Garlic Chive Index and Prediction of China's Stock Market

by

Dongchen Zou

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Undergraduate College

Leonard N. Stern School of Business

New York University

May 2017

Professor Marti G. Subrahmanyam	Professor Yakov Amihud
Faculty Adviser	Thesis Adviser

Garlic Chive Index and Prediction of China's Stock Market

Dongchen Zou
Stern School of Business, New York University

May 10, 2017

Abstract

This paper aims to examine whether the Garlic Chives Index (GCI), the number of newly added brokerage accounts in China, provides information about investor sentiment and can predict stock return. This is the popular belief in China. Results show that GCI are significantly determined by lagged market returns given high lagged trading volumes. Also, GCI has little predictability over future market returns and factor returns, and yet the negative component of GCI changes have significant return predictability before the 2015 market crash. Finally, we compare GCI with China's implied volatility index iVX as a proxy of institutional investor sentiment, and find that similar to GCI, iVX has little return predictability. But iVX relatively better non-linear predictability.

JEL classification: G12

Keywords: Market Sentiment, New Accounts, China

1. Introduction

The number of newly-added brokerage accounts released every week by China Securities Depository and Clearing Co. (CSDC), also known as "Garlic Chives Index" (GCI), was brought to attention among Chinese investors in recent years. The index is named garlic chives with a reason that the chives, representing the retail investors, grow at a crazy speed.

I would like to thank Professor Yakov Amihud for offering meaningful suggestions and guidance in the Stern Honors Program and providing me invaluable inputs that has helped me compile this report. I also sincerely appreciate Professors Jennifer Carpenter, Robert Whitelaw and Venky Venkateswaran in providing thoughtful advice in my thesis.

And each time they grow sufficiently, the farmers, representing the institutional investors, will reap the harvest. The name suggests that many investors believe GCI as a reversal indicator of market, i.e., when the number of new accounts is substantially large and the increase of new accounts is abnormally high, the stock market is likely to crash in the near future. Such belief was strengthened after the 2015 crash in China's stock market.

A number of academic literature relates the number of newly added accounts as a proxy of investor sentiment. Previous studies on new brokerage accounts including Zhu and Niu (2016), Han and Li (2017) focused on its contribution to principal components in market sentiment and examined cross-sectional stock returns based on the methodology from Baker and Wurgler (2006). Other studies such as Li et al. (2008), and Kling and Gao (2008) treated the number of new accounts as a standalone measure of investor sentiment. In particular, given the short-sale constraints in the China's stock market, GCI represents a bullish sentiment among investors: when they are optimistic or speculative about the market, they will have tendency to open trading accounts to purchase stocks, vice versa.

One of the central debates over Investor sentiment including GCI is its stock return predictability. According to Efficient Market Hypothesis (EMH), investor sentiment rarely predicts of future returns, since if its predictive power is sufficiently high, then investors will trade based on sentiment's prediction and will wipe out the expected profits; thus in equilibrium sentiment is priced at present and provides little information about future returns. Brown and Cliff (2004) suggests that sentiment index provides little predictability of near-term stock returns. While research in behavioral finance suggests that abnormal investor sentiment as irrationality in the stock market will lead to deviations of the asset price from its fundamental value in the short-run, and the price adjusts in the long-run when sentiment returns to neutral. Specifically, investor sentiment is a contrarian predictor of future returns, i.e., high sentiment predicts low future stock returns in the long-term. Fisher and Statman (2000), Brown and Cliff (2005), Baker et al. (2012), and Huang et al. (2014) provide well-documented evidence that future returns over long-period horizons are nega-

tively associated with high investor sentiment. Also, investor sentiment has predictability over cross-sectional returns due to market revisions of speculative trades. Baker and Wurgler (2006) finds that bullish sentiment has negatively and significantly forecast subsequent stock returns over small stocks, low profitability stocks, young stocks, low dividend-yield stocks, and high volatility stocks. Similar results are found in China's stock market. Therefore, it will be interesting to see whether GCI as a standalone proxy of investor sentiment provides useful information in forecasting market returns as well as cross-sectional stock returns.

It is also noteworthy that GCI represents trading behaviors of retail investors, in that 99.71% of the brokerage accounts belongs to retail investors and among them 76.7% has less than 100,000 RMB ($14,526) in their accounts, according to SSE 2015 Annual Report. Retail investors are generally viewed as "noise traders", uniformed investors who initiate trades without knowledge of fundamentals. According to Schleifer and Summers (1990), these noise traders increase the limits of arbitrage and cause the asset prices to deviate from their intrinsic values. Based on this theory, China's stock market, in which above 80% of the trading volume comes from retail investors, will barely reflect the fundamentals. Yet recent studies such as Carpenter, Lu and Whitelaw (2015) finds that the stock prices in China's stock market are as informative as those in the developed market. The findings suggest that retail investors in aggregation may behave like informed traders and perform systematic valuation of assets. Thus it is interesting to understand the causes of noise trading in aggregate level, and compare the aggregate noise-trading with informed trading behaviors. By examining the determinants of GCI we may better understand the aggregate retail-trading behaviors, particularly the incentives for retail investors to initiate buy orders; and by comparing GCI with proxies of institutional investor sentiment such as implied volatility index we could sense the similarity or difference between informed trading patterns and aggregate uniformed trading.

In this paper, I examine the determinants of GCI, test its predictability over future market returns and factor returns, and finally compare the index with iVX, the implied

volatility index in China. I find that GCI may be significantly determined by lagged return in the most recent week. Also, the GCI has little predictive power over future market returns or factor returns, while non-linearity of predictability is significant. Finally, we find that iVX has relatively better return predictability over GCI. The thesis is organized as follows: Section 2 examines the determinants of GCI and its predictability on future market returns, together with discussion of its asymmetric predictive power. Section 3 investigates into GCI's predictive power over facto returns. Section 4 compares GCI and iVX and documents the predictability of implied volatility on future market returns.

2. Determinants of GCI and Return Predictability

The first question we would like to ask is what determines the opening of new brokerage accounts. Brown and cliff (2004) finds that investor sentiment is determined by past market returns. As a proxy of investor sentiment, it is likely that GCI is also driven by past index returns. Therefore, it is hypothesized that positive past market returns predicts an increase in GCI. Secondly, it is also possible that agents open new accounts and trade because they observe other agents opened accounts in the past, which we call "herd behavior". According to Bikhchandani and Sharma (2000), traders may imitate the trading behaviors of their similar group given imperfect information, as they believe that the behaviors of others reveal some information about the intrinsic value. Based on the herding theory, it is hypothesized that higher past opening of new accounts and past trading volumes predicts higher GCI. Thirdly, it is likely that past return volatilities attract more speculators or noise traders to participate in trades. Thus in hypothesis, higher past return volatilities predict higher GCI.

The second question is whether GCI is able to predict future market returns. In particular, whether the number of newly added accounts has near-term predictive power (within three weeks). We would also like to examine the asymmetry in predictability exists in GCI, i.e., whether the increase or the decrease in new accounts predicts market returns, and in

which directions respectively. Asymmetry test helps us better interpret the content of GCI.

2.1. Data

Data for the number of newly added brokerage accounts (GCI) are collected from the CSDC weekly announcements (`http://www.chinaclear.cn/zdjs/xmzkb/center_mzkb`) dated from January 11, 2008 to April 7, 2017. Note that starting from April 2015, the Shanghai Exchange allows investors to open multiple accounts under different securities firms. In response to the new policy, the CSDC adjusted the methods of computing GCI since May 2015, from counting the number of newly registered accounts, referred as **Period A** in the later sections, to counting the number of newly registered investors, referred as **Period B**. The new method avoids multiple counts on the same investor who have already opened an account. Although the data collected with traditional method are similar to those collected with new method, we will examine the two groups of data separately by s in the following sections. Weekly data for the Shanghai Stock Exchange (SSE) Index levels and trading volumes are collected from the Wind Financial Terminal.

