PHBS

北京大学汇丰商学院

Big Data

Final Project Report, Group 2

Name	Student ID
沈廷威	1901212631
何隽贤	1901212583
王子雄	1901212644
玉鑫霖	1901212664
李佳辉	1901212603
李方闻	1901212601

Final Project Report

1. Overview

This report is for description of Group 2's final project for Big Data course. Our group developed two strategy to try more within the limited time period. Thus, our group was divided into three sub-groups, i.e. data processing group, strategy I group and strategy II group.

Data processing group has members of Fangwen Li and Junxian Xiu, aimed at cutting the data using different methods in the following effort: 1) Different time scales; 2) Different slicing standards; 3) Non-quantity & price information; 4) Translate transaction data into order data.

Strategy I group has members of Zixiong Wang and Jiahui Li, aimed at factor investment model based on minute -frequency data, using the 178 factor in 国泰君安 191 factors of each stock close price as an index to do stock selection, comparing four-times daily trading and one-time daily trading and concluding that the one-time daily trading performs betters than four-times trading strategy.

Strategy II group has members of Tingwei Shen and Xinlin Yu, aimed at factor investment model based on high frequency data built by low frequency data, conducting map reduce two times to rebuild order data, and during the process we build factors of daily frequency, comparing the relative strength of big orders and big sales, judging the trading trend of the next day with the daily frequency (For stock selection) and concluding that the constructed single factor showed good layered effect.

2. Strategy I

2.1 Strategy Introduction

Background: Traditional factor investment is based on the low frequency trading like daily or monthly trading. Some factors are tested for the trading with higher frequency like minute-level data. Previous empirical study shows that some of them still perform well. Instead of digging special price-volume characters, we focus on using the traditional factors in the minute-level data.

Factor I: skewness factor: In the Amaya et al. (2011) the author research on the skewness and kurtosis effect on daily data. The study shows that kurtosis has positive effect while skewness has negative effect significantly. This phenomenon can be explained by the "gaming theory". We used to hypothesize the distribution of stock price can be explained by the mean and variance. Default skewness is zero. In the reality the distribution used to be skewed. When the skewness is positive. There is a fat-tail in the distribution which means that the possibility of getting extreme high return is bigger. Thus investors will look highly on the stock price and the price premium will be overstated. Skewness is calculated by the following formula:

$$RSkew_i = \frac{\sqrt{N}\sum_{j=1}^{N} r_{ij}^{3}}{RVar_i^{\frac{3}{2}}} \quad j = \{1,2,3,\ldots\ldots N\}$$

Factor II: 178th factor in 国泰君安 191: In the empirical study of the famous 191 factors, the 178 factor shows relatively good performance in daily trade. The formula is:

$$\frac{close - delay(close, 1)}{delay(close, 1)} * volume$$

The principle behind it is simple. High volume deals and price improvement can show the raising expectation of one stock, which means the initial price is underestimated.

Strategy construction: For factor I, we need to calculate each stock skewness for each minute. Each RSkew in the formula is calculated by the previous N days. Since the data is minute-level, we need to use N*60*4 data to calculate one skew. The scale of factor matrix is determined by trade frequency. Considering the calculation problem, we choose to do the daily trading. Since research said that skewness factor has negative relationship with stock expected return, we choose the last ten stocks as the target stocks to deal in the next period. For daily trade we choose the close price as reference and deal in the next day opening time. For factor II, each 178 factor only use the deal time volume, deal time price and deal time price in the last trade date. To simplify the calculation, we also choose top ten score stocks in each signal. We choose to trade four times a day and one time each day as comparison. To make sure we can take the deal, we choose 10:00 am, 11:00 am, 13:30 pm, 14:30 pm for each trading day.

2.2 Data Slicing

To get the data needed by strategy I, we slice the data into daily and minute frequency data. In this process we only need one MapReduce. We use stock id and time stamp as the key, and the values are open, high, low, close, trading volume and amount of that time interval.

After we slice the data, we reshape the data into csv files by property for later analysis. Because each record of data has 6 properties so we reshape it into 6 csv files. In each of those files, each row represents a certain time stamp and each column represents a certain stock. Therefore, each element represents the value of a certain property of a particular stock at a certain time point.

2.3 Result

When testing on the 178th factor, we print out the accumulated PNL picture for both four times a day and one time a day. See Figure 1. Four-time Trade Each Day Accumulated PNL and Figure 2. Daily Trade Accumulated PNL.

