The style and structure of the stock markets in China:
an application to PCA for interval symbolic data

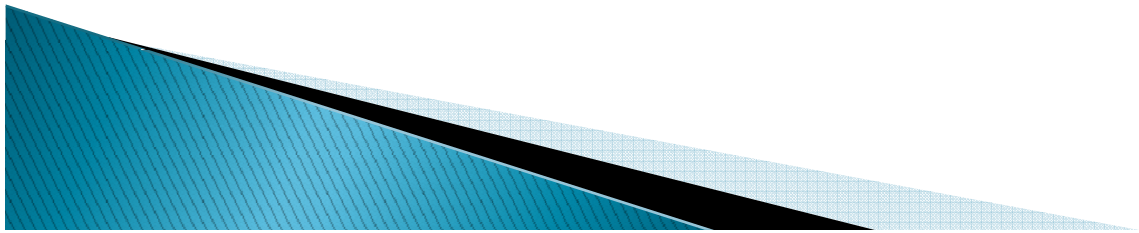
Wen Long

Research Center On Fictitious Economy & Data Science,
Chinese Academy of Sciences

longwen@gucas.ac.cn

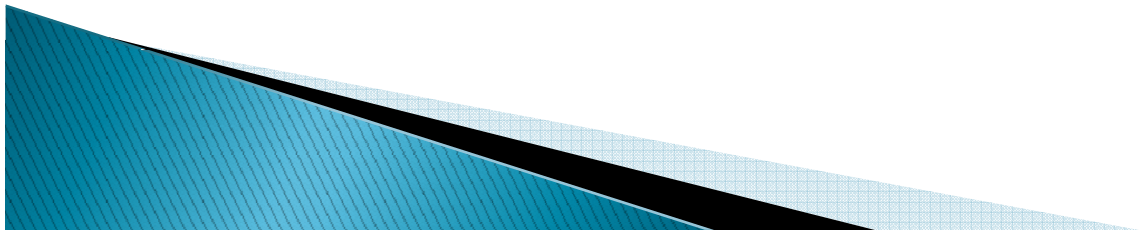
Outline

- Introduction
- The Chinese stock market
- Methodology--SPCA
- Results and discussions
- Conclusions



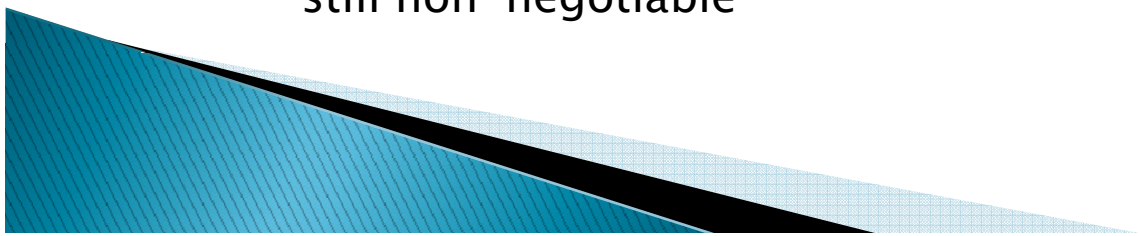
Introduction

- China has two stock markets--in Shanghai and in Shenzhen
- The two emerging stock markets commenced operations in 1990
- Developed rapidly during 30 years
- By the end of Aug. 2011,
  - the total market value of the two Chinese stock markets was 25,461 billion Yuan (1 EUR≈9 CNY)
  - the average value of daily stock turnover is about 161 billion Yuan (≈US$25 billion)
  - there are 2,273 listed companies
  - there are over 137 million registered investors in China



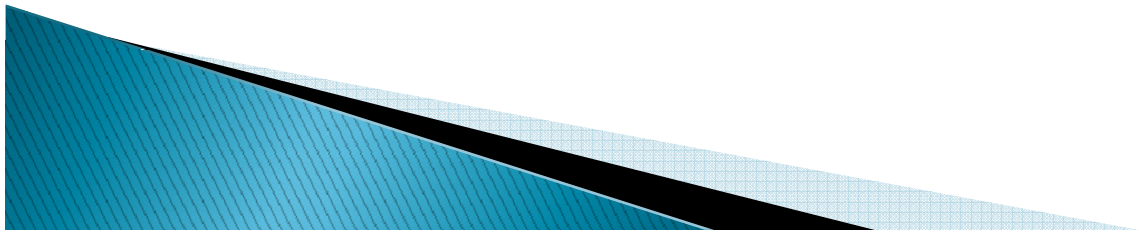
- Though thriving, China's stock market is often criticized for being over-speculative and inefficient
  - The volatility is high comparing with the other world's major stock markets
  - The turnover rate is high