2.2. Methodology

In answering the first question of what determines the opening of new brokerage accounts, we perform a Granger causality test of GCI on past returns, past trading volumes and past return volatilities based on our hypothesis. Notice that high synchronicity exists in GCI with the market index and trading volume which we can see in Figure 1 and Figure 2. Such high synchronicity can be explained in economics sense, as more new registered stock accounts are associated with higher trading volume, and more buyers in the market are associated with higher price.We calculate the compounding index return (R), the change in new accounts (∂GCI), and the change in trading volume ($\partial Volume$) as following:

$$R_t = ln\frac{IDX_t}{IDX_{t-1}}, \quad \partial Volume_t = ln\frac{Volume_t}{Volume_{t-1}}, \quad \partial GCI_t = ln\frac{GCI_t}{GCI_{t-1}}$$

where IDX_t is the stock index closing price at the end of week t, $Volume$ is the index trading volume, and GCI is the number of new-registered accounts. We calculate the return volatility VOL with a 5-day moving window as follows:

$$VOL_t = \frac{1}{5-1}\sum_{i=0}^{4}(R_{t,d-i} - \overline{R_t})^2, \quad \partial VOL_t = ln\frac{VOL_t}{VOL_{t-1}}$$

where day d is the last trading date of week t, and $\overline{R_t}$ is the mean of index returns from $R_{t,d-4}$ to $R_{t,d}$. To examine the determinants of GCI, we first conduct the predictive regression model of GCI_t on one to three weeks' lagged market returns, lagged trading volumes plus interactions with returns, and lagged volatility, controlled for stationarity with GCI_{t-1}, i.e.,

$$GCI_t = a_{GCI} + \sum_{i=1}^{3}\kappa_i GCI_{t-i} + b_{DUM_1}DUM_{1,t} + b_{DUM_2}DUM_{2,t} + \epsilon_{GCI,t} \tag{1a}$$

$$\begin{aligned}\epsilon_{GCI,t} = &\sum_{i=1}^{3} b_{R,i}R_{t-i} + \sum_{i=1}^{3} b_{Volume,i}\partial Volume_{t-i} + \sum_{i=1}^{3} b_{VOL,i}\partial VOL_{t-i} + \\ &b_{int}(R_{t-1} \times \partial Volume_{t-1}) + e_{GCI,t}\end{aligned} \tag{1b}$$

where $DUM_{1,t}$ is the dummy variable which is 1 if the week t is the first week of the month and 0 elsewhere, and $DUM_{2,t}$ is 1 if t is the second week of the month and 0 elsewhere. The two dummy variables control the seasonality in opening new accounts, whose significance may come from the seasonality in salary paychecks among middle-class workers in China. Based on our hypotheses, the coefficients for returns $b_{R,i}$, volume $b_{Volume,i}$ and volatility $b_{VOL,i}$ will be positive and significant. We also add the interaction effect between one-week lagged volume and returns to examine whether past return have larger positive effect when the trading volume is larger. Next, we examine the changes in new brokerage accounts ∂GCI can be predicted with lagged returns, volatilities and trading volume changes. We

first control ∂GCI for serial correlations and seasonality with the following regression:

$$\partial GCI_t = a_{\partial GCI} + \sum_{i=1}^{3} \eta_i \partial GCI_{t-i} + \beta_{DUM_1} DUM_{1,t} + \beta_{DUM_2} DUM_{2,t} + \epsilon_{\partial GCI,t} \tag{2}$$

According to the hypothesis, the lagged returns would be positively and significantly associated with returns, i.e., b_{1wk}, b_{2wk} or b_{3wk} in Equation 3 will be positive and significant. Similar to the step in Equation 1, we collect the residual terms $\epsilon_{\partial GCI,t}$ in Equation 2 and use it as a response variable for the following multiple regression:

$$\begin{aligned} \epsilon_{\partial GCI,t} = & \sum_{i=1}^{3} \beta_{R,i} R_{t-i} + \sum_{i=1}^{3} \beta_{Volume,i} \partial Volume_{t-i} + \sum_{i=1}^{3} \beta_{VOL,i} \partial VOL_{t-i} + \\ & \beta_{int}(R_{t-1} \times \partial Volume_{t-1}) + e_{\partial GCI,t} \end{aligned} \tag{3}$$

in which we examine whether the part of GCI unexplained by the its lags and seasonality can also be explained past volatilities and past trading volume, by looking at the coefficients $\gamma_{VOL,i}$ and $\gamma_{volume,j}$ for $i, j = 1, 2, 3$. According to our hypothesis, positive changes in trading volume and volatility predict higher GCI and thus positive ∂GCI, that γ_{VOL} and γ_{volume} in Equation 3 should be significant and positive.

To answer the next question that whether GCI can predict market reutrns, we perform the OLS regression of returns on lagged ∂GCI. The reason we choose ∂GCI as the explanatory variable is because of matching the moment: since the stock index is co-integrated with the GCI, their first derivatives, returns and ∂GCI, should have the same moment. The regression is as follows:

$$R_t = \beta_{t-1} \partial GCI_{t-1} + \beta_{t-2} \partial GCI_{t-2} + \beta_{t-3} \partial GCI_{t-3} + \epsilon_{R,t} \tag{4}$$

and if GCI has predictability over market returns, then at least one β should be significant. We may also be interested in return predictability among extreme observations in GCI, that whether the abnormal growth or drop in opening new accounts will forecast market returns.

To test this, we will run the following regression:

$$R_t = b_{\partial GCI} * \partial GCI_{t-1} + b_{DUM_\sigma} * DUM_{\sigma,t-1} + e_{R,t}, \tag{5a}$$

$$\text{where } DUM_{\sigma,t} = 1 \text{ if } t > 30 \text{ and } |\partial GCI_t| \geq \overline{\partial GCI_{t-30,t}} + S_{\partial GCI_{t-30,t}}, \quad 0 \text{ otherwise} \tag{5b}$$

where $\overline{\partial GCI_{t-30,t}}$ and $S_{\partial GCI_{t-30,t}}$ are the sample mean and standard deviation of ∂GCI during between weeks $t-30$ and t. We choose $t > 30$ to mitigate the noises of variance in a small sample size. According to our hypothesis, the coefficient b_{DUM} will be significant in the regression, and will probably be negative as extreme ∂GCI may suggest investors' over-reaction or under-reaction that which retrace to normality by adjusting future returns.

Furthermore, we examine the asymmetric linear relation between ∂GCI and future returns, i.e., whether positive changes in new accounts have different predictive power over returns from negative changes. We test the relation by conducting the following regression:

$$R_t = \sum_1^3 \beta^+_{t-i} \partial GCI_{t-i} + \sum_1^3 \beta^-_{t-i} \partial GCI_{t-i} + \sum_1^3 c_{t-1} ILLIQ_{t-i} \tag{6}$$

where $\partial GCI^+ = \partial GCI$ if $\partial GCI > 0$, zero otherwise; and $\partial GCI^- = \partial GCI$ if $\partial GCI < 0$, zero otherwise. $ILLIQ$ is the 5-day moving average Amihud's illiquidity measure introduced by Amihud (2002), i.e.:

$$ILLIQ_t = \frac{1}{5} \sum_{i=0}^{4} \frac{|R_{t,d-i}|}{Volume_{t,d-i}}, \quad \partial ILLIQ_t = ln \frac{ILLIQ_t}{ILLIQ_{t-1}}$$

We use $ILLIQ$ to control for the liquidity effect in ∂GCI, since the changes in GCI bring about changes in the population of investors that affects level of market liquidity. We would like to see whether the asymmetric predictive power in ∂GCI comes from investor sentiment.

Table 1: **Summary statistics of GCI and market variables**. GCI is the number of newly added brokerage accounts (in ten thousands) in the Shanghai Stock Exchange. $Index$ is the index level of the SSE Composite Index. $Volume$ is the trading volume (in 100 million RMB) of SSE index. ∂GCI is the logged change in GCI, R is the logged return of $Index$, $\partial Volume$ is the logged change of $Volume$, and VOL is the 5-day moving average return volatility of R. In May 2015, CSDC implemented new method that counts the number of new investors instead of new brokerage accounts, as one investor could open multiple accounts since April 2015. Period A refers to data collected with traditional method from Jan 11, 2008 to May 29, 2015. Period B is data collected using new method from May 8, 2015 to March 7, 2017.

Variable	Mean	Median	Std. Dev.	Min.	Max.	N
			Period A			
GCI	15.10	10.06	25.42	0.81	245.88	379
$Index$	2627.15	2448.59	607.34	1728.79	5484.68	379
$Volume$	6310.16	4614.75	6640.26	678.29	55373.59	379
∂GCI	0.0052	0.026	0.42	-1.96	2.20	378
R (%)	-0.046	0	3.61	-14.90	13.94	378
$\partial Volume$	0.0051	-0.028	0.43	-1.96	2.13	378
VOL	0.014	0.012	0.0091	0.0019	0.053	378
			Period B			
GCI	41.97	35.49	24.07	10.36	164.44	99
$Index$	3292.90	3154.32	481.69	2737.60	5166.35	99
$Volume$	15548.28	11399.20	11281.39	1080.24	55373.59	99
∂GCI	-0.013	0.0061	0.27	-0.72	1.05	98
R (%)	-0.25	0.084	4.05	-14.29	8.54	98
$\partial Volume$	-0.015	-0.018	0.76	-4.49	4.55	98
VOL	0.015	0.011	0.013	0.0021	0.0669	98

2.3. Analysis of GCI's Deteminants

Table 1 reports the summary statistics of the number of newly added stock accounts (GCI), SSE index level ($Index$), index trading volume ($Volume$), the logged change in new accounts (∂GCI), logged index return (R), logged change in trading volume ($\partial Volume$), and 5-day return volatility (VOL). Period A is dated from January 2008 to May 2015 and refers to GCI data collected with the traditional method mentioned in Section 2.1, while Period B refers to GCI data collected using the new method from May 2015 to April 2017. Period A has 379 observations on average and covers the market crash in 2008 and the market boom in 2015; Period B has 99 observations and covers the market crash in June 2015. The

Table 2: **Correlation matrix of GCI and market variables.** Period A refers to data collected with traditional method from Jan 2008 to May 2015. Period B refers to data collected with new method from May 2015 to March 2017.