It shows that four-time trade doesn’t bring higher return. In our view, short-term data volatility is quite high. Relying more on the short-term data leads to more uncertainty, the return will fluctuate like figure 1 shows. We will keep researching on the improvement of performance.

It is pretty much the same as for daily trade of 178 factor. During the skewness strategy, we faced a problem of high dimension calculation. Since we need to calculate the skewness for each moment. Suppose we use 20 days to do the trackback as on the research paper, we will calculate at least 20*60*4 for each day skewness for three months. To decrease the calculation, we choose the daily trading. See Figure 3. Skewness Factor Accumulated PNL.

Conclusion: With the 178 factors, we find that high-frequency trading may perform worse than daily trading. Sometimes we need to compromise the accuracy of signal to get the efficiency.

3. Strategy II

3.1 Strategy Introduction

For the second part, our strategy is to construct factor that can instruct trading daily from high frequency data. We get the idea from the research report from Haitong Security.

We first construct the traded buy order and sell order from the tickData of transaction, which is aggregate the trade volume and the weighted price from the transaction to the same order. And calculate the mean and variance for one stock per day. We then define the order with volume above one σ to be the big orders. Hence the factor definition we get is：

$$\frac{Amount\ of\ big\ buy\ orders}{Amount\ of\ all\ orders} - \frac{Amount\ of\ big\ sell\ orders}{Amount\ of\ all\ orders}$$

The logic behind this is that if the big buy order is more than the sell orders, we then think the buyers are in a stronger position. The calculation detail is discussed in the data slicing part.

3.2 Data Slicing

Because the data of strategy II is not as usual and it is will be complicated to deal with the raw data in one step, to realize it, we design to MapReduce the data twice.

In the first MapReduce, we need to translate the transaction data into order data. In the raw transaction data, we have the buying side order ID and selling side order ID so we can summary the transaction data according to their order ID.

Here is an example (see Figure 4. The WeChat Pay logo), in the map stage，each transaction record is split into two order record, buying side and selling side. The key is stock id, date and order ID. In the reduce stage, the data with the same order ID will be summarized and we can get its order size, amount and if it is buying or selling side.

In the second MapReduce, we deal with the results of the first MapReduce. In the map stage, we set the key to be the stock code and date. Then in the reduce stage, we will get all the orders of the stock in a single day and do the statistics.

In the reduce stage we use three loops and in the first loop we calculate the average order size. In the second loop we calculate the standard deviation of order size. And in the last loop we use the one sigma principle to judge if each of the order is big or not.

Finally, after we finish the whole process, we can calculate the big buying and selling order transaction amount ratio to all transaction amount of a day.

3.3 Result

After construction, we firstly did a rough test of this factor. The strategy is just holding the 10 stock with the highest factor value of the previous day and find out that it can generate cumulative return about 300% even after getting rid of the stock raising to the limit which is pretty ridiculous. See Figure 5. Rough Test of Strategy II.

After discussion with the teammates, we find that this may be unreasonable for we use $\frac{Close_t - Close_{t-1}}{Close_{t-1}}$ to measure the return, but it is unfair since we can only calculate the factor after close and most of our return is actually coming from the spread of $Open - Close_{t-1}$. So if we consider a more realistic scenario, a more reasonable measure of return would be $\frac{Close_t - Open_t}{Open_t}$ and also remove the stocks that suspended in day t.

After the change the measure of return, things went pretty bad as we can only get the best result from pure short the 10 stocks factors with the lowest factor value. It only generates 17% cumulative return in three months. See Figure 6. Second Test for Strategy II.

But considering the whole market was in a bow rising trend, we can get positive return from shorting strategy is actually proving our strategy to be useful. And the IC mean is still positive, and pure long cumulative return of 10 group also shows a layering power of the factor, just not too obvious. See Figure 7. Daily IC Histogram and Figure 8. Cumulative Income of Pure-Long Strategy (Sorted by Factor Value from High to Low).

Nevertheless, our return is still in weak position comparing to the index like 沪深 300, and is actually not tradable because of the T+1 restriction and limitation on shorting stocks. So, we look back to our original strategy to make it more realistic for trade.

So, the improved one is that we do the data slicing part to the beginning of the final minute rather than the EOD, so we can use the last minute to trade the stocks. In this way, using $\frac{Close_t - Close_{t-1}}{Close_{t-1}}$ to measure the return is actually reasonable. And we also only select stocks with market cap above 50 billion to get higher probability of that we can actually find seller within the last minute.

