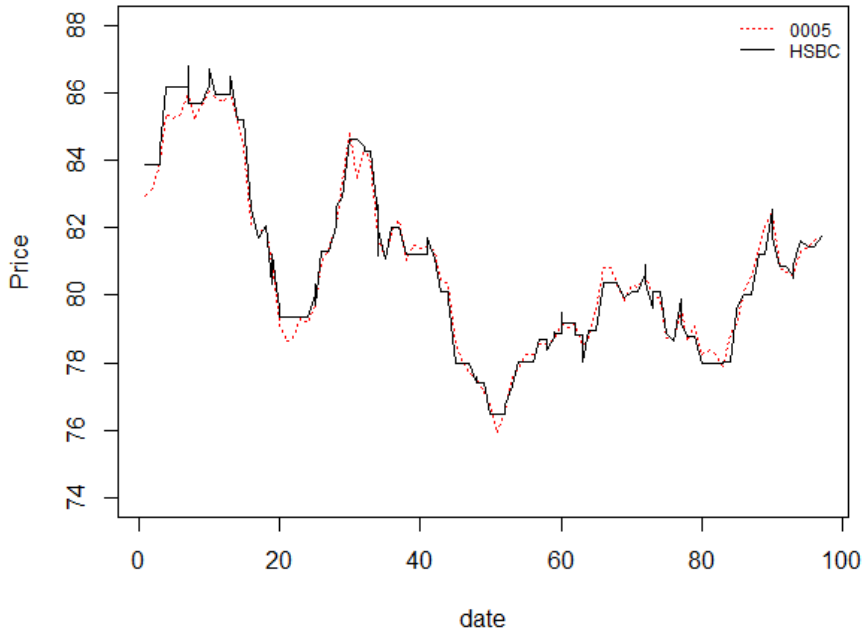

Figure 3.7: ADR-HSBC

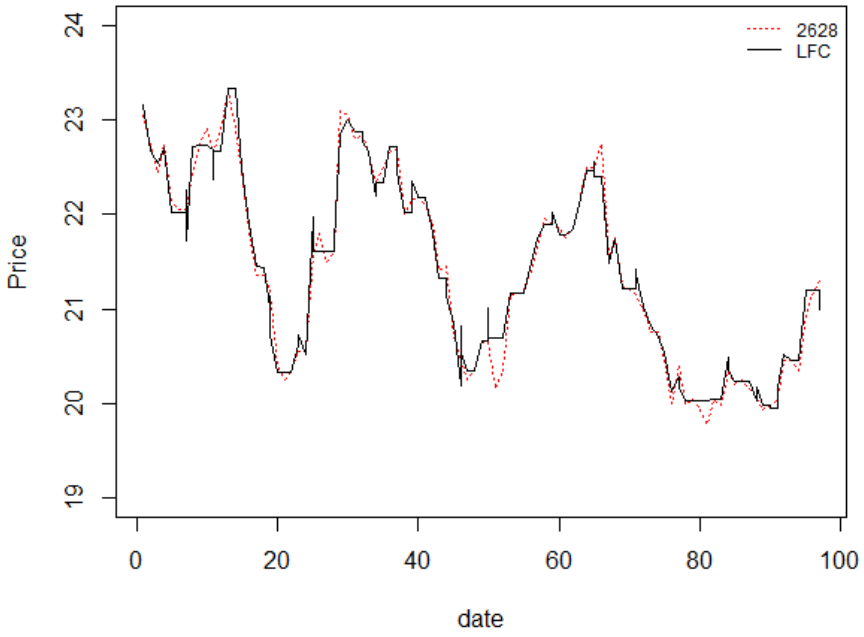

Figure 3.8: ADR-LFC

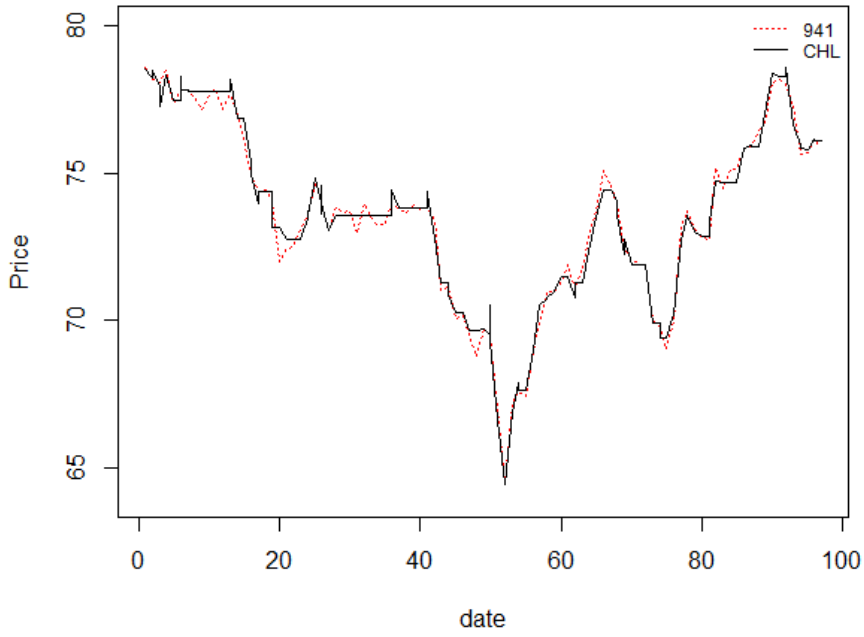

Figure 3.9: ADR-CHL

Chapter 4

Results and Analyses

In this section, we will show the results of the aforementioned methodologies used in this thesis. First, we will present the cluster results and individually analyze them from the statistical and financial perspectives. Then, we will integrate the ANOVA test results, and choose some companies based on the cluster results. We will also discuss how we use dynamic time warping to compare the patterns of A- and H- shares for the same company. Finally, we will present some details for each individual result.

4.1 Clustering analysis

Chinese A- shares are traded in Shanghai and Shenzhen exchanges, and are not available directly to foreign individual investors. On the other hand, Chinese H-shares traded in Hong Kong are open to foreign investors. Historically, the A- shares were traded at significant premium compared with H-shares. However, Hang Seng China A-H Premium Index is a convenient way of handling divergence in the prices of the A- and H- shares of constituent companies. Because the weight and number of industries in the A-H Premium Index are different, the volatility of individual industry also has a different effect on the entire index movement. Hence, analyzing the influence of major factors on different industries is necessary. Nine industries

and about 68 companies were listed in the Hang Seng China AH Premium Index during the training period. Obviously, during the sample period, finance and service industries play important roles, and account for 22.06% and 17.65% respectively in weight, while the other industries are around 8 percent of all, (Tables 2.1 and 2.2). Developing the banking industry may have a significant role in boosting A- and H-share markets.

We first divide the AH Premium Index based on the price difference, see Figure 4.20. The analysis and Figure 4.20 show that the first cluster includes 29 companies which are in the A-H premium, and most of them are the production, energy and service industries. On the other hand, other industries including 21 companies, are belonged to non-premium cluster, the finance and real estate industries dominate in it. The reasons for classifying the aforemention dual-listed companies being in non-premium cluster are considered in the following. First, the financial companies in Mainland China offer more opportunities and abilities to investors, because their performance is stable. Information on these large capitals is open, and communications between investors and the media are more frequent than others. Nevertheless, the non-premium rate is above -0.06, and high premium is around 2.39, the value of the latter is much higher than non-premium rate. Moreover, Figure 4.20 shows that the same industries have opposite performance in terms of price differences. For example, banking stocks, have not only a high premium stocks, such as Minsheng Banking (600016-1988), but also non-premium stocks, such as Ping An (601318-2318), CPIC (601601-2601), ABC (601288-1288), and ICBC (601398-1398). Based on this phenomenon, the industry can not be considered uniquely as a factor for clustering.

We use the general clustering method to divide the AH Premium Index into different groups, based on the price differentials. Rousseeuw (1987) considered it as silhouette clustering, which refers to a method of interpretation and validation of

clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster. Figure 4.1 shows the result of this method.

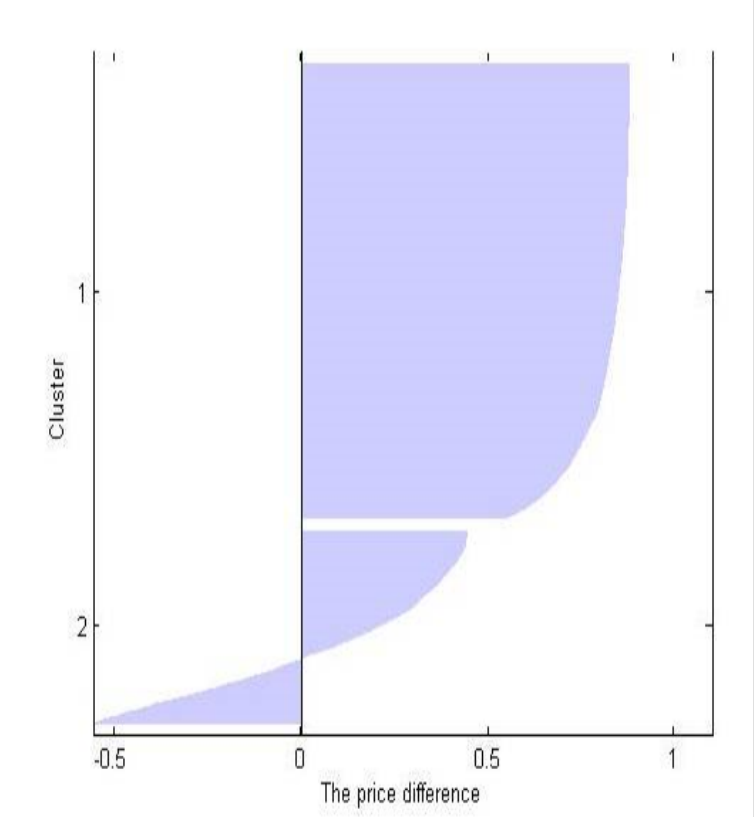

Figure 4.1: Silhouette clustering

Silhouette clustering shows that the AH Premium Index can be divided into three clusters, because of the different degrees of price gaps. However, it can not show the number of the companies that should belong to the same group. Figure 4.2 demonstrates that the capital is a constant factor for every company.

Figures 4.20 and 4.2 show that the level of non-premium cluster is lower, and the related stocks account for large-scale companies, while the small-cap shares dominate in the low- and high premium clusters.

In summary, the entire A- and H- stock markets are mainly composed by premium performance. Accordingly, the research derives an efficient result to analyze the entire AH premium stock market, we classify 50 dual-listed companies into different clusters

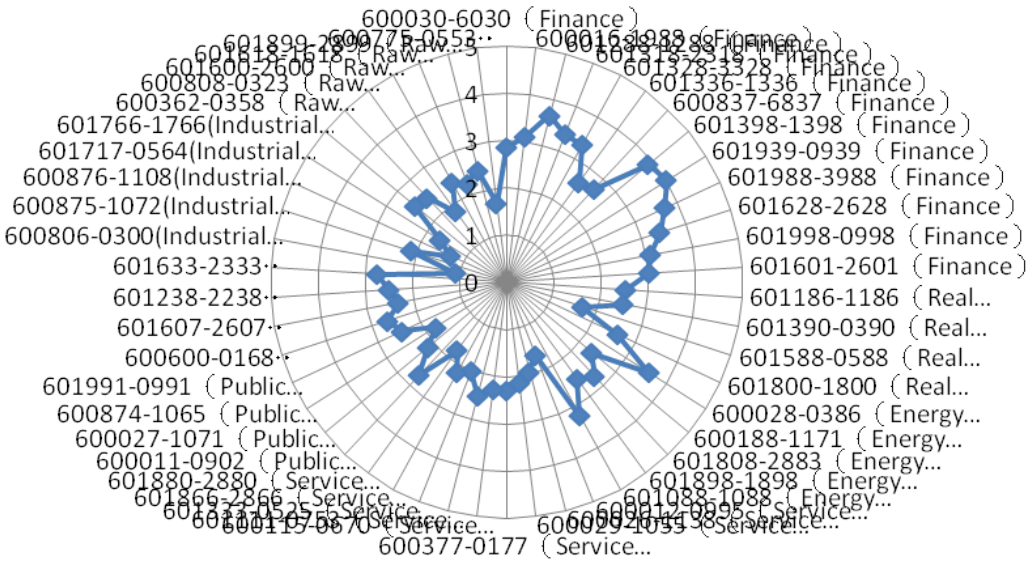

Figure 4.2: The capital of different companies

with individual characteristics using k-means clustering. We assume $k = 3$ based on the result in Figure 4.3 and traditional analysis.

Figure 4.3 suggests that the entire AH Premium Index can be grouped into the non-, low-, and high premium, and include 24, 22 and 4 companies respectively. We use the final cluster result and the corresponding variable intercept model to perform the regression analysis. The details of the statistical results will be shown in the following subsections.

4.1.1 Analysis for the cluster result

We used R program to conduct a statistical description of the explained and independent variables based on the cluster results, and obtain a deeper understanding of the research questions. Three basic statistical descriptions are listed in this subsection.