- Observed that emerging markets have unusual features
  - Negative relation between stock volatility and stock returns in the early operating years of China's stock markets
  - There are essentially two types of Chinese stocks: negotiable and non-negotiable
    - The former represents those outstanding shares in circulation, and the latter is state owned or owned by legal entities and are not tradable
  - Two-thirds of the stocks were held by agents of the state
    - In 2005, 72.36% of the stocks in 1381 listed shares in China were still non-negotiable



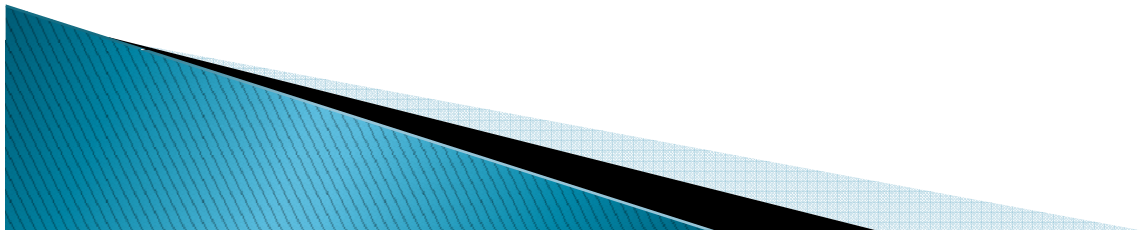
5 financial variables

- Market capitalization of negotiable shares (NMC)
- Return rate (R)
- P/E ratio
- Turnover rate (T)
  - that is equal to trading volume divided by outstanding negotiable shares
- Beta
  - That is calculated from the CAPM model
  - This beta is the non-diversifiable risk of the stock in question