Due to time limit, we only implement this strategy within 2 months' data (Bugs on the Hadoop server).

Finally, we get a pretty decent result, long the first 10 stocks with the highest factor value. And found that it generate about 35% cumulative return within 2 months, while 沪深 300 just generate the 18% within the same period. Consider the prediction power, the rank IC mean is about 0.1, which shows crazily strong prediction power. See Figure 9. Daily IC Histogram (excluding last minute).

As for grouping, it still shows great layering power. See Figure 10. Cumulative Income of Pure Long Strategy (Sorted by Factor Value from High to Low) and Figure 11. Cumulative Returns Grouped by Factors in Two Months

So hurray!!!!! The power of this single factor is astonishing.

4. Future Work

Due to limited time, we have several ideas needed to be explored and implemented.

1) Volume based data. Our data is time series, i.e. price trends over time change. One possible improvement is volume series, i.e. prices trends over volume. We want to find how prices change when volumes change. First, for each order, we want to calculate the corresponding accumulated volume, that is summing up all the historical volume flow. For each key-value pair, we get a new feature called accumulated volume, we need to replace the time stamp by the accumulated volume stamp in the key. Finally, using the volume stamp to repeat our strategy to see there is an improvement.

2) Optimization on data slicing. When calculating the variance of the dataset, our implementation is using one single reducer to calculate the variance. This could be a waste of time and memory. First, we want to implement the map-reduce process to calculate the overall mean value of whole dataset and store it in a new file. Second, we want to implement another map-reduce process to calculate, specifically, loading the mean value from the file into memory, then using mappers to calculate the term $(x_i - \bar{x})^2$ and then using the reducers to calculate the overall variance. Third, to improve efficiency, we want to add some combiners between mappers and reducers

3) Factor construction. We used only one factor to construct our strategy. However, there are possible factors based on high-frequency data.Due to time limit, we can only analyze the performance of one single factor. Our future work is to construct more factors and test their validity.