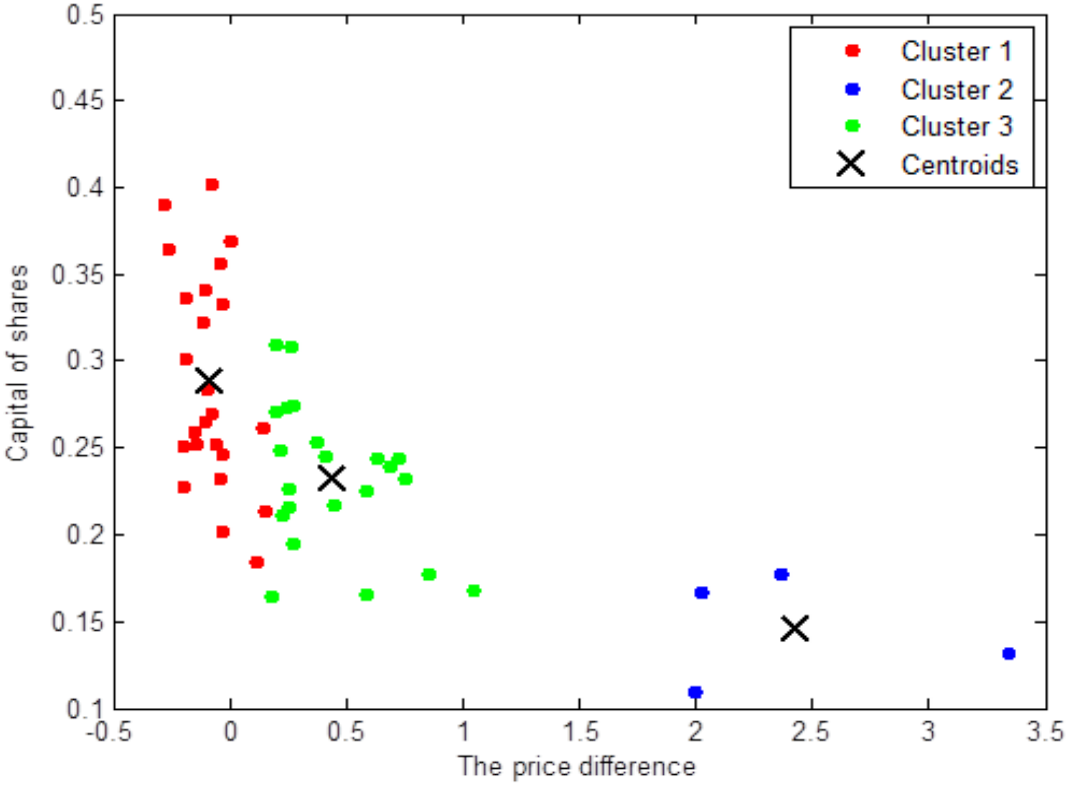

Figure 4.3: Final cluster result by k-means

Previously, we find the minimum values of these results by comparing Tables 4.1, 4.3 and Figure 4.4. The minimum value of the non-premium and high premium clusters are -0.3080 and 1.2996, respectively. The gap between these clusters is obviously wide, which indicates that clustering is valid. However, Table 4.2 shows that the minimum values of A-H low premium cluster is negative, which shows that the inversion phenomenon exists for several industries in any day. Analyzing the factors that lead to the different performance is necessary. As we analyze a portion of the mean part, we find that the mean of the non-premium rate is around -0.0611 with low volatility, which is higher than the general historical performance and thus, the price gap is closer than the previous one. The low- and high- premium clusters with the same indicator are 0.4482 and 2.3868, respectively. Their means demonstrate

Table 4.1: Statistics description about variables-non-premium

Panel 1	Mean	Median	Max	Min	Variance
Non-premium rate	-0.0611	-0.0787	0.4073	-0.3080	0.0139
CAP	2.8869	2.6772	4.0207	1.8515	0.3636
TUR	0.4088	0.2430	10.0440	0.0000	0.3085
V	2.6022	1.3128	105.5656	0.0092	18.4456
TS	1.1548	0.8817	7.3599	0.0000	1.9870
EPS	0.2665	0.1900	1.3300	-0.2559	0.0625
VOL	0.9878	0.8495	5.6880	0.1499	0.3406
ER	0.7971	0.7954	0.8122	0.7858	0.0001
MA	0.0824	0.0771	0.2627	-0.0501	0.0072
MH	0.3470	0.3541	0.5918	0.1367	0.0117

Table 4.2: Statistics description about variables-low premium

Panel 2	Mean	Median	Max	Min	Variance
Low premium rate	0.4482	0.3650	1.4417	-0.0622	0.0801
CAP	2.3273	2.3940	3.0985	1.6546	0.1655
TUR	0.6872	0.3570	28.0000	0.0220	1.0193
V	4.6791	1.6830	319.4888	0.0000	120.4232
TS	1.1619	0.6831	5.9401	0.0000	1.5153
EPS	0.2612	0.1650	2.7000	-0.4200	0.2110
VOL	1.0649	0.9288	9.4562	0.1397	0.3960
ER	0.7971	0.7954	0.8122	0.7858	0.0001
MA	0.1313	0.1334	0.3295	-0.0501	0.0050
MH	0.3059	0.2723	0.5330	0.0000	0.0124

that the levels of the price differentials are considerably different. The comparison is in Figure 4.4.

The comparison of the average total capital of industries indicates that the proxy is larger in the non-premium cluster than that in the low and high premium clusters, which are 2.8869, 2.3273, and 1.4673 respectively. The phenomenon indicates that almost all the blue chips account for the non-premium cluster.

Concurrently, the ratios of their turnover are approximately similar. However, the performance of the turnover in the premium cluster is more active than that in the non-premium cluster. The value of mean and variance are larger than that in

Table 4.3: Statistics description about variables-high premium

Panel 3	Mean	Median	Max	Min	Variance
High premium rate	2.3868	2.2226	3.9582	1.2996	0.3884
CAP	1.4673	1.6748	1.7770	1.0979	0.0745
TUR	1.2440	0.8130	11.2530	0.0730	1.8965
V	11.5859	5.0762	445.3920	0.0000	714.4547
TS	0.9139	1.1215	4.0869	0.3271	0.2615
EPS	0.0329	0.0800	0.5805	-0.2000	0.0105
VOL	1.1859	1.0399	8.2205	0.1049	0.4695
ER	0.7971	0.7954	0.8122	0.7858	0.0001
MA	0.1794	0.1091	0.3295	-0.0014	0.0091
MH	0.2718	0.3208	0.5164	0.1067	0.0270

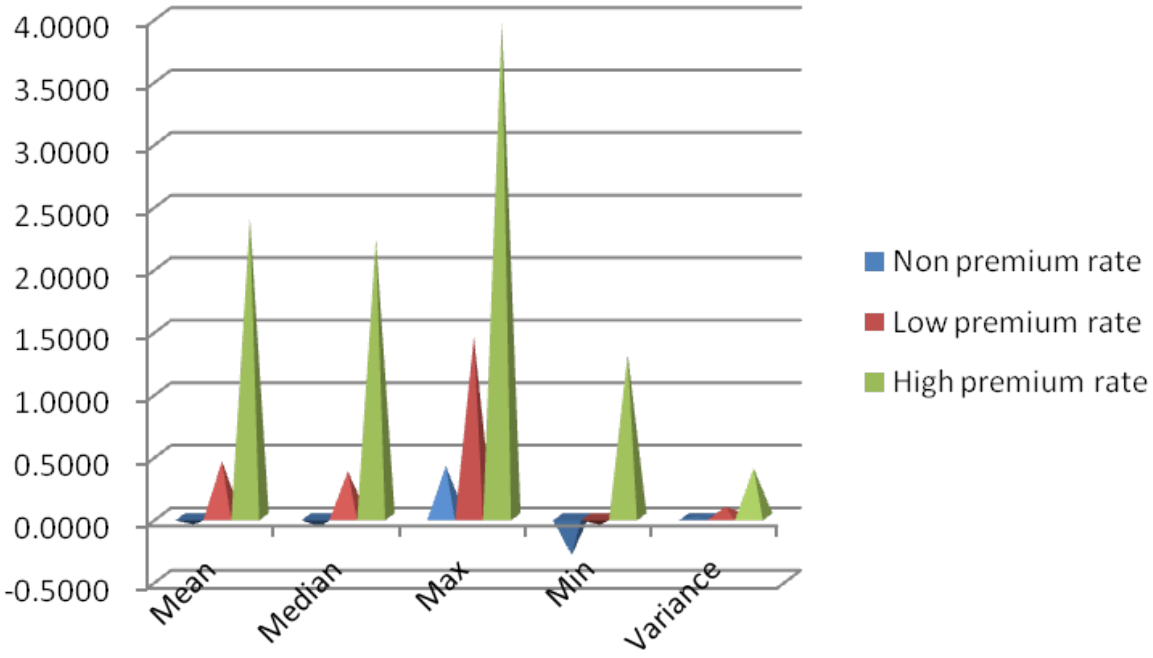

Figure 4.4: The comparing information about the three clusters-1

non-premium cluster. Volume, which is another indicator to measure liquidity, could illustrate some phenomena. The maximum volume, jumps to 105.5656, 319.4888, and 445.3920 for the non-, low and high premium clusters, respectively. These values demonstrate that the volume of A- share is larger than that of H- share for most

industries. Their maximum values are much higher than the minimum values. Thus, their variance values are 18.4456, 120.4232, and 714.4547, respectively. In summary, the liquidity of the A- share is more quick than that of the H- share when referring to the turnover and volume.

With respect to the differential elasticity of demands, the performance is obvious in the H- share market, and the integral level is more than that of the A- share market, because its proxy of the non-premium cluster is more valid and stable than that of other clusters. Hence, the fundamental factors dominate in the non-premium cluster. The performance and descriptions of the EPS are similar to the three clusters.

The next index is volatility, which is relatively low in the three conditions, with a value above or below 1. We creat the portfolio from different clusters to be a nearly fixed value, and to obtain better returns. Another factor is exchange rate, it is found to be relatively stable based on the statistical analysis. The research uses the daily exchange rate, which is the same for both A- and H- share markets. It has a mean value of 0.7971, and the variance of exchange rate is 0.0001, which is quite stable. The performance of the two markets is almost same, and stable for every industry, but the H- share market is more active than the A- share market. Most importantly, the marketable factors explain the price gaps more clearly in the premium clusters. The comparisons of these factors are in Figure 4.5.

4.1.2 Analysis for the economic model

The regression is also ran to test whether the above factors have a significant influence on the A- and H- share price gaps. The variable intercept model is shown in this subsection.

$$\begin{aligned} Y_{i,t} = b_0 + (CAP_{i,t}, TUR_{i,t}, V_{i,t}, TS_{i,t}, EPS_{i,t}, VOL_{i,t}, ER_t, MA_{i,t}, MH_{i,t}) * \\ (b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9) + \varepsilon_{i,t}, i = 1, 2, ..., 50, t = 1, 2, ..., 759. \end{aligned} \tag{4.1}$$

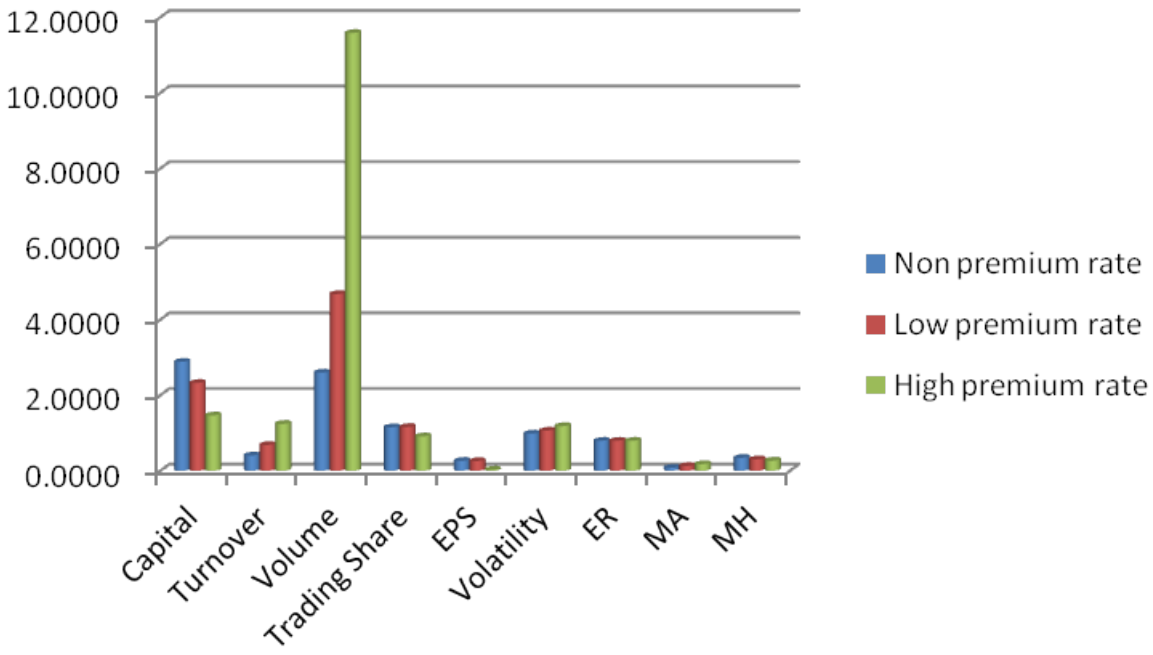

Figure 4.5: The comparing information about the three clusters-2

The study uses R software to carry out statistic analysis. We also try to capture the possible nonlinear effects and make R^2 optimal by conducting regression analysis, which determine the optimal model. In particular, the proxies of market condition including A- and H- share markets appear to exhibit a quadratic behavior in the regression curve, and therefore two terms of the market conditions are added in the proposed regression model. The final model in thesis is considered in the following manners.

$$\begin{aligned} Y_{i,t} = b_0 + b_1 CAP_{i,t} + b_2 TUR_{i,t} + b_3 V_{i,t} + b_4 TS_{i,t} + \\ b_5 EPS_{i,t} + b_6 VOL_{i,t} + b_7 ER_t + b_8 MA_{i,t} + b_9 MH_{i,t} + \\ b_{10} MA_{i,t}^2 + b_{11} MH_{i,t}^2 + \varepsilon_{i,t}, i = 1, 2, ..., 50, t = 1, 2, ..., 759. \end{aligned} \tag{4.2}$$

In order to avoid the problem of spurious regression, Dickey Fuller Test is carried out on $Y_{i,t}$, which tests whether a unit root of time series is presented in an autore-

gressive manner. We find that the DF test has a statistics equal to -3.0151, with a p-value 0.1486, which shows that the time series may be non-stationary. Therefore, the lagged $Y_{i,t-1}$ is needed to remove unit root. We test the differences $Y_{i,t} - Y_{i,t-1}$ again, and the result is given in Table 4.4. It suggests that the time series is stationary without unit root, it is at level of 99% to reject the null hypothesis. Consequently, we carry out regression using $Y_{i,t} - Y_{i,t-1}$.