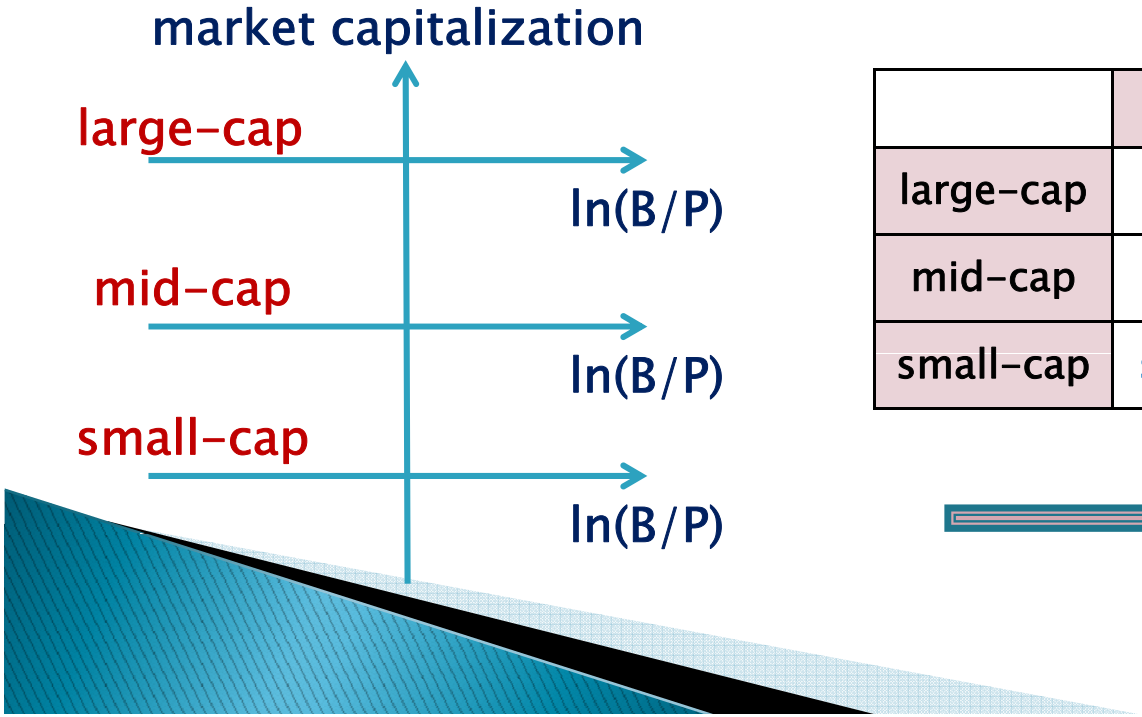
Value and Growth

- Stocks are commonly classified as value and growth
  - Value stocks have high B/P ratios or low P/E ratios
  - Growth stocks are represented by high P/E ratios
  - Value stocks are priced low relative to the company's fundamental value that is measured usually by price-to-earning (P/E) ratio, book-to-price (B/P) ratio, or book-to-market (B/M) ratio



6 stock style portfolios

- In this study, we use six stock styles to classify the two Chinese stock markets
  - large-cap value stocks, mid-cap value stocks, small-cap value stocks, large-cap growth stocks, mid-cap growth stocks, and small-cap growth stocks



	Growth	Value
large-cap	large-cap growth	large-cap value
mid-cap	mid-cap growth	mid-cap value
small-cap	small-cap growth	small-cap value

- We use the aforementioned financial variables to study the latent structure of the Chinese stock markets
- **Problems**
  - Very large amount of data
    - Over 1,000 listed companies, 5 financial index, over 10 years
  - Listed company merger, restructuring, renamed, listing and delisting, which causes
    - the amount of data in a sample change
    - name of the individual change
    - content of the individual change
- Traditional methods have some difficulty in dealing with the above problems

How to solve these problems?

Symbolic data analysis (SDA) offered an effective solution.

Symbolic data

Let $i=1,\ldots,n$ denote n objects described by p interval-type variable $Y_1,\ldots,Y_p$. Without loss of generality, consider $\xi_{ij} = |\underline{x}_{ij}, \overline{x}_{ij}|$ as the interval (interquartile) value of variable j for the object i, where $\underline{x}_{ij}$ and $\overline{x}_{ij}$ represent the lower (first quartile) and upper (third quartile) values. Let $x_1' = (\xi_{1i}, \ldots \xi_{1p}) = \left(\left[\underline{x}_{i1}, \overline{x}_{ip}\right], \ldots, \left[\underline{x}_{ip}, \overline{x}_{ip}\right]\right)$ denote the symbolic data vector obtained for object i. The resulting symbolic data matrix $\underline{X}$ is then given by:

$$\underline{X} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} \zeta_{11} & \cdots & \zeta_{1p} \\ \vdots & \ddots & \vdots \\ \zeta_{n1} & \cdots & \zeta_{np} \end{pmatrix} = \begin{pmatrix} \left[\underline{x}_{11}, \overline{x}_{11}\right] & \cdots & \left[\underline{x}_{1p}, \overline{x}_{1p}\right] \\ \vdots & \ddots & \vdots \\ \left[\underline{x}_{n1}, \overline{x}_{n1}\right] & \cdots & \left[\underline{x}_{np} \cdot \cdot \overline{x}_{np}\right] \end{pmatrix} \qquad (1)$$

Build symbolic data

- We employ the first-quartile and the third-quartile values to build interval data which is an typical type of symbolic data.
  - Interval-type data series conveys more information than the single-value data series in classical data analysis, for it contains both the central tendency and dispersion information of a variable.
  - It could enrich the interpretation of the analysis
  - Using the first-quartile and third-quartile values can avoids the biases introduced by extreme values at both ends of the distribution, and thus enhances the robustness of the estimation
  - It succeeds in avoiding the problems of amount of data change, individuals' name change
- For each of the 10 years from 1996 to 2005, we form six symbolic data vectors of the stock style portfolios (n=6) for the five interquartile financial variables (p=5).