Period A

	GCI	$Index$	$Volume$	VOL	∂GCI	R	$\partial Volume$	∂VOL
GCI	1							
$Index$	0.60	1						
$Volume$	0.89	0.58	1					
VOL	0.20	0.33	0.11	1				
∂GCI	0.13	0.04	0.20	0.03	1			
R	0.11	0.03	0.23	-0.16	0.21	1		
$\partial Volume$	0.08	0.02	0.17	0.03	0.83	0.37	1	
∂VOL	0.07	0.06	0.08	0.44	0.09	-0.07	0.11	1

Period B

	GCI	$Index$	$Volume$	VOL	∂GCI	R	$\partial Volume$	∂VOL
GCI	1							
$Index$	0.71	1						
$Volume$	0.73	0.92	1					
VOL	0.13	0.29	0.49	1				
∂GCI	0.21	-0.04	0.08	0.02	1			
R	0.09	0.10	-0.03	-0.48	0.12	1		
$\partial Volume$	0.36	0.28	0.31	0.04	0.29	0.24	1	
∂VOL	0.00	0.00	0.05	0.37	0.05	-0.29	-0.03	1

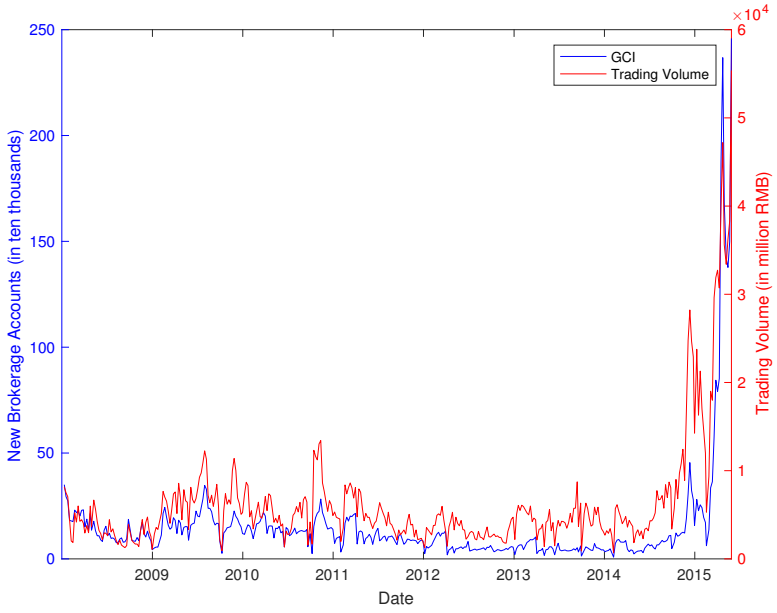
average market returns in both s are negative, as both s contain the observations from the period of market crashes, while the average return in Period A is slightly higher as it covers the bullish period in 2015. The market trading volume in Period B is significantly higher than that in Period A, with an average of 1.56 trillion RMB ($226 billion), which may come from the constant increase in the number of investors in the market as well as the increase in price level. Notice that the average $\partial Volume$ in Period B is negative with an value of -1.5%, possibly due to the decrease in trades during the 2015 market crash. The return volatility is approximately the same in the two periods, with an average of around 1.5%.

Table 2 reports the correlation between the variables and their first derivatives. From the table we can see that GCI and market trading volume are highly correlated in both

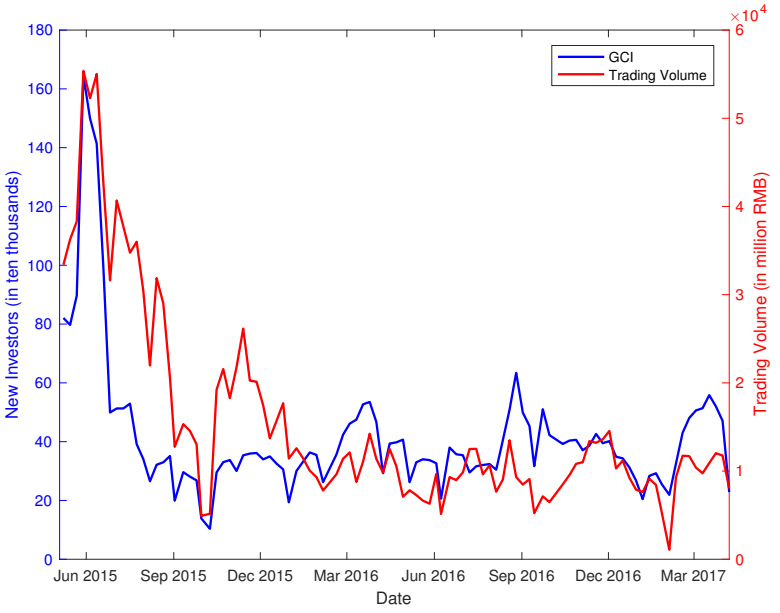
level and speed in Period A, with value of 0.89 and 0.83, respectively; and in Period B volume and GCI still have high correlation of 0.73. This result suggests that there is high synchronicity between newly registered brokerage accounts and trading volume, which is also shown in Figure 1. Such synchronicity is reasonable because new investors who have just opened the accounts are more likely to initiate trades and incur trading volume; furthermore, existing investors are more prone to trade when investor sentiment is high. While the co-movement between new accounts and volume weakens after the crash, as opening accounts is less associated with initiating trades during the crash period. When the market keeps falling, new investors tend not to participate in trades immediately after opening accounts. The market index and GCI are also moderately synchronized, possibly because investor sentiment is high when the index is high, vice versa. In particular, the co-movement between index and GCI is much strong during the period of market boom and burst, as we can see in Figure 2. The return volatility and ∂GCI are barely in either s, suggesting that the speed of opening new brokerage accounts does not affect contemporaneous return volatility of the short run.

Next, we examine the factors that potentially determines the opening of new accounts. The dependent variable we choose is e_{GCI}, the new brokerage accounts or investors after we control for serial correlation and seasonality effect. The independent variables are lagged 1-3 weeks' returns, trading volume changes and volatility changes. We use volume and volatility changes because trading volume and return volatility is non-stationary, which may produce biased result. Next, we use ∂GCI as dependent variable to examine the determinants of GCI change. Figure 3 of ∂GCI shows that the first derivative of new brokerage accounts is mean-reverting and matches the moment of market returns. Before regressing ∂GCI, we first control for its serial correlations and seasonality in the response variable.

Table 3 summarizes the regression result for serial correlations and seasonality of GCI and ∂GCI, with Period A referring to traditional data collecting method and Period B to new method mentioned in Section 2.1. GCI has significant serial correlation up to three