Appendix

Figure 1. Four-time Trade Each Day Accumulated PNL

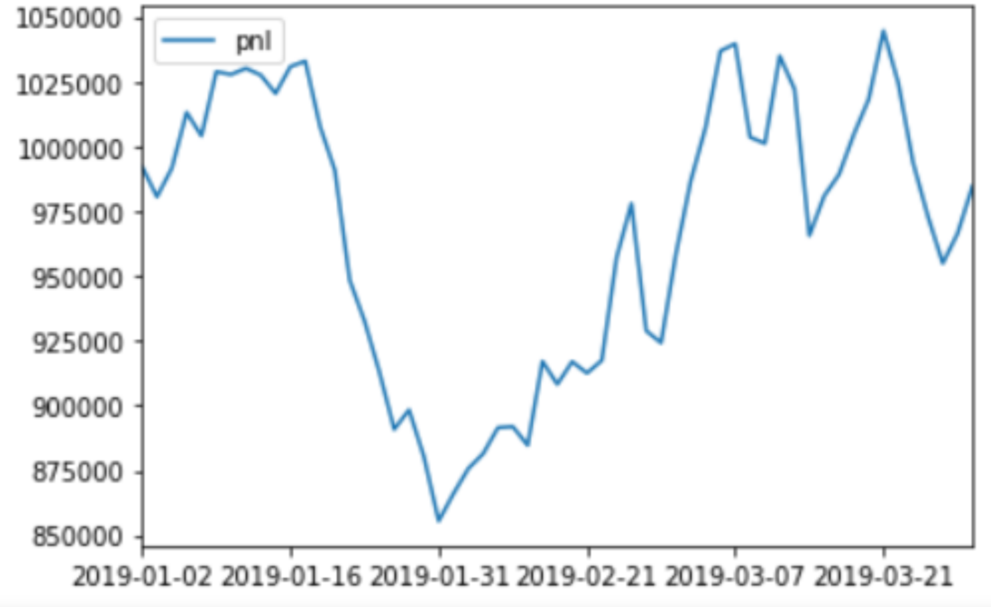


Figure 2. Daily Trade Accumulated PNL

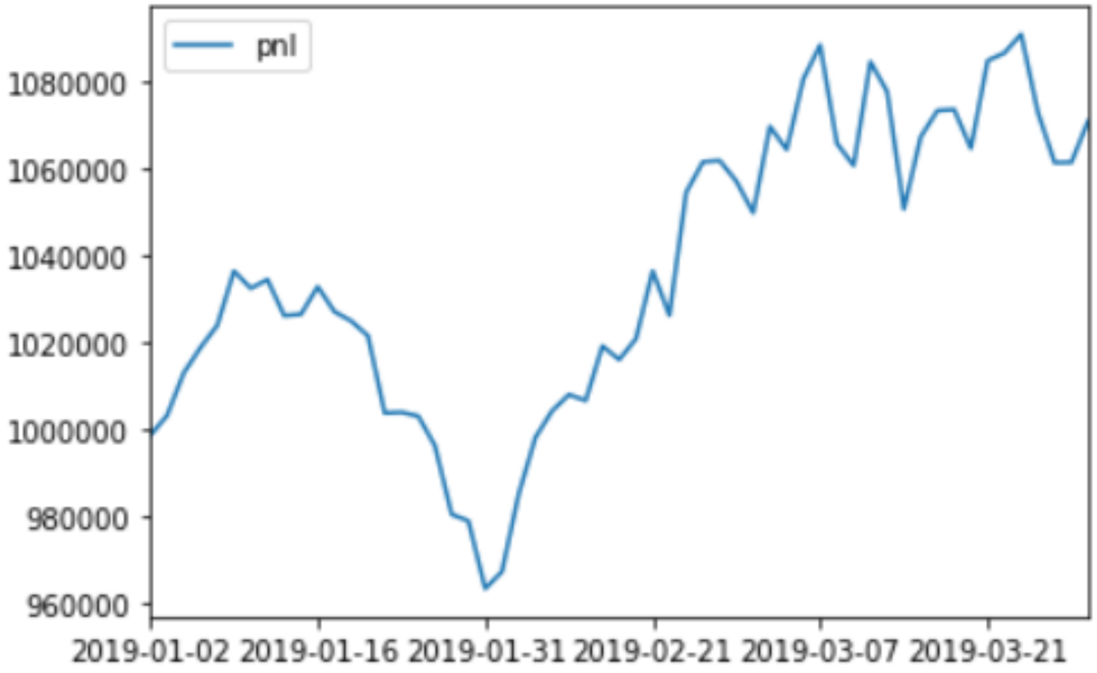


Figure 3. Skewness Factor Accumulated PNL

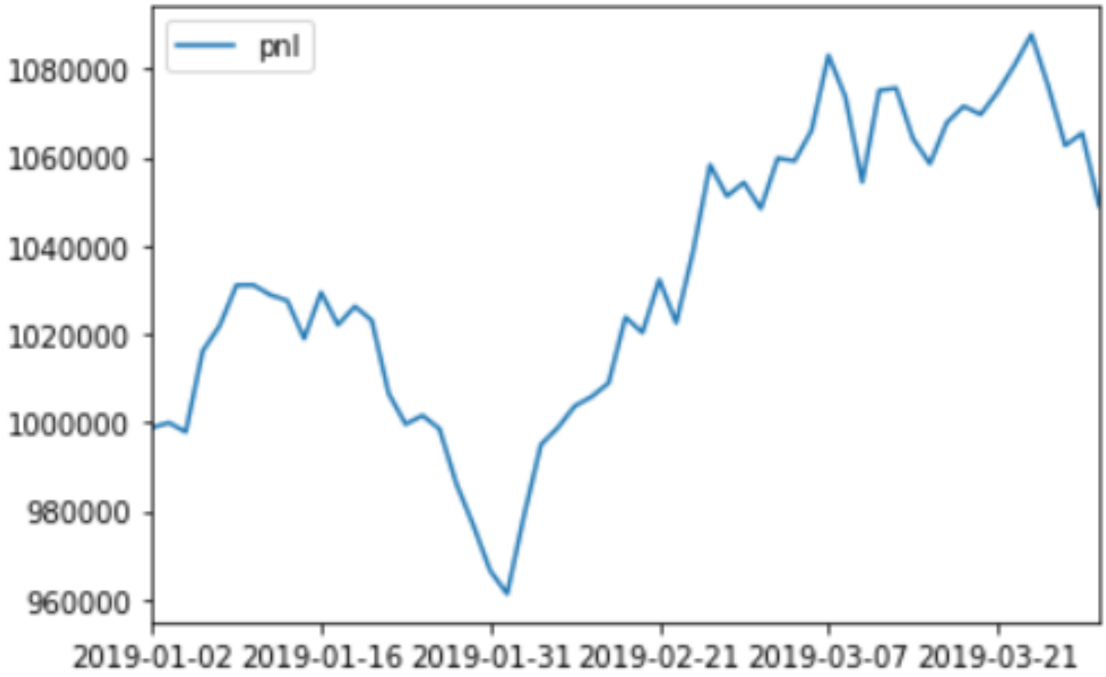
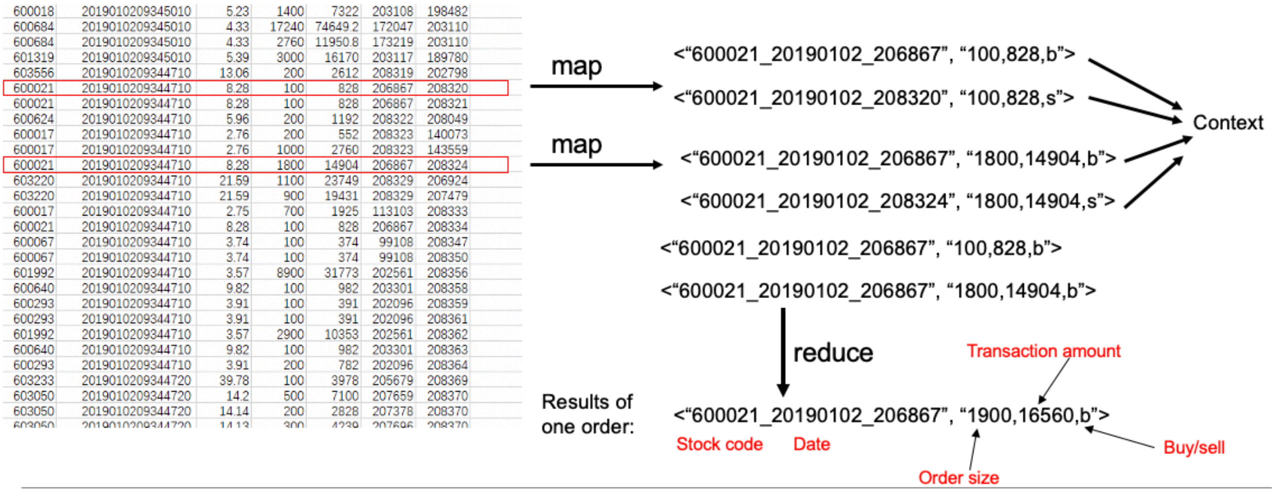


Figure 4. Data Slicing for Strategy II



600018	2019010209345010	5.23	1400	7322	203108	198482
600684	2019010209345010	4.33	17240	74649.2	172047	203110
600684	2019010209345010	4.33	2760	11950.8	173219	203110
601319	2019010209345010	5.39	3000	16170	203117	189780
603556	2019010209344710	13.06	200	2612	208319	202798
600021	2019010209344710	8.28	100	828	206867	208320
600021	2019010209344710	8.28	100	828	206867	208321
600624	2019010209344710	5.96	200	1192	208322	208049
600017	2019010209344710	2.76	200	552	208323	140073
600017	2019010209344710	2.76	1000	2760	208323	143559
600021	2019010209344710	8.28	1800	14904	206867	208324
603220	2019010209344710	21.59	1100	23749	208329	206924
603220	2019010209344710	21.59	900	19431	208329	207479
600017	2019010209344710	2.75	700	1925	113103	208333