Traditional numerical data table

	NMC	T	R	PE	Beta
Stock 1	a_1	b_1	c_1	d_1	e_1
Stock 2	a_2	...	...	...	...
Stock 3	a_3				
Stock 4	a_4				
...	...				
Stock n	a_n	b_n	c_n	d_n	e_n

Symbolic data table

	NMC	T	R	PE	Beta
large-cap growth	$[\underline{a}_1, \overline{a}_1]$	$[\underline{b}_1, \overline{b}_1]$	$[\underline{c}_1, \overline{c}_1]$	...	$[\underline{e}_1, \overline{e}_1]$
large-cap value	$[\underline{a}_2, \overline{a}_2]$	$[\underline{b}_2, \overline{b}_2]$	...		...
mid-cap growth	...	...			
mid-cap value					
small-cap growth					
small-cap value					$[\underline{e}_6, \overline{e}_6]$

SPCA

Similar to the classical PCA, the aim of SPCA is to describe the objects i by a reduced number $s<p$ of the new interval variables.

The $x_i' = (\zeta_{1i}, \ldots, \zeta_{1p}) = \left(\left[\underline{x}_{i1}, \overline{x}_{i1}\right], \ldots, \left[\underline{x}_{ip}, \overline{x}_{ip}\right]\right)$ data are visualized in the description space $\Re^p$ by a hyperrectangle with 2^p vertices. Moreover, a hyperrectangle in the p-dimensional space can be described by a matrix with 2^p rows and p columns, where each row contains the coordinates of one vertex of the hyperrectangle in $\Re^p$.

In the general case of a dimension p, the object i can be characterized by a matrix M_i that has $2^p \times p$ dimension:

$$M_i = \begin{bmatrix} \underline{x}_{i1} & \cdots & \underline{x}_{ip} \\ \vdots & \ddots & \vdots \\ \overline{x}_{i1} & \cdots & \overline{x}_{ip} \end{bmatrix}$$

Thus, the original matrix characterized by Eq. (1) can be described by the following numerical matrix M of interval values:

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_n \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \underline{x}_{11} & \cdots & \underline{x}_{1p} \\ \vdots & \ddots & \vdots \\ \overline{x}_{11} & \cdots & \overline{x}_{1p} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x}_{n1} & \cdots & \underline{x}_{np} \\ \vdots & \ddots & \vdots \\ \overline{x}_{n1} & \cdots & \overline{x}_{np} \end{bmatrix} \end{pmatrix}_{n \cdot 2^p \times p}.$$

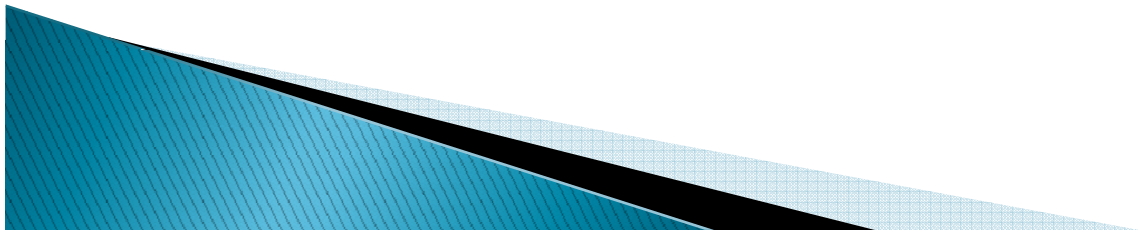
We can then apply the classical PCA method to this $n \cdot 2^p$ rows of the new numerical matrix M of interval (interquartile) values, with a suitable choice of dimension $s \leq p$. Let $Y_1^*, Y_2^*, \ldots, Y_S^*$ denote the first s numerical principal components. We will construct the interval-type principal components $Y_1^I, Y_2^I, \ldots, Y_S^I$ from these numerical principal components $Y_1^*, Y_2^*, \ldots, Y_S^*$.