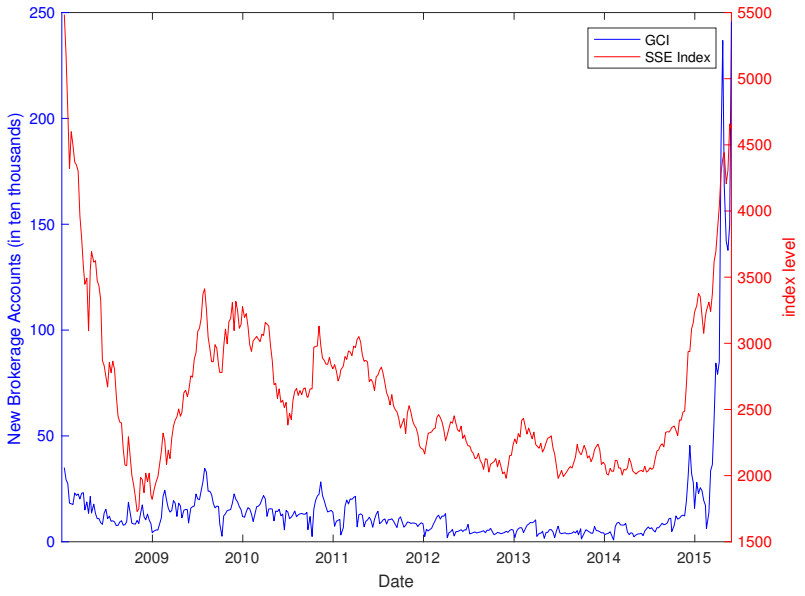


(a) *GCI* and market trading volume from Jan 2008 to May 2015

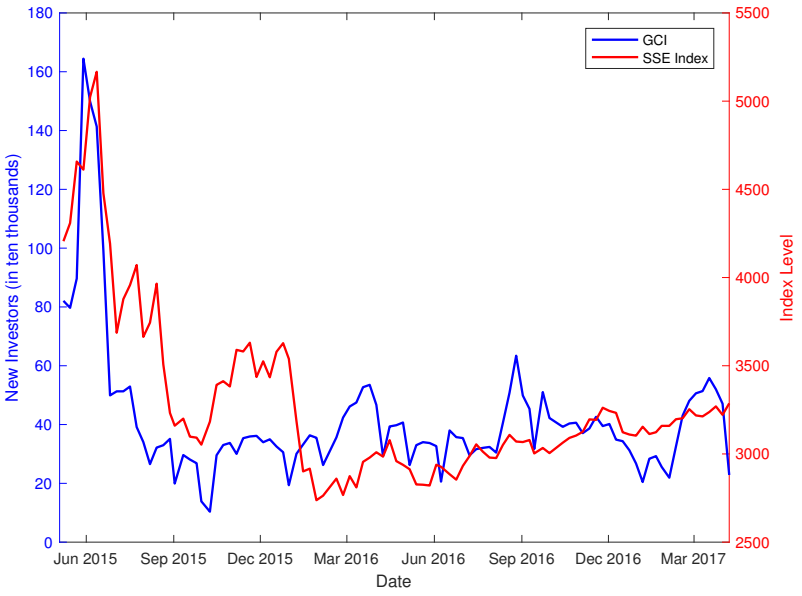


(b) *GCI* and market trading volume from May 2015 to April 2017

Figure 1. Time series plot of Garlic Chive Index (*GCI*) and its strong synchronicity with SSE index trading volume. *GCI* is the weekly number of new brokerage accounts in the Shanghai Stock Exchange, in ten thousands. Trading volume is the SSE Composite Index trading volume in million RMB. Plot 1(a) refers to *GCI* data from Jan 2008 to May 2015, collected using traditional method that surveys the number of newly open accounts. Plot 1(b) refers to data from May 2015 to April 2017 collected with new method that surveys the number of new investors who open trading accounts for the first time, in dealing with the new policy in May 2015 that one investor was enable to open multiple brokerage accounts.

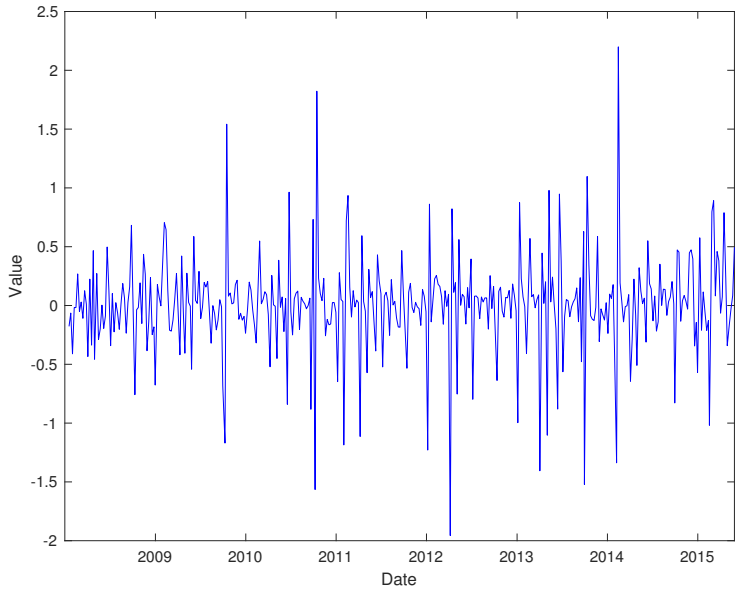


(a) *GCI* and SSE Index from Jan 2008 to May 2015

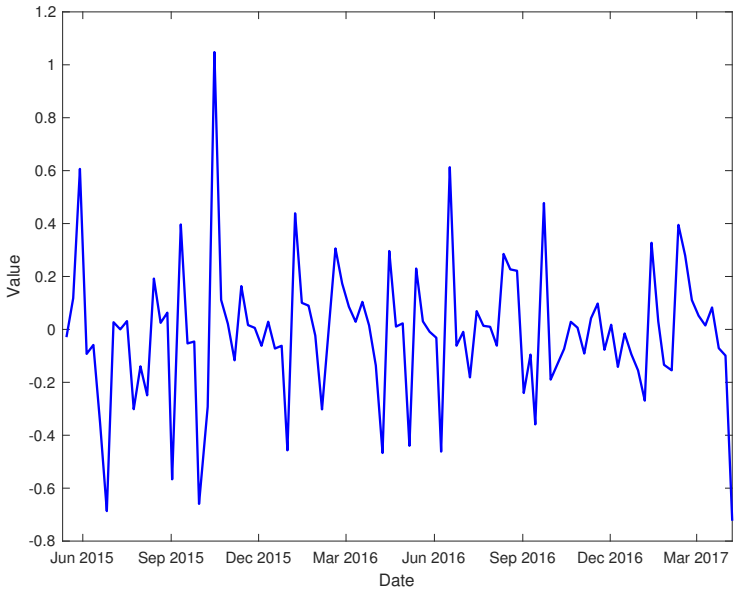


(b) *GCI* and SSE Index from May 2015 to April 2017

Figure 2. Time series plot of *GCI* and SSE Composite Index. Plot 2(a) refers to *GCI* data from Jan 2008 to May 2015, collected using traditional method that counts the number of newly open accounts. Plot 2(b) refers to data from May 2015 to April 2017 collected with new method that counts the number of new investors who open trading accounts for the first time, in dealing with the new policy in May 2015 that one investor was enable to open multiple brokerage accounts.



(a) ∂GCI from Jan 2008 to May 2015



(b) ∂GCI from May 2015 to April 2017

Figure 3. Time series plot of ∂GCI, the logged changes of GCI. ∂GCI in Plot 3(a) is calculated using GCI data from Jan 2008 to May 2015, collected using traditional method that counts the number of newly open accounts. ∂GCI in Plot 3(b) is calculated using GCI data from May 2015 to April 2017 collected with new method that counts the number of new investors who open trading accounts.

Table 3: **Auto-correlation and seasonality of GCI and ∂GCI.** We conduct a time series regression of ∂GCI on its lagged observations from one to 4 weeks. We also control for seasonality using $DUM_{1,t}$ and $DUM_{2,t}$, where $DUM_{1,t}$ is 1 if t is the first week of the month, zero otherwise; $DUM_{2,t}$ is 1 if t the second week of the month, zero otherwise. Period A refers to data collected with traditional method from Jan 2008 to May 2015. Period B refers to data collected with new method from May 2015 to March 2017.

	GCI_{t-1}	GCI_{t-2}	GCI_{t-3}	$DUM_{1,t}$	$DUM_{2,t}$
Period A	1.29***	-0.66***	0.46***	-2.20	-0.16
Adjusted R^2:			0.883		
Period B	1.00***	-0.11	-0.090	-5.08	-1.00
Adjusted R^2:			0.732		

	∂GCI_{t-1}	∂GCI_{t-2}	∂GCI_{t-3}	$DUM_{1,t}$	$DUM_{2,t}$
Period A	-0.39***	-0.19***	-0.14***	-0.20***	-0.14**
Adjusted R^2:			0.179		
Period B	-0.16	-0.12	-0.23	-0.091	-0.032
Adjusted R^2:			0.059		

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

lags in Period A, while in Period B only lagged-one-week lag shows up with coefficient of 1 which implies random walk. The autocorrelation and seasonality effect explains most of the variations in GCI for both periods, with R-squared of 88.3% and 73.2% respectively. From Period A in the table the coefficients for lagged ∂GCI are significantly negative to the three lags, suggesting that there is possible reverting pattern in the logged change of GCI. Yet the coefficients of lags in Period B are not statistically significant, suggesting that the time series is stationary. The dummy variables have negative coefficients, suggesting that controlled for autocorrelation, there are fewer accounts opened in the first and second weeks of the month. This may result from the seasonality in household's salary payment, that many middle-class workers and employees receive their paychecks at the end of the month, while most of the new investors are from the middle class in China.

We documents in Table 4 the predictive regression of GCI over lagged market returns, market trading volumes and lagged market return volatilities plus interaction between volume

and returns, controlled for stationarity with lagged GCI[1]. Lagged one-week returns and lagged one-week trading volume change show up as the most significant explanatory variable in Period, with coefficients of 3.02 and -0.45 in the full model, respectively, suggesting that investors initiate buy orders in response to market momentum, i.e., they open brokerage accounts to trade when the recent past returns are high; but when the volume increases abnormally, investors may enter at a lower speed, which contradicts our hypothesis. The result is in accordance with Brown and Cliff (2004) that investor sentiment is determined by past returns. Similar effects are observed in the subset models in Period A has negative explanatory power in Period. Note that this result is possibly due to the high correlation between one-week lagged GCI and lagged trading volume. The predictive effect of market returns and trading volumes appear much weaker in Period B, suggesting that opening brokerage accounts are less associated with initiating trades and investors' mood after the market crash. Return volatility has little predictive power indicated in the insignificant coefficients. This result implies that new investors are mostly prone to trade with momentum than the incentive to coordinate or bet on volatility.