Table 4.4: Dickey Fuller Test to AH Premium Index

Augmented Dickey-Fuller Test
Data: The index
Dickey-Fuller = -6.3948, P-value = 0.01
Alternative hypothesis: stationary
Result: No unit root

After running regression of the entire testing companies, the first item of A- and H- market conditions does not pass the significance test through regression test. We determine to select the final model based on the backwards regression, see Table 4.13. Finally, we obtain the final model about the entire AH Premium Index. The research also finds the adjusted goodness-of-fit value of the model is equal to 91.04%, and it is improved comparing with previous work, such as Zhanchi (2007). The regression fitting is given in the following manner:

$$\begin{aligned} Y_{i,t} = 0.0261 - 0.0556CAP_{i,t} + 0.0398TUR_{i,t} + 0.0042V_{i,t} - \\ 0.0243TS_{i,t} - 0.0924EPS_{i,t} - 0.124VOL_{i,t} - 1.23ER_t \\ -1.85MA^2_{i,t} + 0.294MH^2_{i,t} + Y_{i,t-1} + \varepsilon_{i,t}, \\ i = 1, 2, ..., 50, t = 1, 2, ..., 759. \end{aligned} \tag{4.3}$$

Although these factors are at significance 1% to explain price gaps in the entire AH Premium Index, we believe that they influence it differently. The details about different explanations of every factor are in Tables 4.5 and 4.14.

Table 4.5: Factors and significance-AH Premium Index

Factor	Coefficients value	ANOVA
CAP	-5.56E-02	16.49%
TUR	3.98E-02	12.17%
V	4.15E-03	9.59%
TS	-2.43E-02	7.06%
EPS	-9.24E-02	7.28%
VOL	-1.24E-01	5.34%
ER	-1.23E+00	9.36%
MA^2	-1.85E+00	13.31%
MH^2	2.94E-01	10.44%

After then, the research has been run the regression program respectively based on cluster results, to show that the clustering is effective. Thus, the test result of it about non-premium model is no unit root, see Table 4.6.

Table 4.6: Dickey Fuller Test to non-premium

Augmented Dickey-Fuller Test
Data: Non-premium
Dickey-Fuller = -30.5132, P-value = 0.01
Alternative hypothesis: stationary
Result: No unit root

Table 4.6 shows that the time series is stationary, because it has small p-value, of which rejects null hypothesis of non-stationary, with at the significance level of 1%. With reference to the non-premium cluster, the regression result shows that most of the companies in the cluster are at the significance level of 5%, and the adjusted goodness-of-fit value is 97.22%, which explains the explained variable well (see Table 4.14). However, some factors are not at significance level of 5%, including exchange rate and A- share market. We use the backward regression through AIC to select the better model in Table 4.15, and Akaike (1987) and Burnham and Anderson (2004) state the AIC algorithm clearly. The study also shows the residual of the model. In

the non-premium cluster, Figure 4.6 shows that the residual standard error is relatively low, and the standard normal distribution is followed. However, some factors, including information asymmetry, liquidity, elasticity of demand difference, investment philosophy difference, risk differences, H-share market condition dominate in the non-premium model. Non-premium model is considered in the following manner:

$$\begin{aligned} Y_{i,t} = 0.0215 + 0.0038CAP_{i,t} + 0.0016TUR_{i,t} + \\ 0.0002V_{i,t} - 0.0003TS_{i,t} - 0.0195EPS_{i,t} - \\ 0.0006VOL_{i,t} - 0.0126MH_{i,t}^2 + Y_{i,t-1} + \varepsilon_{i,t}, \\ i = 1, 2, ..., 24, t = 1, 2, ..., 759. \end{aligned} \tag{4.4}$$

From the fitted result, we obtain the following:

1. The proxy of the information asymmetry and liquidity can explain the significant effects of fundamental factors on the non-premium cluster.

2. The investment philosophy difference is a factor that influences on the non-premium model, and validates the original hypothesis, which can be indicated the price gap as a factor. Given that the investment philosophy difference will bring some differential invest behaviors, the investors can perform some predictions and assessments with significant risk. Thus, its proxy has a greater effect on price of H- share, and price difference.

3. The index of the H- share market condition includes the assumptions as a factor. The assumptions indicate that the relationship between H- share and the Hang Seng Index is closer, because this condition leads to decrease the diversification. Thus, investors want to obtain high returns from the increased price difference. In fact, Hang Seng Index has four main sub-indices, including business, finance, real estate, and utilities. The aforementioned industries also

belong to blue chip shares mainly. According to the previous analyses, the classification is helpful to increase the price of H-share compared with that of A- share in these industries.

4. Equally important, the coefficients of exchange rate and A- share market are not at significance level of 5% to clarify the model. In other words, the proxies do not pass the significance test. After doing backwards regression, the final model of non-premium can be more optimal without them, see Table 4.15. These proxies may be affected by other factors, thereby leading to minimal explanations on the price difference between A- and H- share markets.

According to the above model for non-premium cluster, we observe the result from the regression analysis. The details of the entire regression result can refer to Tables 4.14 and 4.15. Table 4.7 also shows the factors contribute to the non-premium cluster.

Table 4.7: Factors and significance-non-premium

Factor	Coefficients value	ANOVA
CAP	3.84E-03	26.16%
TUR	1.57E-03	13.32%
V	1.63E-04	8.14%
TS	-3.27E-04	6.58%
EPS	-1.95E-02	30.75%
VOL	-6.09E-04	6.13%
MH^2	-1.26E-02	6.14%

Table 4.7 states the fundamental factors explain the price differences better than other factor through ANOVA test. Although the results of the regression analysis explain the model well, the residuals are another indicator to state the model. Residuals and diagnostic statistics allow for the identification of patterns that are either poorly achieved by the model, have a strong influence upon the estimated parame-

ters, or a high leverage. Through residual plots, we can assess whether the residuals are consistent with the stochastic error, which can be helpful in understanding any potential problem with the model by interpreting these diagnostics jointly. The standardized residual results for the non-premium cluster also are considered in Figures 4.6, and 4.7.

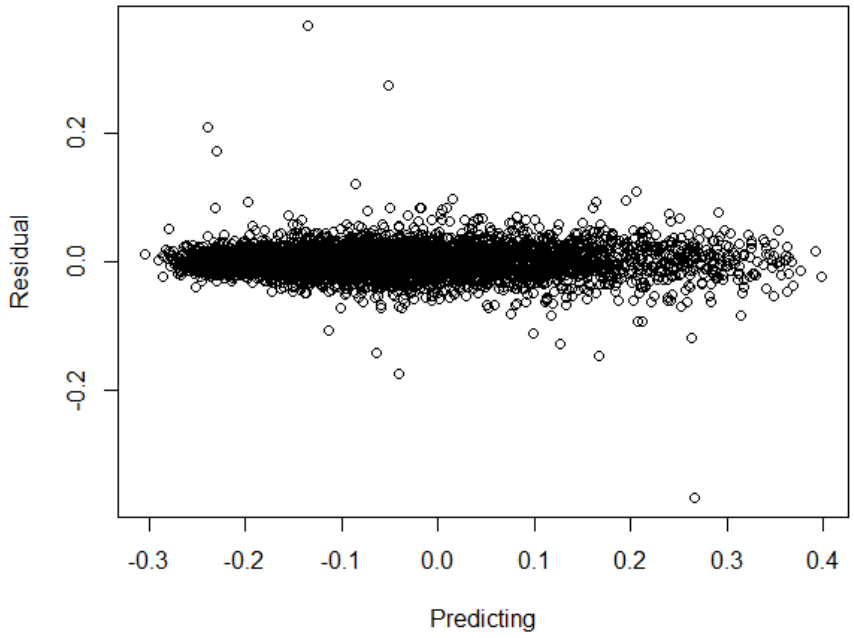

Figure 4.6: The scatter plot of residuals-non-premium

Figure 4.6 plots the residuals of the observed sample against the corresponding residuals of a standard normal distribution $\mathcal{N}(0, 1)$. The result demonstrates the validity of the model. However, this graph suggests a problem in determining the coefficients on non-premium model. Given that, this analysis is a regression, the four points at the top corner and the one at the bottom can cause concerns. These points fell far away from the least square line in the y direction as outliers to identify the problem points, and to determine whether these points stand for some trading days

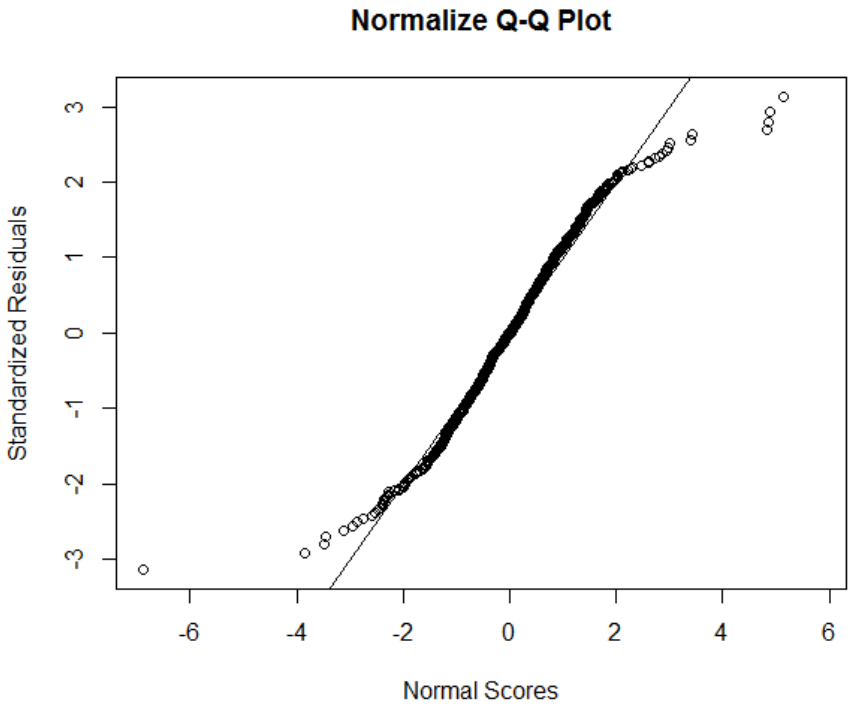

Figure 4.7: Normal probability plot of residuals-non-premium

for any company. According to the statistical sample in the study, having outliers is probably normal.

In the low premium cluster, the unit root test is firstly necessary to do. We use Dickey Fuller Test again to test it, Table 4.8 shows the null hypothesis of a unit root for price differentials is rejected, adjusted t statistic is -26.1794, which is significant at 1% testing levels. Therefore, it suggests that the time series is stationary.

Table 4.8: Dickey Fuller Test to low premium

Augmented Dickey-Fuller Test
Data: Low-premium
Dickey-Fuller = -26.1794, P-value = 0.01
Alternative hypothesis: stationary
Result: No unit root

In addition, the result about the regression of low premium clusters shows us,

information asymmetry, liquidity (turnover), elasticity of demand difference, investment philosophy difference, risk differences, exchange rate and A- share market condition are major factors in the A-H low premium. The regression fitting is given as follows:

$$\begin{aligned} Y_{i,t} = 0.310 - 0.0211CAP_{i,t} + 0.0040TUR_{i,t} - \\ 0.0698TS_{i,t} - 0.0096EPS_{i,t} - 0.0028VOL_{i,t} - \\ 0.278ER_t - 0.277MA^2_{i,t} + Y_{i,t-1} + \varepsilon_{i,t}, \\ i = 1, 2, ..., 22, t = 1, 2, ..., 759. \end{aligned} \tag{4.5}$$

According to the fitted result, and combining Tables 4.14, 4.16 and Figure 4.8, we can observe the following:

1. Statistical regression indicates that nearly all factors are at significance level of 1%, except for volume and H- share market condition. We select the final model by backwards regression. The residual standard error is also relatively low, standard normal distribution is obeyed, and the adjusted goodness-of-fit value is 93.42%. In other words, these indices clarify the A-H low premium model well. The major factors that can influence the A-H premium, including the information asymmetry, liquidity (turnover), elasticity of demand, investment philosophy, exchange rate, and the condition of A- share market.