Let L_i be the set of row indices in matrix M that refers to the matrix M_i corresponding to the i^{th} symbolic object x_i. For $k \in L_i$, let y_{kv} be the value of the numerical principal component Y_v^* with row index k. The value of the interval-type principal component Y_v^I for the i^{th} object is then denoted by $y_{iv} = \left[\underline{y}_{iv}, \overline{y}_{iv}\right]$, where:

$$\underline{y}_{iv} = \min_{k \in L_i} y_{kv} \quad \text{and} \quad \overline{y}_{iv} = \max_{k \in L_i} y_{kv}$$

Results and discussions

- Factor analysis extracts the combinations of variables that explain the most variance in descending order.
- We perform a varimax rotation on the factors extracted by SPCA so that the sum of the variances of the loadings is the maximum possible.
- We have extracted two factors to characterize the innate structure of the two Chinese stock markets. The cumulative contributions of the two factors, which range from 53.7% to 71.4% over the 10 years, are considered satisfactory and significant.



Factor loadings and cumulative contribution percentages—after varimax rotation (1996–2005)

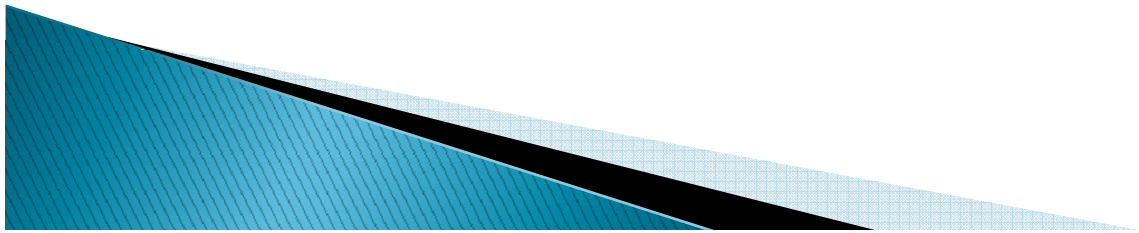
Year	Correlations	NMC	Return	*P/E*	Turnover	Beta	Contribution%	Cumulative contribution%
1996	Factor 1	0.7706*	0.2003	−0.7418*	−0.2111	0.6998*	34.3718	34.3718
	Factor 2	0.1875	0.8727*	0.2044	0.8497*	−0.0024	31.2099	65.5817
1997	Factor 1	0.7595*	0.0758	−0.6768*	−0.5763*	0.6672*	36.3580	36.3580
	Factor 2	0.3273	0.8849*	0.3881	0.0434	0.3124	22.8053	59.1632
1998	Factor 1	−0.7145*	0.6611*	0.7562*	0.8066*	0.1650	43.9467	43.9467
	Factor 2	−0.4910	0.4479	0.2011	0.0600	0.9424*	27.4796	71.4263
1999	Factor 1	0.8223*	−0.0018	−0.6620*	−0.6103*	0.6725*	38.7794	38.7794
	Factor 2	0.0553	0.9955*	0.0749	0.0265	0.0523	20.0611	58.8406
2000	Factor 1	−0.8206*	0.7159*	0.5268*	0.7898*	−0.0356	41.7666	41.7666
	Factor 2	0.1533	0.0819	−0.3390	−0.0742	0.9561*	21.2969	63.0635
2001	Factor 1	−0.7903*	0.0382	0.6799*	0.7230*	0.5996*	39.4101	39.4101
	Factor 2	−0.1144	0.9939*	−0.0709	0.0364	0.0413	20.1795	59.5896
2002	Factor 1	−0.7035*	−0.1665	0.6864*	0.6701*	−0.2891	30.5300	30.5300
	Factor 2	0.2495	0.7557*	0.2338	0.0798	−0.6090*	21.3059	51.8359
2003	Factor 1	0.7648*	0.7897*	−0.5774*	0.0246	0.5355*	36.5873	36.5873
	Factor 2	0.1669	0.0896	0.2835	0.9520*	−0.0626	20.5304	57.1177
2004	Factor 1	−0.8140*	−0.2098	0.5607*	0.1385	0.7040*	30.7144	30.7144
	Factor 2	−0.0307	0.7609*	0.4619	0.5904*	−0.0909	23.0020	53.7164
2005	Factor 1	−0.7354*	0.0480	0.6710*	0.6975*	0.5326*	35.2718	35.2718
	Factor 2	0.3705	0.9182*	0.1627	0.1193	−0.3041	22.2697	57.5415

Notes: NMC = market capitalization of negotiable shares; Return = return rate; *P/E* = price/earning ratio; Turnover = turnover rates, and *Beta* = the systematic risk.