Table 5 reports the regression result of Equation 3, in which we analyze the determinants of the opening of the stock account. The response variable is $\epsilon_{\partial GCI}$, the speed of opening new accounts after controlled for serial correlations and seasonality. Period A refers to GCI data collected with the traditional method, and Period B to data with the new method, described in Section 2.1. the one-week lagged return shows up as positive and the most significant, with a value of 6.38 in the full model in Period A and of 7.76 in Period B, suggesting that given lagged trading volume and return volatility, the brokerage account opening patterns are positively associated with near-term market returns. The result is aligned with our hypothesis that the proxy of investor sentiment goes up when the past returns increases. Also, new investors may enter the market faster to chase the momentum in past returns. But in Period B the returns come with less determining power, because

[1]We have also performed OLS regressions with the same variables, in which the full model explains 90.6% of variations in GCI in Period A, and 83.5% in Period B.

Table 4: **Causality test of GCI_t.** Response variable is ϵ_{GCI_t}, the weekly number of new registered brokerage accounts at week t controlled for serial correlations and seasonalities. We run the bi-square robust regression over lagged market returns R_{t-i}, index trading volume changes $\partial Volume_{t-i}$ and return volatility changes ∂VOL_{t-i}, controlled for serial correlation of GCI, i.e., the regression residuals of $GCI_t = \alpha + \sum_{i=1}^{n} GCI_{t-i} + DUM_{1,t-1} + DUM_{2,t-1} + \epsilon_{GCI,t}$, where DUM_i is the ith week of the month dummy. Period A refers to data collected with traditional method from Jan 2008 to May 2015. Period B refers to data collected with new method from May 2015 to March 2017.

	Period A			**Period B**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
R_{t-1}	3.43***	3.30***	3.02**	1.47	2.22	1.52
	(5.38)	(3.55)	(3.07)	(0.78)	(1.07)	(0.74)
R_{t-2}		0.76	0.77		-1.21	-1.78
		(0.79)	(0.76)		(-0.59)	(-0.88)
R_{t-3}		-0.31	-0.28		-0.10	3.02
		(-0.33)	(-0.28)		(-0.06)	(0.11)
$\partial Volume_{t-1}$	-0.45***	-0.42***	-0.45***	0.060	0.0011	-0.43
	(-8.25)	(-4.82)	(-4.63)	(0.63)	(0.01)	(-0.23)
$\partial Volume_{t-2}$		0.11	0.12		-0.18	-0.14
		(1.14)	(1.18)		(-1.89)	(-1.09)
$\partial Volume_{t-3}$		-0.010	-0.012		-0.060	-0.19*
		(-0.11)	(-0.13)		(-0.61)	(-2.02)
∂VOL_{t-1}	0.02	0.012	0.014	0.056	0.23	0.26
	(0.54)	(0.20)	(0.21)	(0.46)	(1.46)	(1.63)
∂VOL_{t-2}		-0.016	-0.017		0.25	0.25
		(-0.22)	(-0.22)		(1.34)	(1.37)
∂VOL_{t-3}		-0.0074	-0.0085		0.046	0.024
		(-0.11)	(-0.13)		(0.30)	(0.16)
$(R \times \partial Volume)_{t-1}$			2.02			4.42
			(0.11)			(1.82)
N	375	374	374	95	94	94

t statistics in parentheses

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5: **Causality test of ∂GCI.** Dependent variable is $\epsilon_{\partial GCI,t}$, the standardized residual terms of the change in new brokerage accounts ∂GCI after controlling its serial correlation and seasonality. We obtain $\epsilon_{\partial GCI}$ by running an OLS regression $\partial GCI_t = \alpha + \sum_{i=1}^{3} \partial GCI_{t-i} + DUM_{1,t-1} + DUM_{2,t-1} + \epsilon_{\partial GCI,t}$. Explanatory variables include lagged index returns R_{t-i} and return volatilities ∂VOL_{t-i}, and lagged trading volume changes $\partial Volume_{t-i}$. Period A refers to data collected with traditional method from Jan 2008 to May 2015. Period B refers to data collected with new method from May 2015 to March 2017.

	Period A			**Period B**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
R_{t-1}		5.34***	6.38***		6.62*	7.76*
		(3.77)	(4.09)		(2.60)	(2.54)
R_{t-2}		3.31*	4.09*		0.17	3.17
		(2.34)	(2.55)		(0.068)	(1.05)
R_{t-3}		1.07	1.28		-1.06	-0.058
		(0.76)	(0.63)		(-0.42)	(-0.02)
$\partial Volume_{t-1}$	0.078		-0.39*	0.035		-0.045
	(0.61)		(-2.56)	(0.26)		(-0.23)
$\partial Volume_{t-2}$	0.21		-0.062	-0.11		-0.16
	(1.49)		(-0.39)	(-0.79)		(-1.12)
$\partial Volume_{t-3}$	0.08		-0.040	-0.27*		-0.17
	(0.64)		(-0.27)	(-2.00)		(-1.20)
∂VOL_{t-1}			0.21*			0.28
			(1.99)			(1.19)
∂VOL_{t-2}			0.19			0.49
			(1.53)			(1.81)
∂VOL_{t-3}			0.015			0.26
			(0.15)			(1.15)
$(R \times \partial Volume)_{t-1}$			6.34*			-1.00
			(2.20)			(-0.28)
N	375	375	375	95	95	95
Adj. R^2	0	0.048	0.060	0.017	0.041	0.031

t statistics in parentheses

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

after the market crash the opening of brokerage accounts are less related to trade initiation. The lagged one-week trading volume change show up in Period A as negative, with a value of -0.39, suggesting that new investors enter at lower speed when trading volume increased abnormally in the past week. The interaction effect is positive and significant in Period A, of 6.34, which indicates that the lagged return effect is more positive when the lagged trading volume increase is higher. This is aligned with our hypothesis that the investors initiate more trades when market was excited most recently.

2.4. Analysis of GCI's Predictability on Market Returns

Table 6 reports the predictive regression of market return R_t on lagged ∂GCI and its extreme observation dummy $DUM_{\sigma,t-i}$. We can see in both s of the table that lagged ∂GCI barely predicts stock returns. None of the variables' coefficients are significant as well when we put in DUM_σ to test for extreme observations. These findings strongly reject our hypothesis, and yet it is reasonable because the market is efficient, and in equilibrium all the present information will be embedded in the current stock prices, thereby not affecting future returns. In Section 2.5 we will further examine whether positive and negative components of ∂GCI have different predictability over future market returns. The result, together with the causality test in the previous section, implies that lagged returns cause GCI to change while GCI change does not cause future returns.

2.5. Analysis of Asymmetry in GCI's Return Predictability

From Table 7, we observe that one-week lagged negative change in GCI has significant predictive power over stock returns in all four models, while positive change has no significant coefficients. The result stays the same after controlling liquidity variables $\partial ILLIQ$. We control for illiquidity to see whether the asymmetric predictability comes from the liquidity effect of increasing participants in the market. We also test for controlling return volatility in seeing whether asymmetry comes from the volatility effect, which is not shown in the

Table 6: **Predictive regression of index return R_t on ∂GCI_{t-i} and DUM_{t-i}.** Response variable is R_t, the market return. To test whether ∂GCI and its extreme observations have predictive power over stock return, we conduct the OLS regressions of SSE index returns R_t on lagged changes in new accounts ∂GCI_{t-i}, controlled for dummy variable DUM_σ, which is 1 if the observation of GCI is at least one standard deviation away from the 30-week average ∂GCI and 0 otherwise. Period A refers to data collected with traditional method from Jan 2008 to May 2015. Period B refers to data collected with new method from May 2015 to March 2017.

	Period A			**Period B**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
∂GCI_{t-1}	-0.0067	-0.0065	-0.0077	0.011	0.019	0.019
	(-1.53)	(-1.3)	(-1.55)	(0.68)	(1.09)	(1.06)
∂GCI_{t-2}		0.0029	0.0037		0.021	0.018
		(0.59)	(0.067)		(1.31)	(1.01)
∂GCI_{t-3}			0.0052			-0.0078
			(1.07)			(-0.47)
$DUM_{\sigma,t-1}$	0.0072	0.0073	0.0068	-0.010	-0.0036	-0.0033
	(1.55)	(1.42)	(1.33)	(-0.88)	(-0.30)	(-0.26)
$DUM_{\sigma,t-2}$		0.0015	-0.00015		-0.011	-0.013
		(0.29)	(-0.027)		(-0.84)	(-0.96)
$DUM_{\sigma,t-3}$			0.011*			0.0050
			(2.13)			(0.37)
N	377	376	375	97	96	95
Adj. R^2	0.004	0.004	0.016	-0.008	-0.007	-0.026

t statistic in parentheses
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

table due to its insignificance. The coefficient of negative GCI change is negative and significant, of -0.021, in the full model of Period A. We can also observe from the table that the coefficients of ∂GCI^- are negative and significant for the subset models. The outcome suggests that the linear relation between ∂GCI and future return is significantly asymmetric in Period A. When negative change in GCI was observed in the past week, holding all else fixed, the more negative the change is, the more positive the next week's return will be. It indicates that, surprisingly, the GCI has predictive power on future returns if we consider negative and positive change in GCI separately. The predictive power possibly comes from asymmetric behaviors among investors: when stocks goes up, retail investors trade more; when stocks goes down, retail investors trade less and wait for the prices to bounce back. The negative coefficient suggests that negative change in GCI is a reversal indicator of future stock returns, which partially agrees on our hypothesis[2]. While the predictability of negative change in GCI becomes very weak in Period B, suggesting that GCI is less meaningful in predicting returns after the market crash.