2. The index of information asymmetry is negative to the low premium rate, which is in accordance with the previous hypothesis as a factor. Various industries gather in A-H low premium cluster, because of the differential levels of information asymmetry. However, the cluster contains main companies from the service industry. In particular, as long as the capital controls are maintained in the mainland market, the prices of A- shares of several companies substan-

tially can exceed the prices of their H- shares, and the price gaps between them can increase.

3. The proxy of liquidity (turnover) displays a positive effect on the A-H low premium rate, which is in accordance with the previous hypothesis as a factor to influence the price gap.

4. The index of elasticity of demand is also at significance level of 1%, and it shows that the elasticity of demand is negative at A-H low premium rate. The result is the same as that with the original hypothesis as a factor.

5. The investment philosophy statement indicates that if the company has high EPS, the rational investors are willing to pay for the high stock price, thereby leading to decrease the price differences. It consists of the original assumptions, and clarifies the A-H low premium rate well as a factor.

6. The exchange rate and A- share market environment play a role in the A-H low premium, and they have a negative relationship that become an explanatory variable. The A- share market is also highly influenced by the mainland market. The higher the correlation is, the closer the value of investments can become, thereby resulting in diversification can reduce the risk.

7. The other factors are of less importance in explaining the A-H low premium. Moreover, only two factors, namely volume, and H- share market condition do not pass significance test. And after doing backwards regression, we select the final model without them in the low premium cluster, see Table 4.16.

According to aforementioned model for the low premium cluster, we can conduct the analysis for the statistical regression. The table 4.9 explains the factor performance in the low premium group.

Table 4.9: Factors and significance-low-premium

Factor	Coefficients value	ANOVA
CAP	-2.11E-02	27.82%
TUR	4.03E-03	11.96%
TS	-6.98E-02	13.57%
EPS	-9.63E-03	10.22%
VOL	-2.83E-03	4.56%
ER	-2.78E-01	7.87%
MA^2	-2.77E-01	17.42%

Table 4.9 explains that the fundamental and marketable factors influence the price gaps similarly in the low premium cluster. The other details about regression results are shown in Table 4.14.

In particular, the residual result suggests that regardless of whether the model is systematically incorrect, we have an opportunity to improve the model though the result. The residual result for the low premium part is considered as follows, (see Figure 4.8).

First, the residuals come from a normal distribution, as a result of the scatter plot of the standardized residuals against the fitted values. The residuals vary randomly at zero, and its spread is about the same throughout the plot. However, one point is at the right bottom corner, and the other three points are at the middle as outliers. The points usually have an unusually large residual. Nevertheless, this residual is normal because of the large sample size used in the research. The preceding analysis shows that the model of the low premium is valid.

At the same time, A-H high premium has the approximate description, whereas, the main factors are different. Before doing regression analysis, unit root should be tested firstly through Dickey Fuller Test. Obviously, we obtain a test statistic that is significant at 1% level, see Table 4.10.

Thus, the regression fitting is given as follows:

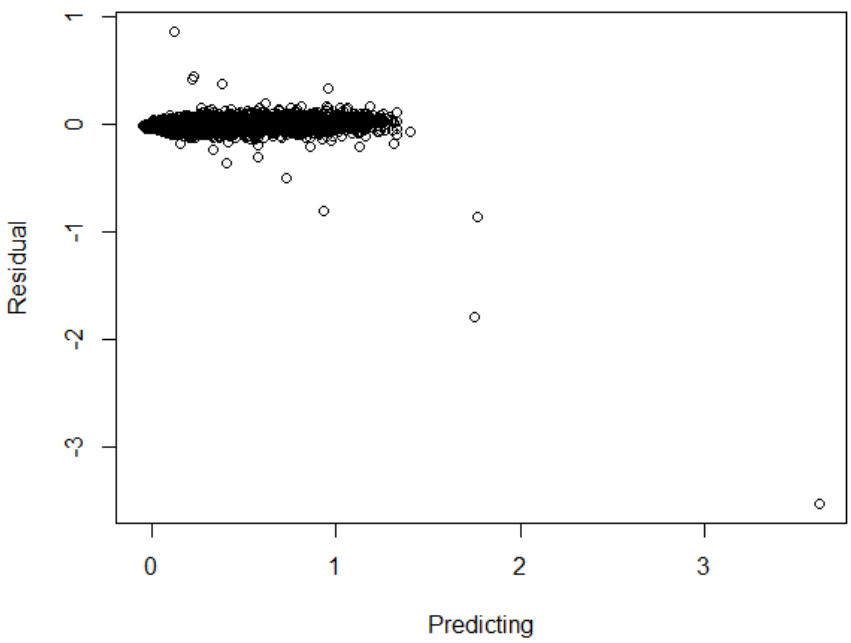

Figure 4.8: The scatter plot of residuals-low premium

Table 4.10: Dickey Fuller Test to high premium

Augmented Dickey-Fuller Test
Data: High-premium
Dickey-Fuller = -14.9009, P-value = 0.01
Alternative hypothesis: stationary
Result: No unit root

$$Y_{i,t} = -0.663 - 0.0321CAP_{i,t} + 0.0084TUR_{i,t} + 0.0280TS_{i,t}$$
$$+0.8860ER_t - 0.536MA^2_{i,t} + 0.0129MH^2_{i,t} + Y_{i,t-1} + \varepsilon_{i,t}, \quad (4.6)$$
$$i = 1, 2, ..., 4, t = 1, 2, ..., 759.$$

Tables 4.11, 4.14 and 4.17 show the results of the regression analysis, based on the preceding model of the high premium cluster, through AIC testing. The adjusted goodness-of-fit value of high premium model is 96.31%. Specially, the model shows

Table 4.11: Factors and significance-high premium

Factor	Coefficients value	ANOVA
CAP	-3.21E-02	19.84%
TUR	8.38E-03	20.55%
TS	2.80E-02	4.96%
ER	8.86E-01	5.97%
MA^2	-5.36E-01	21.58%
MH^2	1.29E-02	23.41%

that the market conditions have high correlation with the high premium cluster. The values of the factors contribution to the high premium model are 21.58% and 23.41% by ANOVA test, respectively. Information asymmetry also influences the high premium cluster in the same manner as in the other two clusters.

By combining regression analysis, the residuals can be centered on zero throughout the range of fitted values. In other words, the model is correct on average for all fitted values. The residual result for high premium cluster is considered at the bottom, see Figure 4.9.

Figure 4.9 shows that the residuals are not correlated. Therefore, the results of residual plot also show that the model is valid. However, the three points are located at left top corner as outliers. The outliers are normal to have in the residual plot, because of the large sample size in studying. Outliers occur very frequently in real data, and are often not noticed because most data nowadays are processed by computers, without careful inspection or screening. Outliers may be a result of misplaced decimal points, recording or transmission errors.

After the regression and residual analysis, we find another phenomenon in the price gaps between the A- and H- stock markets. In particular, the level of non-premium cluster is lower than that of the premium cluster. This thesis uses analysis of variance (ANOVA) to effectively measure the phenomenon, see Gelman (2005).

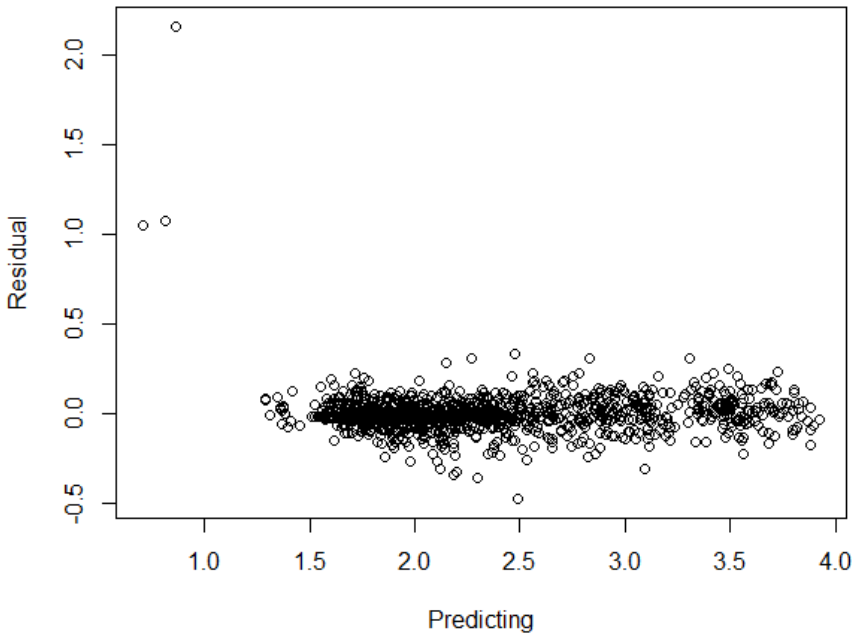

Figure 4.9: The scatter plot of residuals-high premium

Table 4.12: ANOVA result

Anova: Single Factor				
SUMMARY				
Groups	Count	Sum	Average	Variance
Non premium rate	18216.0000	-1113.2619	-0.0611	0.0139
Low premium rate	16698.0000	7484.4452	0.4482	0.0801
High premium rate	3036.0000	7246.4684	2.3868	0.3884
P-value	0.0000	F crit	3.8417	

The results of k-means clustering and ANOVA test are combined, see Figure 4.3, and Table 4.12. As demonstrated, dividing the entire AH Premium Index to three clusters is efficient. The differences between the group means and their associated procedures are analyzed based on the ANOVA test. The p value is equal to 0, and the F ratio is also large. Hence, the ANOVA test shows that the three clusters are different significantly. Other features can be sorted, we analyze it from a technical

angle, and the technical analysis and pattern similarity can be considered in the next subsection.

4.1.3 Analysis for an example-Vanke company

Figures 4.2 and 4.3 show the company explains our predictions further. Vanke is the largest residential real estate developer in the Mainland China, and has expanded to Hong Kong, United States, and Singapore since 2012. The reason for choosing this company as sample is that Vanke is a dual-listed company in Mainland (1991) and Hong Kong (2014). It belongs to the research scale as a new company listed in A- and H- share markets.

The company had the largest market capitalization in 2006 on the Shenzhen Stock Exchange. As of 28 May 2013, its capital market was HK165 billion. Moreover, this company can be guessed to exist in the non-premium cluster based on the research results. The thesis chooses the date period from July 2014 to May 2015.

The study uses the non-premium model to obtain the price difference between the A- and H- shares of the Vanke company. The results are shown in Figure 4.10.

Figure 4.10 shows that Vanke belongs to the non-premium cluster exactly, the price gap of it is always negative. The simulation result further determines a small error between the patterns of model- and truth- data. In addition, the trend of pattern is nearly similar, which explains the model effectively. The model can be used to predict the trend of pattern, and determine the company belonging to which cluster. We will do much more tests for the new members that will be listed in A- and H- share markets further. The subsection just lists an example (Vanke company), because it is stable in research period, without considering some operation or policy factors.