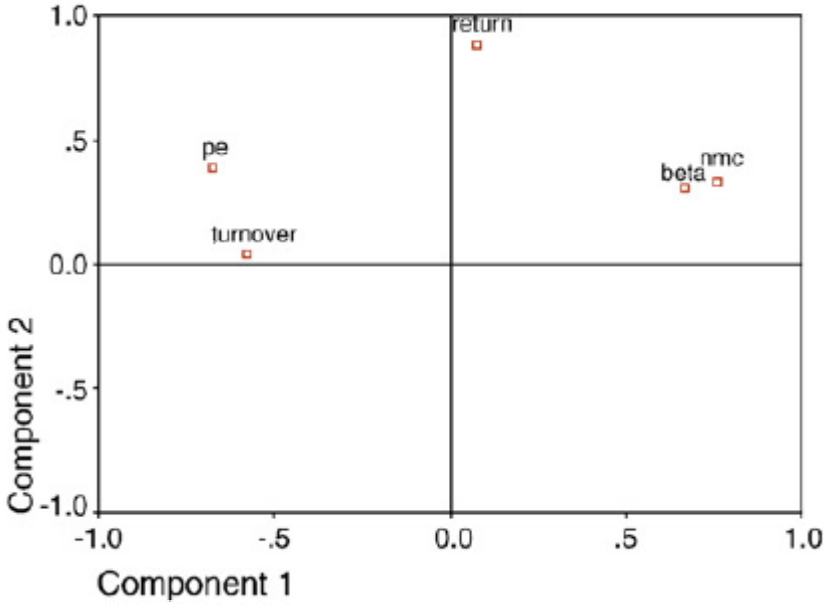
* Variables with factor loading of ±0.5 are considered constitute members of the factor. Factor loadings are the correlation of each variable and the factor.

- We identify the highly correlated group of variables (NMC, P/E ratios, turnover rates, and beta) as a “corporate-style” factor because of their high factor loadings in the first factor in seven out of the ten years of study.

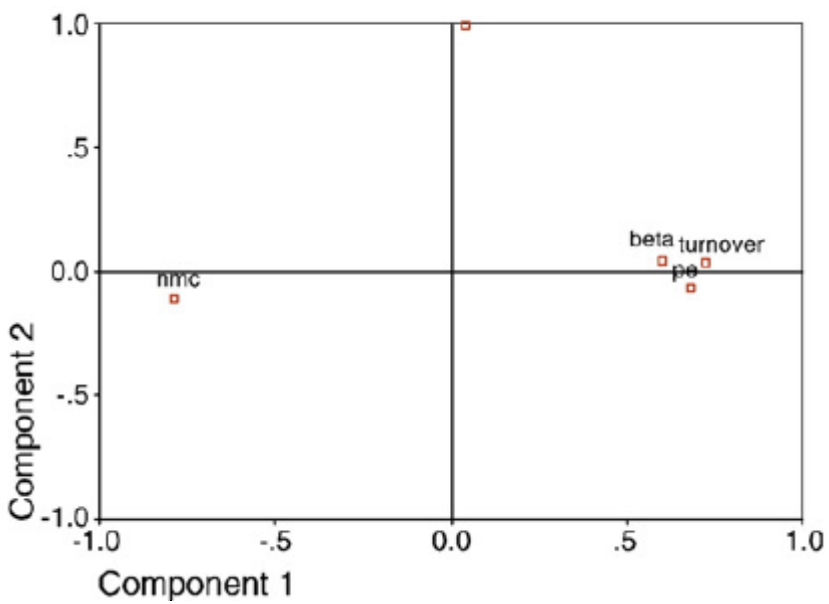
- we identify "corporate-performance" as a surrogate factor for the return variable as it obtains high factor loadings in the second factor in seven out of ten years.