3. GCI in Predicting Factor Returns

The next question we would like to think about is whether the behavior of registering brokerage accounts can predict factor returns. Baker and Wurgler (2006) finds that the higher sentiment predicts negative returns on smaller stocks, younger stocks, less profitable stocks, lower dividend stocks, higher volatility stocks, and the lower sentiment predicts positive returns. This happens because in high sentiment investors tend to overprice the stocks that have the above characteristics, and their prices will adjust when the sentiment cools down. Based on the findings, we would like to see whether GCI can predict factor returns, i.e., the zero-net investment portfolio returns based on factor of size, book-to-market ratio (HML), age, dividend yield, return volatility and past cumulative returns.

[2]We also control for volatility and perform bi-square robust regressions with the same variables in Table 7. The results are very similar to the OLS regression results in both periods, and the significance of coefficients remain.

Table 7: **Linear asymmetry test of ∂GCI_{t-i} in return predictability**. Response variable is R_t, the market return. We run the OLS regressions of R_t on lagged values of $\partial GCI^+$and ∂GCI^-, where $\partial GCI^+ = \partial GCI$ if $\partial GCI > 0$, zero otherwise; and ∂GCI^- equals ∂GCI if $\partial GCI < 0$, zero otherwise. The regressions are controlled with $\partial ILLIQ$, the logged change in 5-day moving average Amihud's illiquidity measure: $ILLIQ_t = \frac{1}{5}\sum_{t=1}^{5} \frac{|R|_t}{Volume_t}$, where $|R|$ is the absolute value of daily SSE index return and $Volume$ is the daily index trading volume. Period A refers to data collected with traditional method from Jan 2008 to May 2015. Period B refers to data collected with new method from May 2015 to March 2017.

	Period A			**Period B**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
∂GCI^+_{t-1}	0.0077	0.014	0.014	-0.016	-0.0050	-0.012
	(1.03)	(1.36)	(1.36)	(-0.63)	(-0.17)	(-0.41)
∂GCI^+_{t-2}		0.0021	0.0028		0.010	0.0031
		(0.28)	(0.27)		(0.39)	(0.10)
∂GCI^+_{t-3}			0.012			-0.014
			(1.53)			(-0.52)
∂GCI^-_{t-1}	-0.020**	-0.020**	-0.021**	0.040	0.036	0.040
	(-2.87)	(-2.98)	(-3.05)	(1.44)	(1.25)	(1.34)
∂GCI^-_{t-2}		0.0086	0.0080		0.020	0.012
		(0.91)	(0.86)		(0.26)	(0.38)
∂GCI^-_{t-3}			0.0011			-0.017
			(0.11)			(-0.50)
$\partial ILLIQ_{t-1}$	0.002	0.0028	0.0038	-0.012	-0.0077	-0.013
	(0.71)	(0.73)	(0.95)	(-1.69)	(-0.98)	(-1.47)
$\partial ILLIQ_{t-2}$		-0.0014	-0.0015		0.0059	-0.00072
		(-0.37)	(-0.34)		(0.76)	(-0.08)
$\partial ILLIQ_{t-3}$			-0.0032			-0.011
			(-0.81)			(-1.27)
N	377	376	375	97	96	95
Adj. R^2	0.015	0.012	0.015	0.015	-0.004	-0.015

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.1. Construction of Factor Returns

In this section, we would like to examine whether ∂GCI have predictive power over subsequent factor returns. We construct the factor returns based on the method in Fama and French (2015). At the end of December of each year from 2006 to 2015, I divide all A-share firms listed in the Shanghai Stock Exchange into 5 quantile groups based on their market value, B/M ratios, age, dividend yield, return volatility (12-month), and past cumulative returns (12 month). The portfolios are kept unchanged for the following twelve months, from January to December next year. Returns for the 5 portfolios are calculated as the value-weighted average of individual stock returns. Then I construct the zero-investment portfolio returns for the factors as follows:

1. **SMB**: the return differences between smallest-size portfolio (1st quantile) and biggest-size portfolio (5th quantile). Size is defined as tradable market value, i.e., price times tradable shares outstanding.

2. **HML**: the return differences between highest-B/M portfolio (5th quantile) and lowest-B/M portfolio (1st quantile)

3. **YMO**: the return differences between youngest-age portfolio (1st quantile) and oldest-age portfolio (5th quantile)

4. **HDMLD**: the return differences between highest-dividend-yield portfolio (5th quantile) and lowest-dividend-yield portfolio (1st quantile)

5. **HVMLV**: the return differences between highest-volatility portfolio (5th quantile) and lowest-volatility portfolio (1st quantile)

6. **WML**: the return differences between highest-past-cumulative-return portfolio (5th quantile) and lowest-past-cumulative-return portfolio (1st quantile)

Next, I perform regressions of the factor returns with ∂GCI_{t-1}, the lagged change of the new registered brokerage accounts in the Shanghai stock exchange (GCI). I further analyze the

asymmetries in ∂GCI's predictability by splitting the variable into its positive component ∂GCI^+ and negative component ∂GCI^-. The regressions are as follows:

$$FACTOR_t = \beta^{\partial GCI} \partial GCI_{t-1} + e_t \tag{7}$$

$$FACTOR_t = \beta_+^{\partial GCI} \partial GCI_{t-1}^+ + \beta_-^{\partial GCI} \partial GCI_{t-1}^- + \xi_{Volume} \partial Volume_{t-1} + \xi_{VOL} \partial VOL_{t-1} + u_t \tag{8}$$

where $FACTOR_t$ is the zero-investment portfolio return for a given factor at month t. We also control for lagged volatility, lagged market trading volume to see whether GCI's predictability comes from volatility bet, sentiment, or other than both. Monthly data of returns, dividend yields, B/M ratios, ages, and market values for Shanghai A-share companies are collected from the Wind Financial Terminal.

3.2. Analysis of GCI's Predictability on Factor Returns

Table 8 reports the summary statistics and correlation matrix of monthly factor returns. From the table the average monthly return in SMB is significantly positive, with an annualized value of 25.08%. This suggests that the small stocks outperform big stocks substantially during the sample period. In the correlation matrix, SMB is highly negatively correlated with YMO and $HDMLD$, with values of -0.88 and -0.89 respectively, implying that young firms listed in the Shanghai Stock Exchange are usually big firms, and small firms usually pay lower dividends. HML and $HDMLD$ has a positive correlation of 0.63, suggesting that value firms usually pay higher dividends. Figure 4 shows the accumulative returns of the six factors from January 2008 to May 2015. The SMB factor has the highest accumulative payoffs during the period, with a return of nearly 500%; trading with momentum earns the lowest returns, with approximately -80% in WML during the period.

Table 9 summarizes the predictive regression results of factor returns on one-month lagged ∂GCI, its positive components ∂GCI^+, and negative components ∂GCI^- [3]. From the

[3]We also perform bi-square robust regressions for dependent variables SMB, HML, YMO and $HDMLD$.