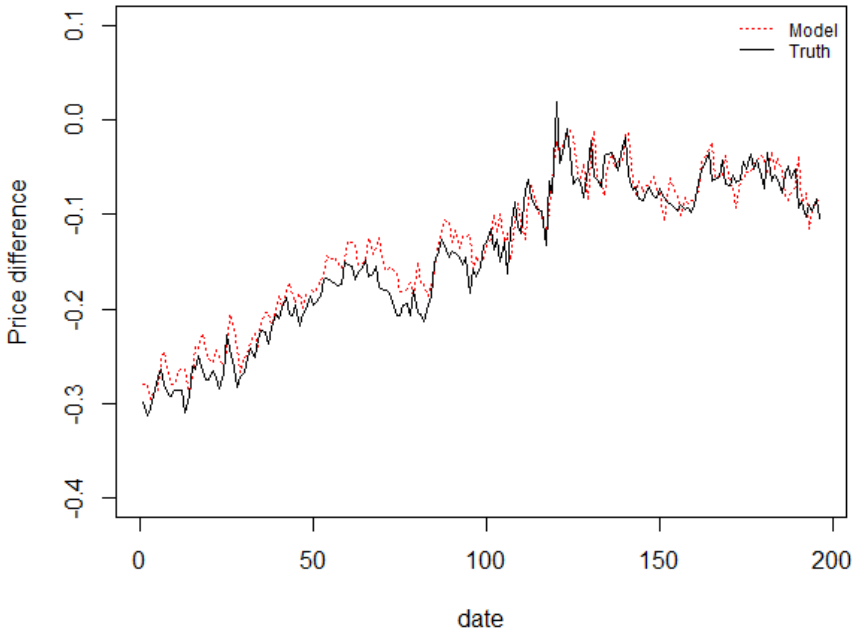

Figure 4.10: Vanke company

Table 4.13: Regression result for separate samples-AH Premium Index

	AH Premium Index	AH Premium Index	AH Premium Index	AH Premium Index
Intercept	2.61E-02	2.52E-02	2.19E-02	2.22E-02
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
CAP	-5.56E-02	-5.43E-02	-5.57E-02	-5.59E-02
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
TUR	3.98E-02	3.87E-02	3.91E-02	4.01E-02
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
V	4.15E-03	4.17E-03	4.18E-03	4.11E-03
	(0.007)**	(0.006)**	(0.006)**	(0.009)**
TS	-2.43E-02	-2.38E-02	-2.40E-02	-2.60E-02
	(0.029)*	(0.033)*	(0.032)*	(0.021)*
EPS	-9.24E-02	-9.26E-02	-9.19E-02	-9.24E-02
	(0.033)*	(0.031)*	(0.038)*	(0.033)*
VOL	-1.24E-01	-1.16E-01	-1.21E-01	-1.11E-01
	(0.023)*	(0.033)*	(0.025)*	(0.036)*
ER	-1.23E+00	-1.35E+00	-1.17E+00	-1.28E+00
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
MA		-9.06E-01		-9.01E-01
		(0.095).		(0.101)
MH			5.41E-02	4.81E-02
			(0.293)	(0.344)
MA^2	-1.85E+00	-1.45E+00	-1.71E+00	-1.78E+00
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
MH^2	2.94E-01	2.01E-01	2.11E-01	2.06E+00
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Adjusted R-squared	0.9104	0.9094	0.9091	0.9087
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1				

Table 4.14: Regression result for separate samples

	AH Premium Index	Non-premium	Low premium	High premium
Intercept	2.61E-02	2.25E-02	3.09E-01	-6.48E-01
	(0.000)***	(0.005)**	(0.001)**	(0.007)**
CAP	-5.56E-02	3.83E-03	-2.07E-02	-3.16E-02
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
TUR	3.98E-02	1.57E-03	3.91E-03	8.21E-03
	(0.000)***	(0.000)***	(0.012)*	(0.000)***
V	4.15E-03	1.65E-04	6.65E-05	3.95E-05
	(0.007)**	(0.026)*	(0.353)	(0.559)
TS	-2.43E-02	-3.38E-04	-7.00E-02	2.55E-02
	(0.029)*	(0.052).	(0.008)**	(0.136)
EPS	-9.24E-02	-1.95E-02	-8.91E-03	8.03E-02
	(0.033)*	(0.000)***	(0.010)*	(0.587)
VOL	-1.24E-01	-6.16E-04	-2.71E-03	7.28E-03
	(0.023)*	(0.090).	(0.251)	(0.206)
ER	-1.23E+00	2.95E-02	-2.77E-01	8.23E-01
	(0.000)***	(0.377)	(0.018)*	(0.610)
MA^2	-1.85E+00	3.45E-03	-2.71E-01	-3.12E-01
	(0.000)***	(0.516)	(0.000)***	(0.000)***
MH^2	2.94E-01	-1.01E-02	1.01E-02	1.66E-02
	(0.000)***	(0.066).	(0.394)	(0.000)***
Adjusted R-squared	0.9104	0.9720	0.9340	0.9628
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1				

Table 4.15: Regression result for separate samples for non-premium

	Non-premium/1	Non-premium/2	Non-premium/3
Intercept	9.92E-04	9.93E-04	2.15E-02
	(0.037)*	(0.036)*	(0.007)**
CAP	3.88E-03	3.89E-03	3.84E-03
	(0.000)***	(0.000)***	(0.000)***
TUR	1.64E-03	1.64E-03	1.57E-03
	(0.000)***	(0.000)***	(0.000)***
V	1.59E-04	1.57E-04	1.63E-04
	(0.032)*	(0.034)*	(0.029)*
TS	-3.31E-04	-3.26E-04	-3.27E-04
	(0.074).	(0.080).	(0.079).
EPS	-1.92E-02	-1.92E-02	-1.95E-02
	(0.000)***	(0.000)***	(0.000)***
VOL	-5.43E-04	-5.56E-04	-6.09E-04
	(0.126)	(0.109)	(0.092).
ER		2.69E-02	
		(0.484)	
MA^2	3.28E-03		
	(0.654)		
MH^2	-1.02E-02	-1.22E-02	-1.26E-02
	(0.066).	(0.032)*	(0.029)*
Adjusted R-squared	0.9684	0.9700	0.9722
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1			

Table 4.16: Regression result for separate samples for low premium

	Low premium/1	Low premium/2	Low premium/3
Intercept	3.21E-01	3.08E-01	3.10E-01
	(0.000)***	(0.001)**	(0.001)**
CAP	-2.11E-02	-2.06E-02	-2.11E-02
	(0.000)***	(0.000)***	(0.000)***
TUR	3.69E-03	3.93E-03	4.03E-03
	(0.031)*	(0.012)*	(0.008)**
V		7.36E-05	
		(0.066).	
TS	-6.96E-02	-6.98E-02	-6.98E-02
	(0.009)**	(0.009)**	(0.009)**
EPS	-8.57E-03	-9.60E-03	-9.63E-03
	(0.011)*	(0.000)***	(0.000)***
VOL	-2.37E-03	-2.92E-03	-2.83E-03
	(0.101)	(0.041)*	(0.047)*
ER	-2.78E-01	-2.77E-01	-2.78E-01
	(0.018)*	(0.018)*	(0.018)*
MA^2	-2.72E-01	-2.75E-01	-2.77E-01
	(0.000)***	(0.000)***	(0.000)***
MH^2	1.10E-02		
	(0.351)		
Adjusted R-squared	0.9331	0.9337	0.9342
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1			

Table 4.17: Regression result for separate samples for high premium

	High premium/1	High premium/2	High premium/3
Intercept	-6.50E-01	-6.79E-01	-6.63E-01
	(0.006)**	(0.003)**	(0.004)**
CAP	-3.17E-02	-3.23E-02	-3.21E-02
	(0.000)***	(0.000)***	(0.000)***
TUR	8.19E-03	8.06E-03	8.38E-03
	(0.000)***	(0.000)***	(0.000)***
V			
TS	2.34E-02	2.36E-02	2.80E-02
	(0.015)*	(0.013)*	(0.009)**
EPS	8.07E-02		
	(0.459)		
VOL	7.28E-03	7.29E-03	
	(0.206)	(0.205)	
ER	8.25E-01	8.34E-01	8.86E-01
	(0.140)	(0.117)	(0.092).
MA^2	-3.21E-01	-5.27E-01	-5.36E-01
	(0.000)***	(0.000)***	(0.000)***
MH^2	1.19E-02	1.11E-02	1.29E-02
	(0.000)***	(0.000)***	(0.000)***
Adjusted R-squared	0.9589	0.9592	0.9631
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1			

4.2 Dynamic time-warping results

In this section, we formulate the pattern between the A- and H- share markets for the same company through dynamic time-warping algorithm. This method is useful to measure the shape-similarity between the two markets. We choose the nine companies from the three clusters individually to conduct empirical analyses, and use the backwards regression test them one by one, see Tables 4.27, 4.28 and 4.29. The details for these studies are analyzed in the following subsections.

4.2.1 Performance of the non-premium cluster

In the non-premium cluster, we choose three proper stocks to show the performance of the non-premium cluster. The first company, called the China Pacific Insurance Group, is listed in A- share market, named 601601, and called 2601 in H- share marked. The company is chosen because of its outstanding performance and stability during this period. The company belongs to the financial industry. Most of the blue chips account for the non-premium cluster, particularly the financial industry. The company has all the necessary characteristics in the non-premium cluster, Hence, it can be considered as an example to discuss.

Figure 4.11 shows that the patterns between A- and H- stocks are nearly similar, except for the period from the end in 2013 to the middle in 2014, (see Chung et al. (2004)), it can be called pattern dissimilarity part. In other words, nearly all H- share prices outperformed A- share prices. We try to test the reason for this phenomenon by conducting to compare the different factors, dividing the sample period into pattern dissimilarity and similarity, see Chan (1993). The results are shown in Table 4.18.

Table 4.18 indicates that several fundamental factors influence the pattern differences of this company, including information asymmetry, liquidity, investment philosophy difference, etc. In particular, several fundamental factors are usually as-

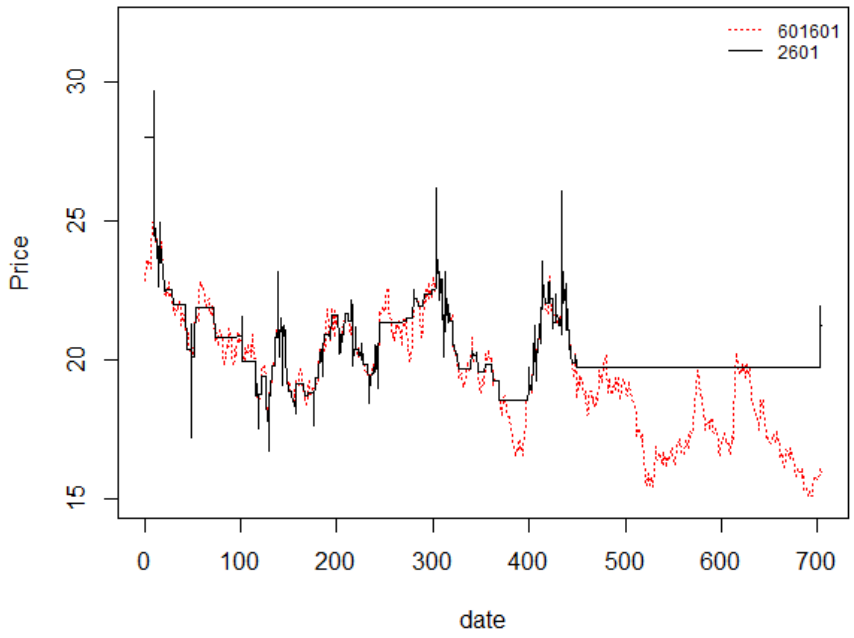

Figure 4.11: Patterns comparsion-601601-2601

sociated with non-premium rather than market condition or other technical factors, through comparing the dissimilarity, similarity parts and the entire non-premium cluster. This performance is stable. In addition, these fundamental factors about this company fluctuate a little, which can influence the volatility of price differences more widely than others.

Table 4.18: Factors comparsion-601601-2601

	Dissimilarity	Similarity	Non-premium
CAP	√	√	√
TUR	√	√	√
V	√	√	√
TS			√
EPS	√	√	√
VOL		√	√
ER			
MA^2			
MH^2		√	√

The section has another example to explain the phenomenon in the non-premium cluster. The second company also belongs to the financial industry. The company is called the China Construction Bank (CCB) Corporation, which is one of the big four banks in the People′s Republic of China. In 2011, CCB was the second largest bank in the world based on market capitalization and 13th largest company in the world, see DeCarlo (2011). Thus, it is listed in A- share market named 601939 and in H- share market named 939. The pattern of CCB indicates another finding, see Figure 4.12.

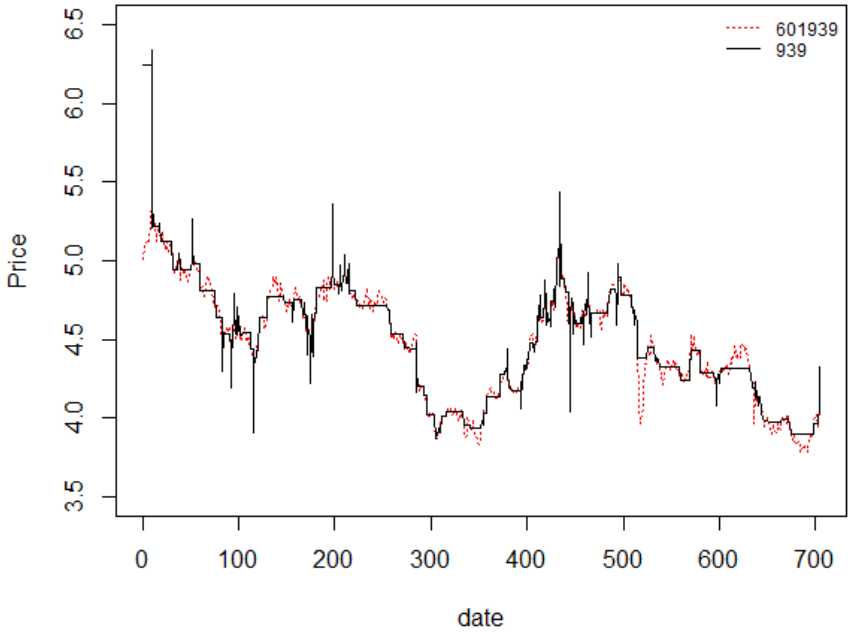

Figure 4.12: Patterns comparsion-601939-939

The pattern of CCB indicates us that the trending of A- and H- shares is similar, and the price of H- share is slightly higher than that of the A- share. The dynamic time-warping method can match the patterns between them, the patterns of them are nearly similar. Nevertheless, the reason for the pattern dissimilarity can be found

in Table 4.19.