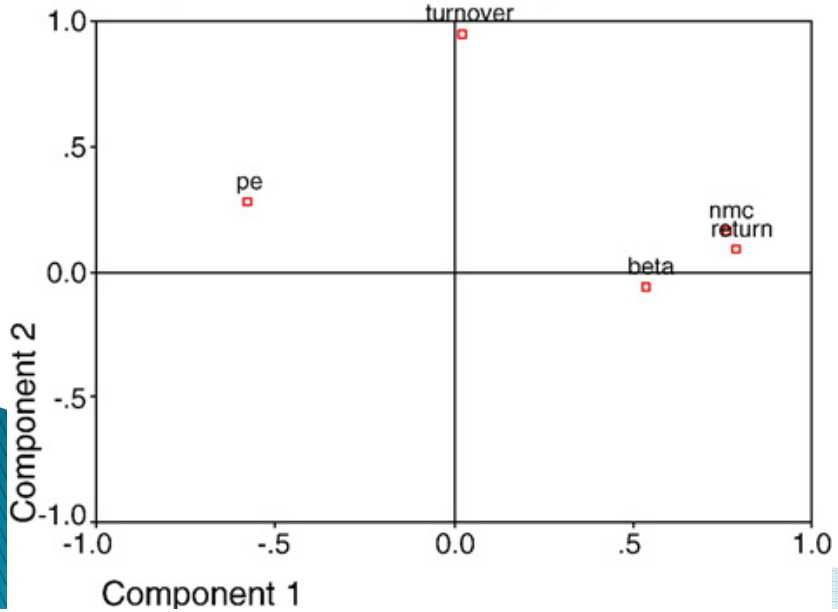
Factor Plot for 1997



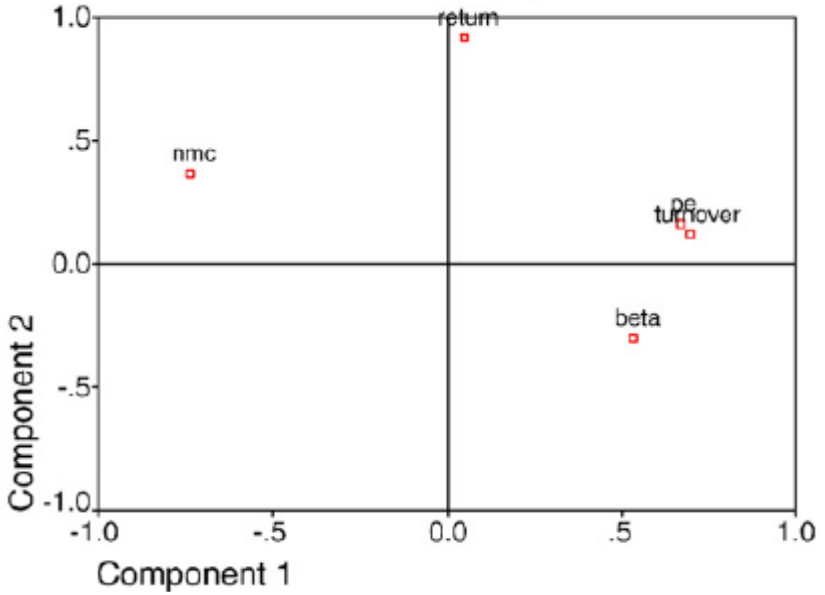
Factor Plot for 2001



Factor Plot for 2003



Factor Plot for 2005



- Among the constituent members of the corporate-style factor, the NMC variable is usually negatively related to other members of that factor.

- This negative relation seems to suggest that the lower the market size, the higher the values in P/E ratios, turnover rates, and beta.

- Since growth stocks are represented by high P/E ratios, we infer that small growth stocks tended to be actively traded and were associated with high beta in most of the 10 years of the study. In other words, small growth stocks, though more risky, are more attractive to investors in China.

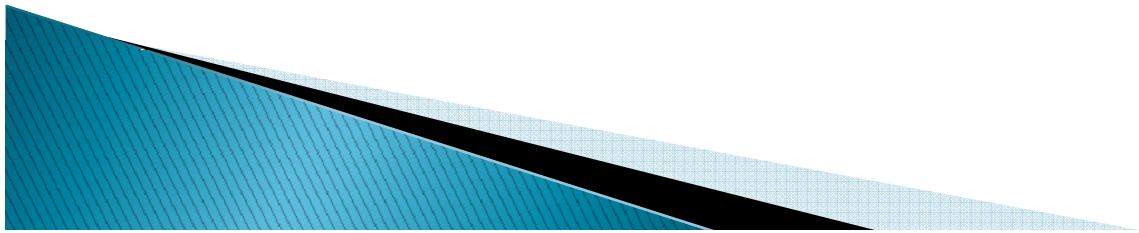
- The actively-traded small-cap growth stocks might have become quite speculative, meaning that the prices of the small-cap growth stocks might have been pushed above their intrinsic values.