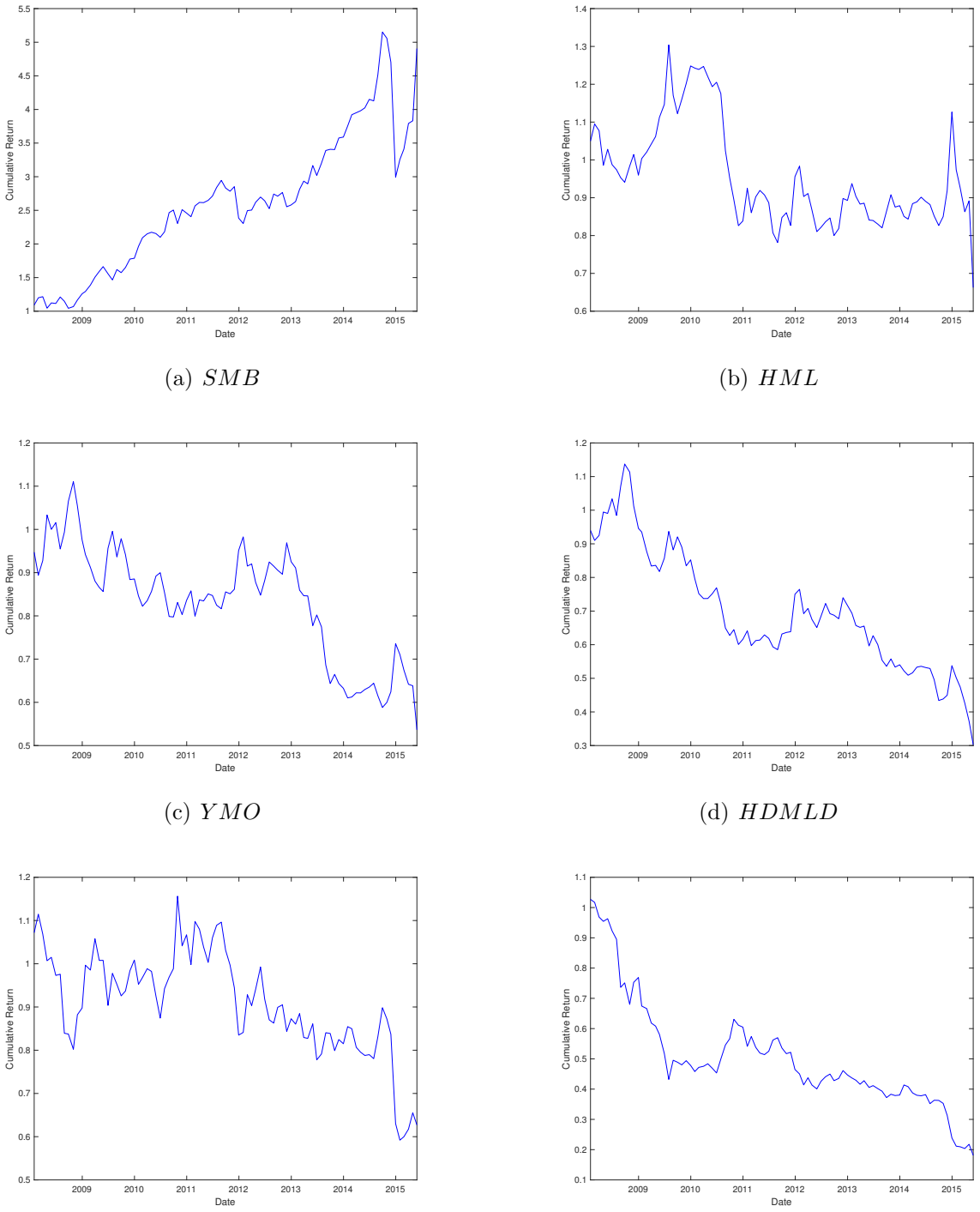
Table 8: **Summary statistics (in percentage point) and correlation matrix of monthly factor returns.** Based on Fama and French (2015), at the end of December of each year from 2007 to 2014, all A-share firms listed on the Shanghai stock exchanges are divided into 5 equally populated groups on the basis of their market value, B/M ratios, age, dividend yield, return volatility (12-month), and past cumulative returns (12 month). The portfolios are kept unchanged for the following twelve months, from January to December next year. Returns for the 5 portfolios are calculated as the equal-weighted average of individual stock returns. SMB is the return difference between smallest-size portfolio and biggest-size portfolio. HML is the return difference between highest-B/M portfolio and lowest-B/M portfolio. YMO is the return difference between youngest-age portfolio and oldest-age portfolio. $HDMLD$ the return difference between highest-dividend-yield portfolio and lowest-dividend-yield portfolio. $HVMLV$ is the return difference between highest-volatility portfolio and lowest-volatility portfolio. WML is the return difference between the highest-past-cumulative-return portfolio and the lowest-past-cumulative-return portfolio.

Variable (%)	**Mean**	**Median**	**Std. Dev.**	**Min.**	**Max.**	**N**
SMB	2.09	2.40	7.54	-36.24	28.27	89
HML	-.26	-0.28	6.36	-25.69	22.58	89
YMO	-.57	-1.01	5.15	-15.95	17.83	89
$HDMLD$	-1.15	-1.41	6.04	-18.83	19.56	89
$HVMLV$	-.33	-0.22	6.13	-24.70	16.98	89
WML	-1.54	-1.64	6.46	-26.23	13.61	89

Corr.	SMB	HML	YMO	$HDMLD$	$HVMLV$	WML
SMB	1					
HML	-0.62	1				
YMO	-0.88	0.56	1			
$HDMLD$	-0.89	0.63	0.88	1		
$HVMLV$	0.48	-0.27	-0.57	-0.58	1	
WML	0.28	-0.28	-0.24	-0.30	0.43	1

table, ∂GCI has little predictive power. Yet to some response variables the positive and negative components of ∂GCI have significant and opposite coefficients in the regressions, indicating that ∂GCI has nonlinear predictability over future factor returns. We also report the control for lagged 30-day market return volatility and lagged monthly market trading volume, marked as VOL and $Volume$ in the explanatory variables. With the control, the coefficients for ∂GCI^+ has less significance, while significance for ∂GCI^- barely changes. In particular, ∂GCI^-'s coefficients are negative for SMB, positive for YMO and $HDMLD$.

The significance further increases for coefficients of negative GCI changes.



(a) SMB (b) HML

(c) YMO (d) $HDMLD$

(e) $HVMLV$ (f) WML

Figure 4. Time series plot of cumulative factor returns from Jan 2008 to May 2015. We divide all tradable A-share companies in the Shanghai Stock Exchange into five quantile with criteria of size (SMB), B/M ratio (HML), age (YMO), dividend yield ($HDMLD$), return volatility ($HVMLV$), and past cumulative returns (WML). The factor returns are the difference between the highest (lowest) and the lowest (highest) quantiles. The cumulative returns are the cumulative products of one plus factor returns.

Table 9: **Predictive regression of subsequent factor return on ∂GCI, ∂GCI^+ and ∂GCI^-.** Response variable is monthly return of the factor on bold. We examine whether zero-net investment factor returns including SMB, HML, YMO, $HDMLD$, $HVMLV$, and WML at month t can be predicted using lagged one-month ∂GCI, together with lagged one-month ∂GCI^+ and ∂GCI^-, where $\partial GCI^+ = \partial GCI$ if $\partial GCI > 0$, zero otherwise; and ∂GCI^- equals ∂GCI if $\partial GCI < 0$, zero otherwise. We control the regressions of SMB and HML on ∂GCI^+ and ∂GCI^- with 30-day SSE index return volatility ∂VOL as control for volatility effect and monthly trading volume $\partial Volume$ for liquidity effect in the test of asymmetric predictive power. The data for regression is monthly from Jan 2008 to May 2015.

Variable	**SMB**	**SMB**	**SMB**	**HML**	**HML**	**HML**	**YMO**	**YMO**	**HDMLD**	**HVMLV**	**WML**
∂GCI_{t-1}	-0.18			-0.29							
	(-0.11)			(-0.21)							
∂GCI^+_{t-1}		5.27*	6.07		-4.47*	-5.05	-3.87*	-3.86	-5.79*	2.29	-1.27
		(1.49)	(1.96)		(-2.01)	(-1.93)	(-2.18)	(-1.81)	(-2.39)	(1.07)	(-0.54)
∂GCI^-_{t-1}		-6.84*	-6.66*		4.82	4.79	5.17*	5.25*	6.84**	-5.45	-0.76
		(-2.28)	(-2.14)		(1.88)	(1.82)	(2.54)	(2.46)	(2.76)	(-1.07)	(-0.28)
$\partial Volume_{t-1}$			-0.97			0.60		-0.10	0.21		
			(-0.36)			(0.26)		(-0.06)	(0.10)		
∂VOL_{t-1}			4.83			-5.55*		-1.41	-1.86		
			(0.47)			(-2.01)		(-0.63)	(-0.71)		
N	87	87	87	87	87	87	87	87	87	87	87
Adj. R^2	-0.011	0.052	0.056	-0.011	0.039	0.063	0.067	0.049	0.0824	0.033	-0.017

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4. GCI vs. iVX: Comparison

In the introduction we demonstrate that the constituents of new registered investors are mostly retail investors, and thus GCI generally represents the sentiment among retail investors. One interesting thing we would like to investigate into is the disparity of moods between retail investors and institutions. Do retail investors have similar sentiment from institutional investors? And do institutional investor sentiment predict the stock returns better? In answering these questions, we first need to find a proxy of institutional investor sentiment. In this paper, the proxy we use is iVX, the implied volatility index in China that is comparable to the VIX index in the US. The iVX is constructed from SSE 50 ETF option implied volatiles with a similar method to the CBOE's VIX index, and it is viewed as a benchmark to predict future market return volatility. Different from GCI which is considered as a bullish sentiment, iVX is reported as the "fear index" in China, a bearish indicator of investor sentiment. Another difference is that the underlying ETF options in iVX have relatively high trading barriers, with which investors have to pass the exam for securities trading and deposit at least 500,000 RMB ($71,400) to initiate trades, while opening brokerage accounts have nearly no barriers. Moreover, the traders in SSE 50 ETF options are mostly institutional investors performing insurance and arbitrage trades, according to the Shanghai Futures Exchange. Therefore, it is reasonable to consider iVX as a proxy of institutional investor sentiment in our analysis. The iVX data are collected from the Shanghai Stock Exchange (`http://www.sse.com.cn/assortment/options/volatility/`), from May 2015 to April 2017.