600021	2019010209344710	8.28	100	828	206867	208334
600067	2019010209344710	3.74	100	374	99108	208347
600067	2019010209344710	3.74	100	374	99108	208350
601992	2019010209344710	3.57	8900	31773	202561	208356
600640	2019010209344710	9.82	100	982	203301	208358
600293	2019010209344710	3.91	100	391	202096	208359
600293	2019010209344710	3.91	100	391	202096	208361
601992	2019010209344710	3.57	2900	10353	202561	208362
600640	2019010209344710	9.82	100	982	203301	208363
600293	2019010209344710	3.91	200	782	202096	208364
603233	2019010209344720	39.78	100	3978	205679	208369
603050	2019010209344720	14.2	500	7100	207659	208370
603050	2019010209344720	14.14	200	2828	207378	208370
603050	2019010209344720	14.13	300	4239	207696	208370

map
<“600021_20190102_206867”, “100,828,b”>
<“600021_20190102_208320”, “100,828,s”>
Context
map
<“600021_20190102_206867”, “1800,14904,b”>
<“600021_20190102_208324”, “1800,14904,s”>
<“600021_20190102_206867”, “100,828,b”>
<“600021_20190102_206867”, “1800,14904,b”>
reduce
Transaction amount
Results of one order:
<“600021_20190102_206867”, “1900,16560,b”>
Stock code
Date
Buy/sell
Order size