Conclusions

- We apply the symbolic principal component analysis (SPCA) to 10 years (1996–2005) of interquartile values to identify the underlying structure in China's stock markets.
- Via SPCA, we extracted two factors, namely a corporate-style factor and a surrogate corporate-performance factor to characterize the innate structure in the Shanghai and Shenzhen stock markets.
  - The corporate-style factor consists of NMC, P/E ratio, turnover rates, and beta, while the market-performance is return rates.
- We observe a negative relation between beta (systematic risk) and return in all stock styles and in all years, with the exception of 2003.

- We draw on additional evidence from the six stock styles and the inter-temporal zoom-star diagrams of beta and return, and compute the return per beta indices (RPB) to measure the relative return adjusted for systematic risk.

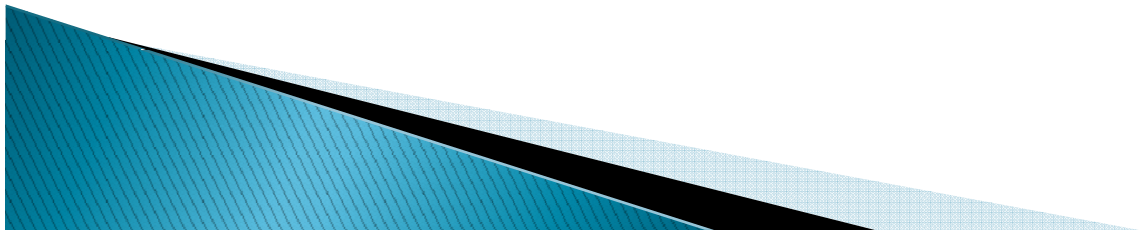
- The negative returns in recent years coincided with the bearish stock market that started in mid-2001. Overall, the seemingly peculiar negative beta/return relation is not truly anomalous but is due to a very high risk-free rate, pervasive negative returns, and stock styles.



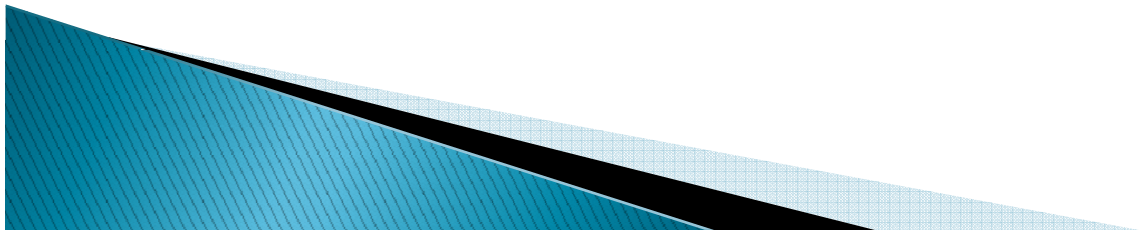
- Evidence also suggests a trend in recent years that the growth stocks of small market capitalization were more attractive to investors in China, and thus their trading is active, resulting in high speculation.

- This is alarming as small-cap stocks generally have an inferior EPS compared to large-cap stocks, and are also more prone to price manipulations.

- Such a trend would be precarious to the stable development of the stock markets in China.



- We traced China's stock market after 2005, using the same method.
- There are many changes in later stock market.
  - More and more new listed companies enter stock market, which means the market capitalization is much huger than before, and the investors have more choices.
  - Non-negotiable shares began to be tradable
- There are some similar results can be obtained.
  - Small-cap stock has higher turnover rate, P/E ratio, and lower EPS.
  - The phenomenon of investors pursuit of small-cap stocks that are easy to speculation haven't be improved in recent years.



Thank you