Table 4.19: Factors comparsion-601939-939

	Dissimilarity	Similarity	Non-premium
CAP	√	√	√
TUR	√	√	√
V			√
TS	√	√	√
EPS	√	√	√
VOL			√
ER			
MA^2		√	
MH^2	√	√	√

This company is approximately same as that of the China Pacific Insurance Group, which can also stand for the financial industry. We divide the period into two parts again, named pattern dissimilarity and similarity parts. The dissimilarity of two patterns is caused by the significantly fundamental factors again, including the information asymmetry, liquidity, investment philosophy difference. In addition, the study finds other industries show that the non-premium cluster has much higher relationship with some fundamental factors. The company called the China Railway Construction Corporation, belongs to the real estate and building industry. The company is the second largest state-owned construction enterprise in China, just after China Railway Engineering Corporation and is also the top construction contractor in the world by total revenue. We can compare the patterns of them by dynamic time-warping algorithm, see in Figure 4.13.

Table 4.20 shows that the company has nearly the correlation with the two aforementioned companies, which shows the later period belongs to the pattern dissimilar part. Therefore, the pattern dissimilarity in the non-premium cluster focuses on several fundamental factors.

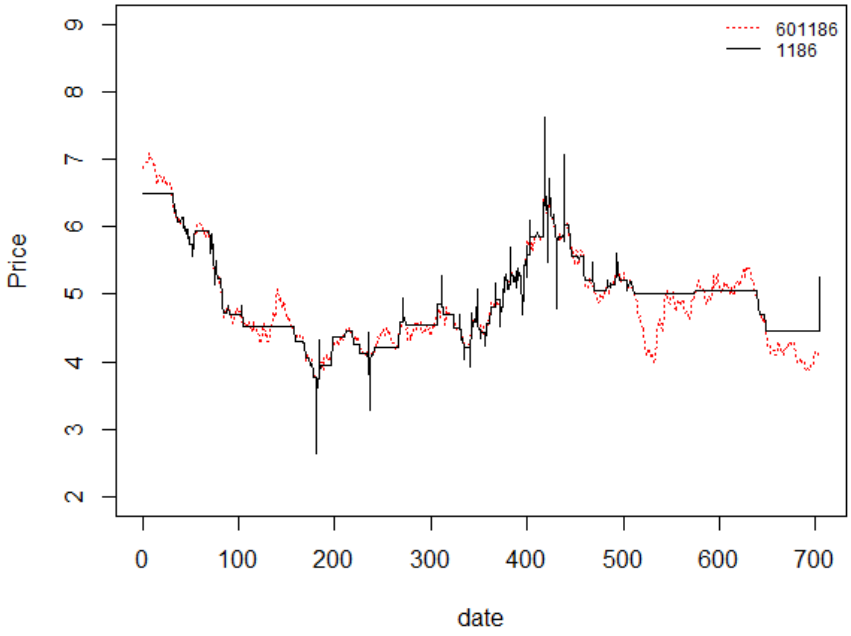

Figure 4.13: Patterns comparsion-601186-1186

4.2.2 Performance of the low premium cluster

In the low premium group, China National Coal Group Corporation is traded as the first sample. The company is listed in A- and H- share markets, and its stock codes are 601898 and 1898, respectively. In particular, compared with the non-premium cluster, small cap stocks account for this cluster, and they often include companies

Table 4.20: Factors comparsion-601186-1186

	Dissimilarity	Similarity	Non-premium
CAP	√	√	√
TUR	√	√	√
V	√	√	√
TS			√
EPS	√	√	√
VOL			√
ER		√	
MA^2			
MH^2	√		√

in the energy and service industries. China Coal has all the characteristics of the low premium cluster, as it belongs to the energy industry and represents the low premium cluster well, see Figure 4.14.

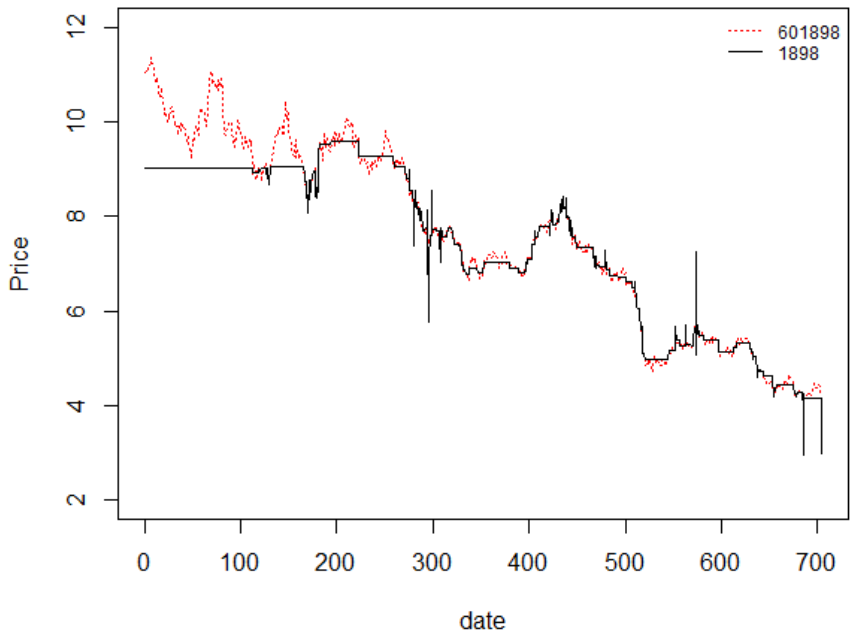

Figure 4.14: Patterns comparsion-601898-1898

Figure 4.14 shows that the dynamic time warping is used to match the patterns of two shares for the same company. The fitting percent of the pattern similarity is high, but the price of H- stock is obviously lower than that of the A stock at the beginning of training period. The phenomenon in the low premium cluster is explained again, which indicates that the levels of the price gaps are not same. The reasons for the pattern dissimilarity in 2011 are caused by several market factors, including exchange rate, CSI300 index, and others. The details about the factors contribute to the unfitting, fitting part and the entire low premium cluster are presented in Table 4.21.

Table 4.21: Factors comparsion-601898-1898

	Dissimilarity	Similarity	Low premium
CAP	√	√	√
TUR	√	√	√
V			
TS			√
EPS	√	√	√
VOL			√
ER	√	√	√
MA^2	√	√	√
MH^2	√		

The study chooses the service industry to explain the pattern performance in the low premium sort. The stock codes of the company are 600012 and 995 for A- and H- stock markets respectively. Additionally, Anhui Expressway Company Limited listed in H- share market in 13 November 1996, which was much earlier than that in A share market (7 January 2003). The pattern dissimilarity of this company is shown in Figure 4.15.

Figure 4.15 shows that nearly most of the patterns fit well, and that the dissimilarity is indicated by some several factors, including the information asymmetry, liquidity, investment philosophy difference, exchange rate, and A- share market. Table 4.22 shows the details on the comparison of pattern dissimilarity, dissimilarity and the entire low premium cluster.

Table 4.22: Factors comparsion-600012-995

	Dissimilarity	Similarity	Low premium
CAP	√	√	√
TUR	√	√	√
V	√		
TS	√	√	√
EPS	√	√	√
VOL			√
ER	√	√	√
MA^2	√	√	√
MH^2			

Lastly, the energy and service industries are not the only fields that account for

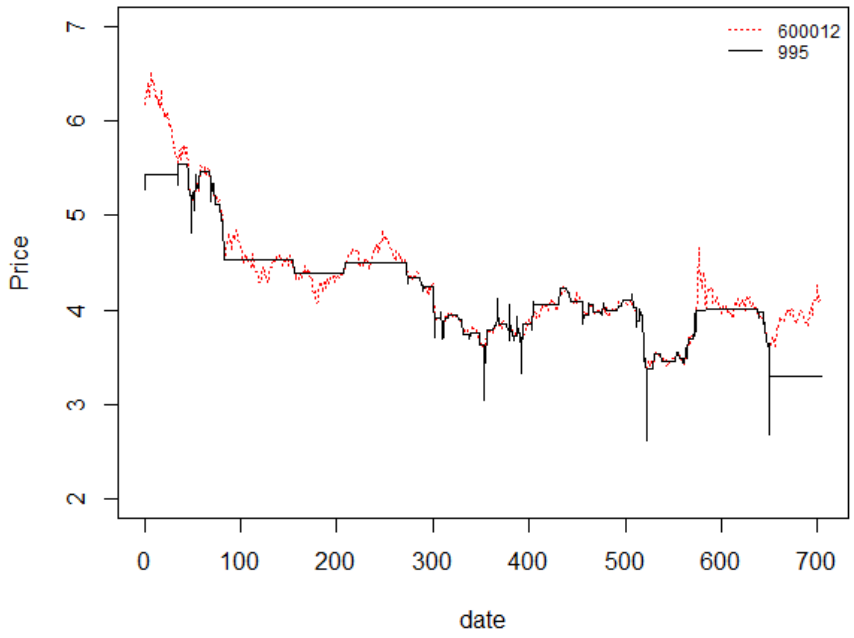

Figure 4.15: Patterns comparsion-600012-995

the low premium cluster. A company that belongs to the financial field also has the same performance with this group. Thus, we use this company, which is assigned to the low premium cluster through k-means clustering. This company called the China International Trust and Investment Corporation (CITIC) Bank, it is a nationally comprehensive and internationally oriented commercial bank. Almost 130 countries cover it currently, while still maintaining a strong foothold on the mainland banking industry. China CITIC Bank is a wholly owned subsidiary of CITIC, with assets of USD 474.73 billion. As of 2006, CITIC Bank has had a non-performing loan ratio of 2.5% (RMB 11.1 billion). The capital adequacy ratio of this bank is 9.1%.

Figure 4.16 illustrates that the pattern of A- share is more volatile than that of H- share. Moreover, the price of A stock is obviously higher than that of H- stock for the same company. However, its pattern dissimilarity part is related with several

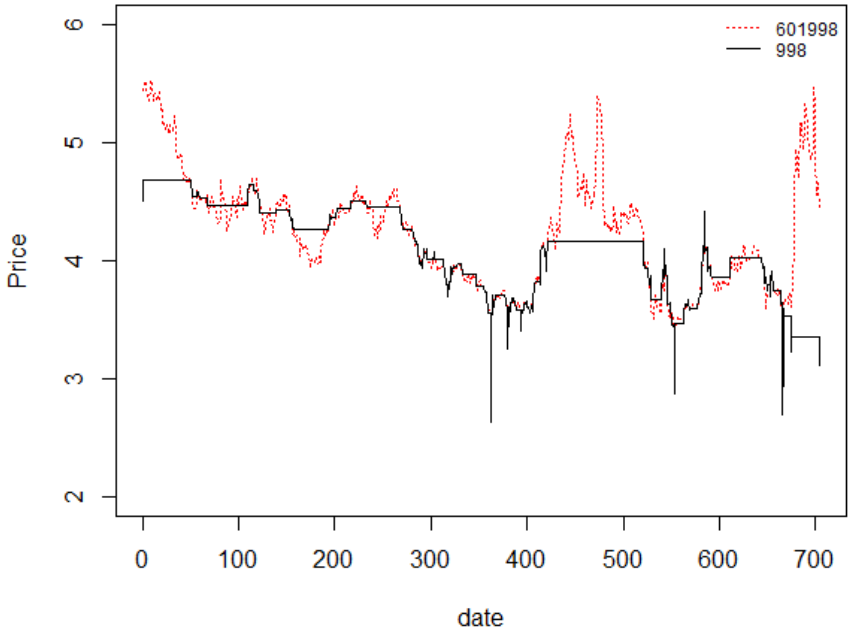

Figure 4.16: Patterns comparsion-601998-998

marketable factors, such as exchange rate, and market conditions, see Table 4.23.

4.2.3 Performance of the high premium cluster

The stock code named 600775 and 0553, is listed in A- and H- stock markets. The company is chosen as the first sample in the high premium cluster. In addition, this

Table 4.23: Factors comparsion-601998-998

	Dissimilarity	Similarity	Low premium
CAP	√	√	√
TUR	√	√	√
V			
TS		√	√
EPS			√
VOL			√
ER	√	√	√
MA^2	√	√	√
MH^2	√		

cluster has merely four stocks and they have same shape. Thus, we randomly choose Nanjing Panda Electronics as the first one.

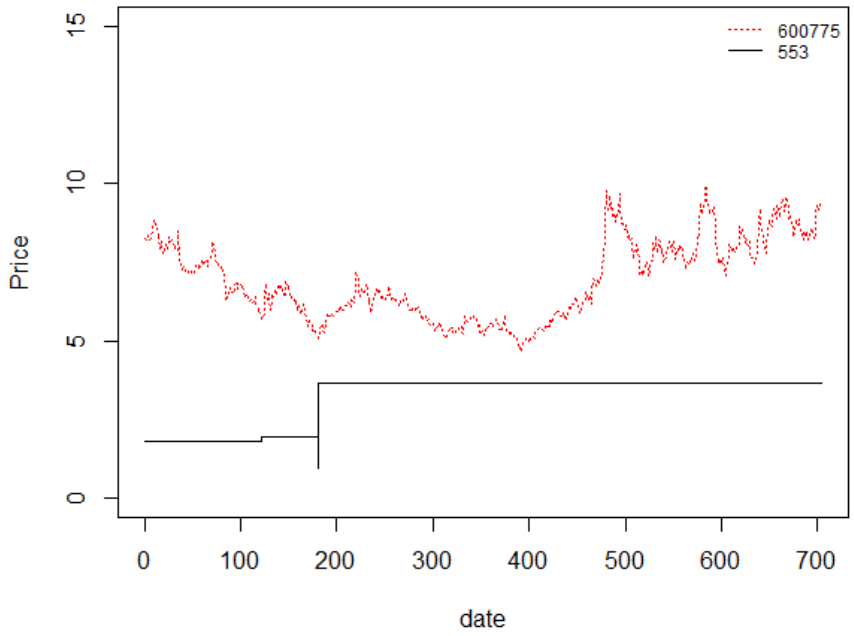

Figure 4.17: Patterns comparsion-600775-553

In the high premium cluster, the dynamic time-warping algorithm can not match the pattern of these stocks, because the distance between the prices of both A- and H- shares is too far. In other words, the price difference between these stocks is too large. The pattern dissimilarity exists concurrently in the entire training period. The study tests the factors, comparing the company with the high premium cluster. The result is the market conditions of the A- and H- share markets can influence the dissimilarity obviously. The high premium cluster growth rate is obtained by controlling for the changes in the overall market conditions. Other factors have an effect on the high premium growth rate, such as the trading shares. The details on this growth rate are described in Table 4.24.