Figure 5. Rough Test of Strategy II

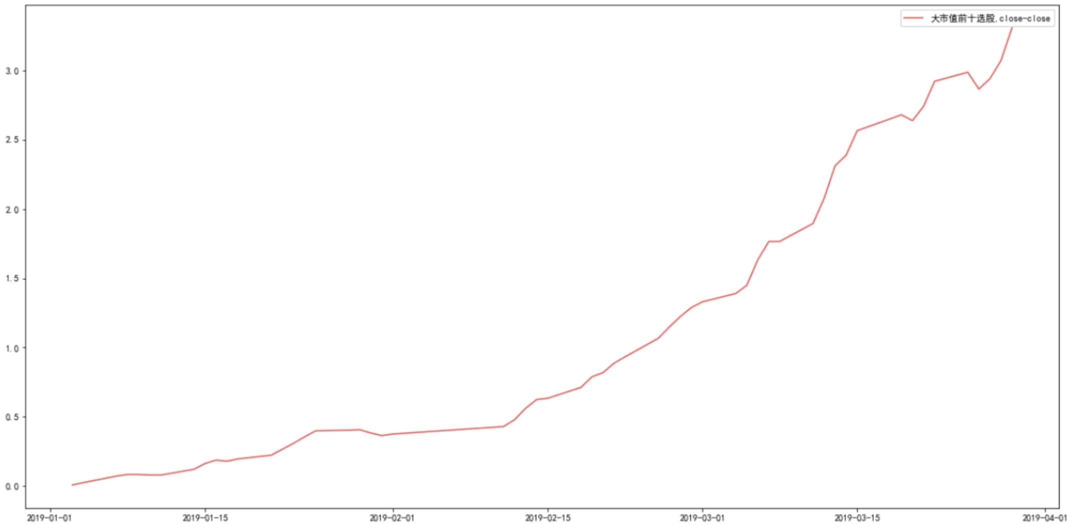


Figure 6. Second Test for Strategy II

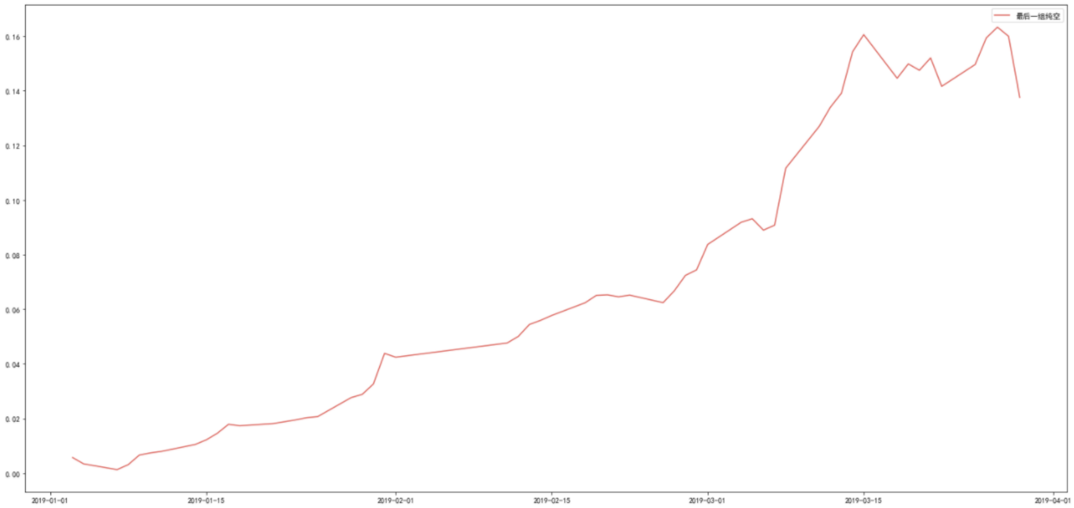


Figure 7. Daily IC Histogram