4.1. Analysis of iVX's Predictability on Market Returns

Figure 5 shows the time series plot of iVX together with GCI in the sampling period. In the plot the implied volatility index has a downward-sloping trend, which is similar to the price pattern in VXX ETF as the product keeps losing money by purchasing expensive longer-

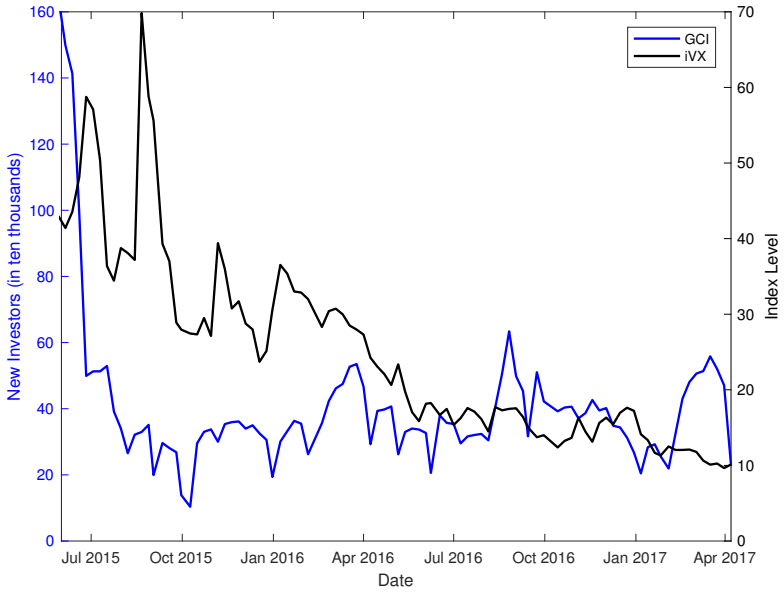


Figure 5. Time series plot of GCI and iVX from May 2015 to April 2017. iVX is the implied volatility index released by the Shanghai Stock Exchange. The implied volatility is derived from SSE 50 ETF options traded in China's stock market. The index calculation method is similar to the method from CBOE of constructing VIX index.

term contracts and selling cheap short-term contracts. There is barely any synchronicity between the opening pattern of new brokerage accounts and the implied volatility index.

Table 10 reports the summary statistics and correlation matrix of iVX. The correlation between GCI and iVX is 0.24, which is relatively low. It suggests that retail investor sentiment has little association with institutional investor sentiment. The correlation between implied volatility and market return volatility is substantially high, with a value of 0.83. This suggests that one of the most influential factors of implied volatility is the historical volatility, and this contemporaneous relation is observed in the US as well. iVX is also positively correlated with market trading volume, possibly because disagreement in price causes both trading volume and implied volatility to increase.

Table 11 summarizes the predictive regression result of SSE index return R on weekly lagged implied volatility change ∂iVX, defined as the logged change of iVX. We perform both OLS regression and bi-square robust regression to check the robustness of the results. For the model of OLS regression, lagged one-week implied volatility changes have significant

Table 10: **Summary statistics and correlation table of iVX.** iVX is the China's implied volatility index (weekly) released by the Shanghai Stock Exchange. The implied volatility is derived from SSE 50 ETF options traded in China's stock market. Note that the actual calculation method of iVX has not been fully released, and yet it is similar to the CBOE's method for VIX index. R refers to the logged change of the SSE Composite Index. $Volume$ is the trading volume (in 100 million RMB) of SSE index. GCI is the weekly number of new brokerage accounts, and VOL is the 5-day moving average return volatility of R. The date of observations is ranged from June 2015 to April 2017.

Variable	**Mean**	**Median**	**Std. Dev.**	**Min.**	**Max.**	**N**
iVX	25.10	21.36	12.90	9.65	69.83	96

Corr.	iVX	GCI	VOL	R	$Volume$
iVX	1				
GCI	0.24	1			
VOL	0.83	0.14	1		
R	-0.38	0.09	-0.48	1	
$Volume$	0.68	0.74	0.50	0.92	1

and negative predictive power, suggesting that when institutional investor feels more bearish about the market, market return will be lower next week. This result, however, is not robust, as we can see in the results of bi-square regression that none of the coefficients are significant. The significant results in OLS regression may be due to the outlier effect during the market crash, in which return volatilities largely increased while market returns were negative. Nonetheless, the positive change in iVX, or ∂iVX^{+}, is significant and negative in both OLS and robust regressions, with coefficients of -0.15 and -0.092 in the full model respectively. The result suggests that a larger increase in implied volatility predicts a lower subsequent market returns. This result implies that the return predictability in iVX is also asymmetric, similar to GCI.

The proxy of institutional investor sentiment iVX has as little return predictability as retail investor sentiment, implying that institutions may not predict better than retail investors over the market returns. Such result is aligned with the efficient market hypothesis. However, iVX have relatively better non-linear predictability of returns.

Table 11: **Predictive regression of market return R_t on ∂iVX_{t-i} and ∂GCI_{t-i}.** Response variable is R_t, the market return. In particular, ∂iVX_t is calculated as the logged change of iVX_t from iVX_{t-1}. We conduct both OLS regressions and bi-square robust regressions on the explanatory variables to document the robustness of the predictive power in the implied volatility. We also control liquidity and volatility effect with variables $\partial ILLIQ$ and ∂VOL in both regressions.

	OLS			**Robust**		
Variable	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
∂iVX_{t-1}	-0.077*	-0.069*		0.011	-0.0050	
	(-2.51)	(-2.03)		(0.536)	(-0.19)	
∂iVX^{+}_{t-1}			-0.15**			-0.092*
			(-2.87)			(-2.13)
∂iVX^{-}_{t-1}			0.04			0.052
			(0.58)			(1.01)
∂GCI_{t-1}		-0.0029			-0.025*	
		(-0.18)			(-2.00)	
∂GCI^{+}_{t-1}			-0.040			-0.031
			(-1.55)			(-1.44)
∂GCI^{-}_{t-1}			0.038			-0.016
			(1.45)			(-0.71)
$\partial ILLIQ^{-}_{t-1}$		-0.0074	-0.0020		-0.024**	-0.021*
		(-0.65)	(-0.18)		(-2.66)	(-2.15)
∂VOL^{-}_{t-1}		0.0021	0.0010		0.0092	0.0092
		(0.20)	(0.10)		(1.10)	(1.03)
N	93	93	92	93	93	92
Adj. R^2	0.054	0.029	0.082			

t statistic in parentheses

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

5. Conclusion

To summarize, the weekly number of newly registered accounts or investors GCI is significantly determined by past market returns, especially when past market trading volumes were high. Yet, the determining power in returns is weaken after the market crash, possibly because the account registering activities are less associated with initiating trades. GCI as well as the logged change in GCI, or ∂GCI, has little predictability over future market returns. However, the negative components of ∂GCI, or ∂GCI^- has significant and robust before the market crash. After the market crash, though, the predictive power disappears. ∂GCI^- also has predictability over factor returns, that a decrease in ∂GCI^- predicts higher subsequent returns in smaller stocks and lower returns in older stocks and lower dividend stock. Finally, we compare GCI with China's implied volatility index iVX, and find that iVX has insignificant predictability over market returns, similar to GCI, while iVX have relatively better non-linear predictive power that positive changes in iVX negatively forecast market returns.

This paper provides several implications in the use of GCI as a proxy of investor sentiment. First, after the 2015 market crash, new registered investors have weaker connections with trade initiations, particularly buy orders, rendering less information about investor sentiment and trading behaviors. Second, the useful component in GCI is its negative speed of changes, or ∂GCI^-, that positive changes may come from natural population growth rather than sentiment change thus has less valuable information about investors' mood. Moreover, the insignificant predictability in iVX and GCI imply that institutional investors are no better than the aggregated retail investors in predicting over the market.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of financial markets*, 5(1):31–56.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680.

Baker, M., Wurgler, J., and Yuan, Y. (2012). Global, local, and contagious investor sentiment. *Journal of Financial Economics*, 104(2):272–287.

Bikhchandani, S. and Sharma, S. (2000). Herd behavior in financial markets. *IMF Staff papers*, pages 279–310.

Brown, G. W. and Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 11(1):1–27.

Brown, G. W. and Cliff, M. T. (2005). Investor sentiment and asset valuation. *The Journal of Business*, 78(2):405–440.

Carpenter, J. N., Lu, F., and Whitelaw, R. F. (2015). The real value of china's stock market. Technical report, National Bureau of Economic Research.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fisher, K. L. and Statman, M. (2000). Investor sentiment and stock returns. *Financial Analysts Journal*, pages 16–23.

Han, X. and Li, Y. (2017). Can investor sentiment be a momentum time-series predictor? evidence from china. *Journal of Empirical Finance*.

Huang, D., Jiang, F., Tu, J., and Zhou, G. (2014). Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies*, page hhu080.

Kling, G. and Gao, L. (2008). Chinese institutional investors' sentiment. *Journal of International Financial Markets, Institutions and Money*, 18(4):374–387.

Li, X. and Zhang, B. (2008). Stock market behavior and investor sentiment: Evidence from china. *Frontiers of Business Research in China*, 2(2):277–282.

Shleifer, A. and Summers, L. H. (1990). The noise trader approach to finance. *The Journal of Economic Perspectives*, 4(2):19–33.

Zhu, B. and Niu, F. (2016). Investor sentiment, accounting information and stock price: Evidence from china. *Pacific-Basin Finance Journal*, 38:125–134.