Table 4.24: Factors comparsion-600775-553

	600775-553	High premium
CAP	√	√
TUR	√	√
V		
TS	√	√
EPS		
VOL		
ER	√	√
MA^2	√	√
MH^2	√	√

Another company, also listed in A- and H- share markets, it is called 600874 and 1065 respectively. It can be classified into the public utilities industry. The company has a common performance about the pattern in high premium cluster, with a pattern fitting of 0, because the distance between them is too far, see Figure 4.18.

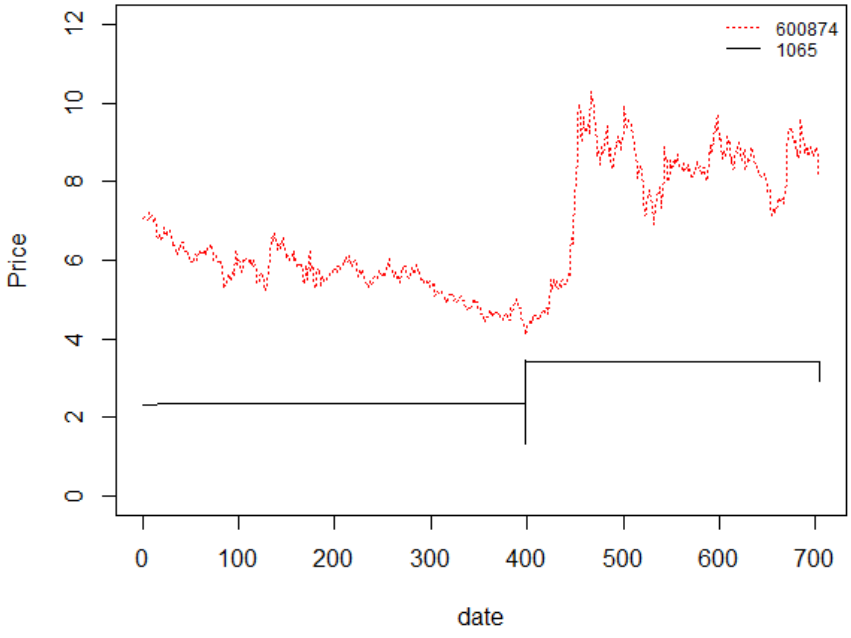

Figure 4.18: Patterns comparsion-600874-1065

Table 4.25: Factors comparsion-600874-1065

	600874-1065	High premium
CAP	√	√
TUR	√	√
V		
TS		√
EPS		
VOL	√	
ER		√
MA^2	√	√
MH^2	√	√

The comparison result is shown in Table 4.25, and it explains that the pattern dissimilarity is caused by the market factors rather than fundamental factors. At the same time, the study chooses the final case to state the pattern performance of the high premium cluster. The final company, called the Shenji Group Kunming Machine Tool Company Limited, was listed in H- share market in 1993, after one year it was listed in A- share market. Its stock codes are 600806 and 300 of the A- and H- share markets respectively. Furthermore, this company has the same performance as that of the two preceding companies. The price gaps are also too large of them, and the price of the A- share is much higher than that of H- share. Thus, the pattern is dissimilar during the period, see Figure 4.19.

Table 4.26: Factors comparsion-600806-300

	600806-300	High premium
CAP	√	√
TUR	√	√
V	√	
TS		√
EPS		
VOL		
ER	√	√
MA^2	√	√
MH^2	√	√

Additionally, this performance indicates that the market factors have more influence on the high premium cluster than others, such as the turnover, exchange rate

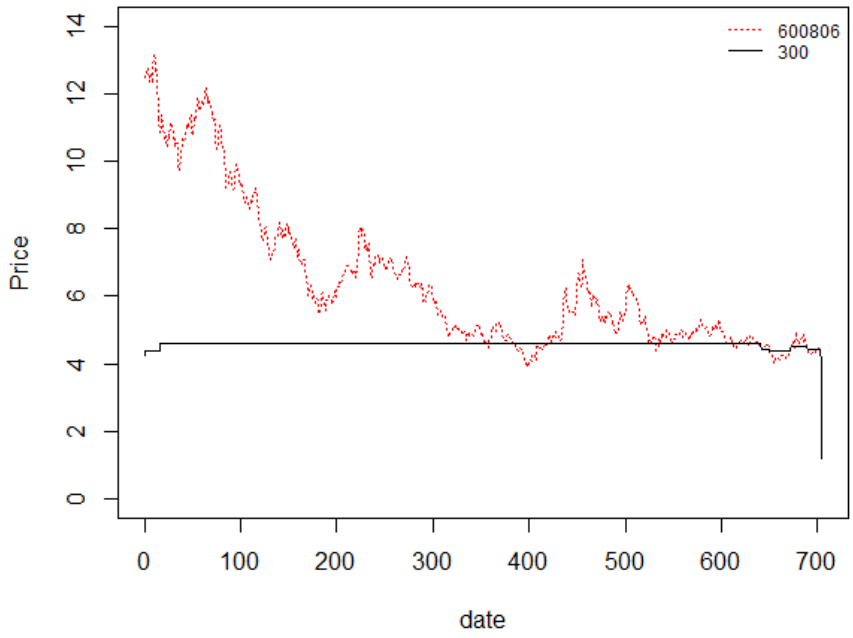

Figure 4.19: Patterns comparsion-600806-300

and the sensitivity to the A- and H- share markets. Table 4.26 shows the details about it.

This subsection shows that almost companies have the pattern dissimilarity or similarity, and the pattern matching of non-premium sort is better than that of others, including low and high premium groups. In addition, the pattern dissimilarity of the non-premium cluster is caused by mainly fundamental factors of the company, and marketable factors influence the low and high premium obviously.

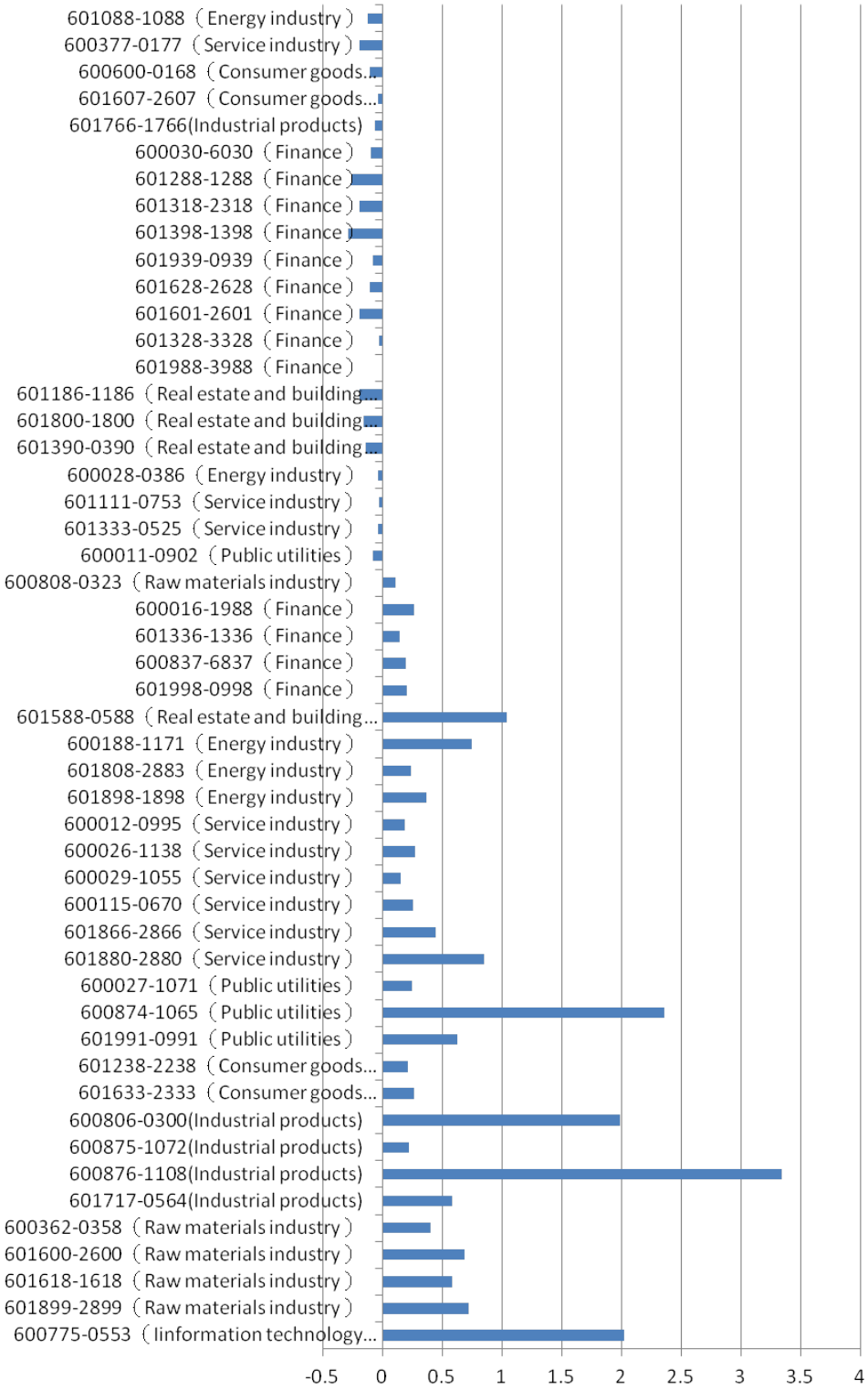

Figure 4.20: Performance of different industries based on the price differences

Table 4.27: The results of three companies in non-premium

	601601-2601/D	601601-2601/S	601939-939/D	601939-939/S	601186-1186/D	601186-1186/S
CAP	4.58E-03	3.13E-03	5.36E-03	4.24E-03	3.29E-03	2.11E-03
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
TUR	1.18E-03	2.01E-03	5.32E-03	7.80E-03	1.31E-03	1.94E-03
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
V	1.99E-04	1.77E-04			8.14E-05	7.23E-05
	(0.000)***	(0.011)**			(0.002)**	(0.011)*
TS			-2.17E-04	-2.95E-04		
			(0.086).	(0.052).		
EPS	-3.10E-02	-4.08E-02	-3.04E-02	-4.23E-02	-2.95E-02	-3.99E-02
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
VOL		-3.32E-04				
		(0.038)*				
ER						3.10E-02
						(0.071).
MA^2				2.51E-02		
				(0.068).		
MH^2		-4.60E-03	-3.44E-03	-3.60E-03	-3.28E-03	
		(0.056).	(0.064).	(0.053).	(0.064).	

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Table 4.28: The results of three companies in low premium

	601898-1898/D	601898-1898/S	600012-995/D	600012-995/S	601998-998/D	601998-998/S
CAP	-9.34E-03	-1.29E-02	-1.27E-02	-1.98E-02	-2.11E-02	-2.08E-02
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
TUR	5.86E-03	5.02E-03	4.02E-03	3.98E-03	2.14E-03	1.84E-03
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
V			7.30E-05			
			(0.066).			
TS			-2.98E-02	-3.40E-02		-5.50E-02
			(0.019)*	(0.013)*		(0.018)*
EPS	-1.72E-03	-2.18E-03	-4.08E-03	-4.59E-03		
	(0.004)**	(0.000)***	(0.031)*	(0.018)*		
VOL						
ER	-2.96E-01	-3.32E-01	-1.96E-01	-2.29E-01	-1.79E-01	-1.44E-01
	(0.004)**	(0.000)***	(0.011)*	(0.007)**	(0.000)***	(0.000)***
MA^2	-1.55E-01	-3.34E-01	-1.74E-01	-2.95E-01	-1.48E-01	-2.63E-01
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
MH^2	2.43E-02				1.07E-02	
	(0.040)*				(0.012)*	

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Table 4.29: The results of three companies in high premium

	600775-553	600874-1065	600806-300
CAP	-2.99E-02	-3.44E-02	-3.15E-02
	(0.000)***	(0.000)***	(0.000)***
TUR	8.44E-03	1.95E-02	1.01E-02
	(0.000)***	(0.000)***	(0.000)***
V			5.15E-05
			(0.076).
TS	3.14E-02		
	(0.002)**		
EPS			
VOL		9.85E-03	
		(0.076).	
ER	8.46E-01		8.40E-01
	(0.087).		(0.098).
MA^2	-2.98E-01	-3.04E-01	-3.69E-01
	(0.000)***	(0.000)***	(0.000)***
MH^2	1.78E-02	1.76E-02	1.10E-02
	(0.000)***	(0.000)***	(0.000)***
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1			

Chapter 5

Conclusions and Future work

In this chapter, we draw conclusions on the thesis, and point out some possible research directions.

5.1 Conclusions

This research attempted to use cluster analysis to divide the AH Premium Index into the three different clusters. Three models were built through factor analysis and regression test to explain the differences among them. The study then used the dynamic time-warping algorithm to fit the patterns of the different clusters. Several results were discovered in the following areas:

1. In general, the thesis introduced two applications of data mining and traditional statistical methods to analyze price behaviors in the segmented stock markets. The cluster analysis for the AH Premium Index is useful and valuable in investigating the characteristics of different industries, and the application of dynamic time-warping algorithm is helpful in fitting the pattern in the segmented stock markets from the technical angle.

2. The panel data regressions were run to examine the different effects of the main factors on the non-premium, low premium and high premium clusters. In par-

ticularly, the information asymmetry, trading liquidity, and market conditions are three prominent factors for the different clusters. It is observed that the A- and H- share price gaps of large-cap stocks are generally smaller than those of small-cap stocks. Based on these common factors, the shares in the premium cluster generally prefer small-cap stocks. Another finding is the two premium clusters are comprised of the industrial manufacturing industry and the energy industry. However, blue-chip stocks as well as the financial industry and real estate form the non-premium cluster.

3. More specifically, ANOVA shows that the price of A-share is much higher than that of H-share in premium clusters. Additionally, information asymmetry, liquidity differences, market condition of A-share, are main factors that affect the premium clusters. Conversely, in the non-premium cluster, the price of H-share is marginally higher than that of A-share. Some fundamental factors dominate in the non-premium sector, including information asymmetry, and investment philosophy differences.

5.2 Future work

Related topics for future research work are listed below.

1. In the future, more features of the pricing behaviors can be explored, and the study can distinguish and analyze them based on different markets, such as bull and bear markets.

2. Other dual-listed companies in other markets, such as A-S, A-N, and A-L share markets can also be investigated.

3. Given the price difference, making the trading portfolios and different arbitrage strategies for different clusters will be an available option.

4. Future studies will improve the stable and profitable trading strategy further, by applying the dynamic time-warping algorithm and various technical indicators in the stock market.

Bibliography

Achtert, E., Böhm, C., Kröger, P., and Zimek, A. (2006), "Mining hierarchies of correlation clusters," in *18th International Conference on Scientific and Statistical Database Management*, pp. 119–128, IEEE.

Admati, A. R. and Pfleiderer, P. (1988), "A theory of intraday patterns: Volume and price variability," *Review of Financial Studies*, 1, 3–40.

Aharony, J., Lee, C. W. J., and Wong, T. J. (2000), "Financial packaging of IPO firms in China," *Journal of Accounting Research*, 38, 103–126.

Akaike, H. (1987), "Factor analysis and AIC," *Psychometrika*, 52, 317–332.

Amihud, Y. (2002), "Illiquidity and stock returns: cross-section and time-series effects," *Journal of Financial Markets*, 5, 31–56.

Amihud, Y. and Mendelson, H. (1986), "Asset pricing and the bid-ask spread," *Journal of Financial Economics*, 17, 223–249.

Amihud, Y. and Mendelson, H. (1989), "The effects of beta, bid-ask spread, residual risk, and size on stock returns," *The Journal of Finance*, 44, 479–486.

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. M. (2014), "A tensor approach to learning mixed membership community models," *The Journal of Machine Learning Research*, 15, 2239–2312.

Arquette, G. C., Brown, W. O., and Burdekin, R. C. K. (2008), "US ADR and Hong Kong H-share discounts of Shanghai-listed firms," *Journal of Banking & Finance*, 32, 1916–1927.

Bahlmann, C. and Burkhardt, H. (2004), "The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 299–310.

Bailey, W. and Jagtiani, J. (1994), "Foreign ownership restrictions and stock prices in the Thai capital market," *Journal of Financial Economics*, 36, 57–87.

Ballard, D. H. (1981), "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition*, 13, 111–122.

Bergström, C. and Tang, E. (2001), "Price differentials between different classes of stocks: An empirical study on Chinese stock markets," *Journal of Multinational Financial Management*, 11, 407–426.

Berndt, D. J. and Clifford, J. (1994), "Using dynamic time warping to find patterns in time series," in *Workshop on Knowledge Discovery and Data Mining*, vol. 10, pp. 359–370, Seattle, WA.

Bonzo, D. C. and Hermosilla, A. Y. (2002), "Clustering panel data via perturbed adaptive simulated annealing and genetic algorithms," *Advances in Complex Systems*, 5, 339–360.

Burnham, K. P. and Anderson, D. R. (2004), "Multimodel inference understanding AIC and BIC in model selection," *Sociological Methods and Research*, 33, 261–304.

Chan, K. (1993), "Imperfect information and cross-autocorrelation among stock prices," *The Journal of Finance*, 48, 1211–1230.

Chan, K., Menkveld, A. J., and Yang, Z. (2008), "Information asymmetry and asset prices: Evidence from the China foreign share discount," *The Journal of Finance*, 63, 159–196.

Chang, E. C., Cheng, J. W., and Yu, Y. (2007), "Short-sales constraints and price discovery: Evidence from the Hong Kong market," *The Journal of Finance*, 62, 2097–2121.

Chen, G. M., Lee, B. S., and Rui, O. (2001), "Foreign ownership restrictions and market segmentation in China's stock markets," *Journal of Financial Research*, 24, 133–155.

Chen, H. (1998), "Price limits, overreaction, and price resolution in futures markets," *Journal of Futures Markets*, 18, 243–263.

Chen, Z. and Xiong, P. (2001), "Discounts on illiquid stocks: Evidence from China," *Yale International Center for Finance Working Paper*, pp. 00–56.

Chung, F. L., Fu, T. C., Ng, V., and Luk, R. W. P. (2004), "An evolutionary approach to pattern-based time series segmentation," *IEEE Transactions on Evolutionary Computation*, 8, 471–489.

DeCarlo, S. (2011), "The world's biggest public companies," *Forbes*.

Domowitz, I., Glen, J., and Madhavan, A. (1997), "Market segmentation and stock prices: Evidence from an emerging market," *The Journal of Finance*, 52, 1059–1085.

Efrat, A., Fan, Q., and Venkatasubramanian, S. (2007), "Curve matching, time warping, and light fields: New algorithms for computing similarity between curves," *Journal of Mathematical Imaging and Vision*, 27, 203–216.

Errunza, V. and Losq, E. (1985), "International asset pricing under mild segmentation: Theory and test," *The Journal of Finance*, 40, 105–124.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996), "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Knowledge Discovery and Data Mining*, vol. 96, pp. 226–231.

Fama, E. F. and French, K. R. (1993), "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

Fernald, J. and Rogers, J. H. (2002a), "Puzzles in the Chinese stock market," *Review of Economics and Statistics*, 84, 416–432.

Fernald, J. and Rogers, J. H. (2002b), "Puzzles in the Chinese stock market," *Review of Economics and Statistics*, 84, 416–432.

Fisher, D. H. (1987), "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, 2, 139–172.

Frey, B. J. and Dueck, D. (2007), "Clustering by passing messages between data points," *Science*, 315, 972–976.

Gelman, A. (2005), "Analysis of variancewhy it is more important than ever," *The Annals of Statistics*, 33, 1–53.

Han, D. (2006), "Study on factors of soft segmentation of A and H share market and on steps and time of setting up QDII," *Business Economics and Administration*, 173, 42–46.

Hietala, P. T. (1989), "Asset pricing in partially segmented markets: Evidence from the finnish market," *The Journal of Finance*, 44, 697–718.

Hjaltason, G. R. and Samet, H. (2003), "Index-driven similarity search in metric spaces," *ACM Transactions on Database Systems*, 28, 517–580.

Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynov, M., Kysel, J., and Tveito, O. E. (2008), "Classifications of atmospheric circulation patterns," *Annals of the New York Academy of Sciences*, 1146, 105–152.

Huttenlocher, D. P. and Ullman, S. (1987), "Object recognition using alignment," in *Proc. ICCV*, vol. 87, pp. 102–111.

Jacobs, C. E., Finkelstein, A., and Salesin, D. H. (1995), "Fast multiresolution image querying," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 277–286, ACM.

Jianping, Z. and Minken, C. (2007), "The cluster analysis of panel data and its application," *Statistical Research*, 4, 11–14.

Karolyi, G. A. (2006), "The world of cross-listings and cross-listings of the world: Challenging conventional wisdom," *Review of Finance*, 10, 99–152.

Kim, M., Szakmary, A. C., and Mathur, I. (2000), "Price transmission dynamics between ADRs and their underlying foreign securities," *Journal of Banking & Finance*, 24, 1359–1382.

Koller, D. and Friedman, N. (2009), *Probabilistic graphical models: principles and techniques*, MIT press.

Kriegel, H. P., Kröger, P., Sander, J., and Zimek, A. (2011), "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 231–240.

Leigh, W., Purvis, R., and Ragusa, J. M. (2002), "Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support," *Decision Support Systems*, 32, 361–377.

Li, Y., Yan, D., and Greco, J. (2006), "Market segmentation and price differentials between A shares and H shares in the Chinese stock markets," *Journal of Multinational Financial Management*, 16, 232–248.

Lin, X. (2004), "An empirical study on discount of H Shares and information asymmetry," *The Study of Finance and Economics*, 30, 39–49.

Mei, J. P. and Chen, L. (2010), "Fuzzy clustering with weighted medoids for relational data," *Pattern Recognition*, 43, 1964–1974.

Melvin, M. (2003), "A stock market boom during a financial crisis?: ADRs and capital outflows in Argentina," *Economics Letters*, 81, 129–136.

Michalski, R. S. and Stepp, R. E. (1983), *Learning from observation: Conceptual clustering*, Springer.

Mok, H. M. K. and Hui, Y. V. (1998), "Underpricing and aftermarket performance of IPOs in Shanghai, China," *Pacific-Basin Finance Journal*, 6, 453–474.

Nandyala, S. P. and Kumar, T. K. (2010), "Real time isolated word speech recognition system for human computer interaction," *International Journal of Computer Applications*, 12, 1–7.

Obizhaeva, A. A. and Wang, J. (2013), "Optimal trading strategy and supply/demand dynamics," *Journal of Financial Markets*, 16, 1–32.

Peng, W., Miao, H., and Chow, N. (2008), "Price convergence between dual-listed A and H shares," *Macroeconomic Linkages between Hong Kong and Mainland China*, pp. 295–315.

Rai, P. and Daume, H. (2010), "Infinite predictor subspace models for multitask learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 613–620.

Rousseeuw, P. J. (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, 20, 53–65.

Sakoe, H. and Chiba, S. (1978), "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26, 43–49.

Salvador, S. and Chan, P. (2007), "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, 11, 561–580.

Seasholes, M. S. and Liu, C. (2011), "Trading imbalances and the law of one price," *Economics Letters*, 112, 132–134.

Shorten, G. P. and Burke, M. J. (2011), "The application of dynamic time warping to measure the accuracy of ECG compression," *International Journal of Circuits, Systems and Signal Processing*, 5, 305–313.

Silber, W. L. (1991), "Discounts on restricted stock: The impact of illiquidity on stock prices," *Financial Analysts Journal*, 47, 60–64.

Solnik, B. H. (1974), "The international pricing of risk: An empirical investigation of the world capital market structure," *The Journal of Finance*, 29, 365–378.

Stulz, R. M. and Wasserfallen, W. (1995), "Foreign equity investment restrictions, capital flight, and shareholder wealth maximization: Theory and evidence," *Review of Financial Studies*, 8, 1019–1057.

Su, D. and Fleisher, B. M. (1999), "Why does return volatility differ in Chinese stock markets?" *Pacific-Basin Finance Journal*, 7, 557–586.

Sun, Q. and Tong, W. H. S. (2000), "The effect of market segmentation on stock prices: The China syndrome," *Journal of Banking & Finance*, 24, 1875–1902.

Tappert, C. C., Suen, C. Y., and Wakahara, T. (1990), "The state of the art in online handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12, 787–808.

Umeyama, S. (1993), "Parameterized point pattern matching and its application to recognition of object families," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 136–144.

Vattani, A. (2011), "K-means requires exponentially many iterations even in the plane," *Discrete and Computational Geometry*, 45, 596–616.

V.Estivill-Castro (2002), "Why so many clustering algorithms: a position paper," *ACM SIGKDD Explorations Newsletter*, 4, 65–75.

Wang, S. S. and Jiang, L. (2004), "Location of trade, ownership restrictions, and market illiquidity: Examining Chinese A-and H-shares," *Journal of Banking & Finance*, 28, 1273–1297.

Wolfson, H. J. and Rigoutsos, I. (1997), "Geometric hashing: An overview," *Computing in Science and Engineering*, 4, 10–21.

Yang, J. (2003), "Market segmentation and information asymmetry in Chinese stock markets: a VAR analysis," *Financial Review*, 38, 591–609.

Zhanchi, W. (2007), "Price difference between H Share and A Share based on the difference in value ideas: Interpretation and evidence," *Finance & Economics*, 6, 16–23.