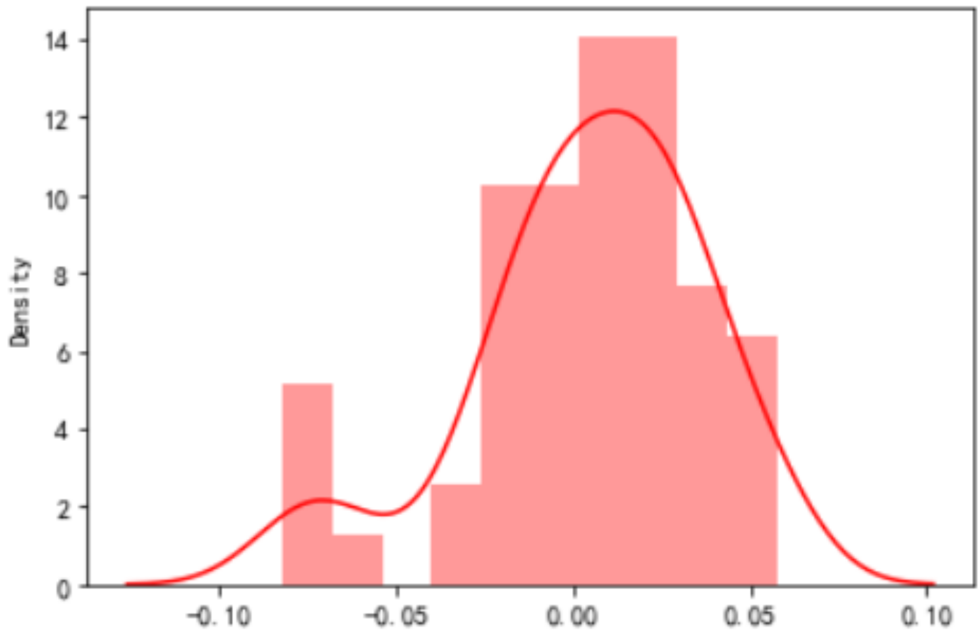


Figure 8. Cumulative Income of Pure-Long Strategy (Sorted by Factor Value from High to Low)

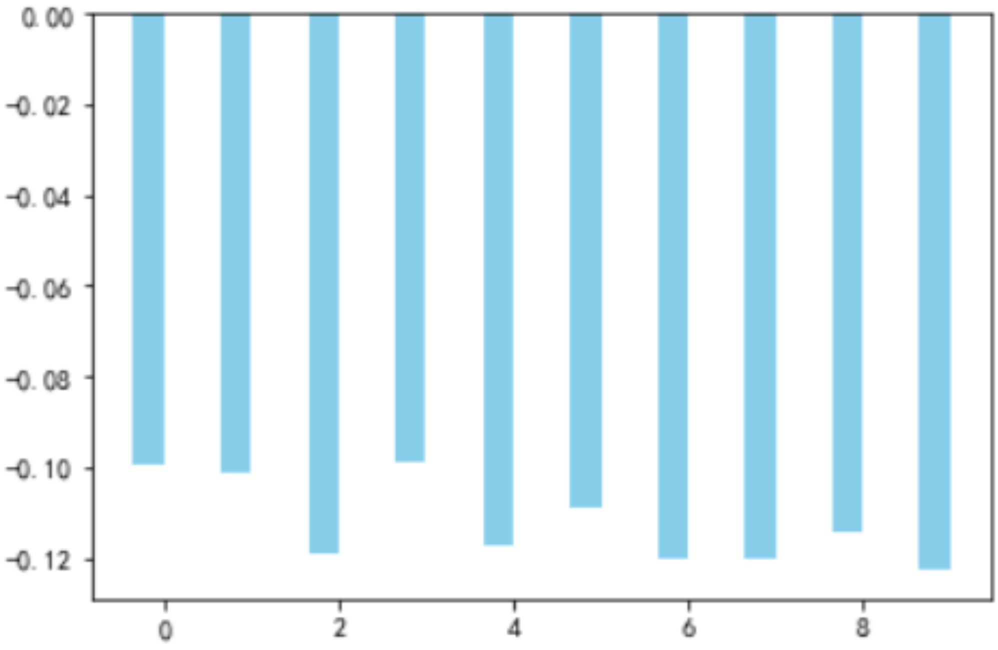


Figure 9. Daily IC Histogram (excluding last minute)

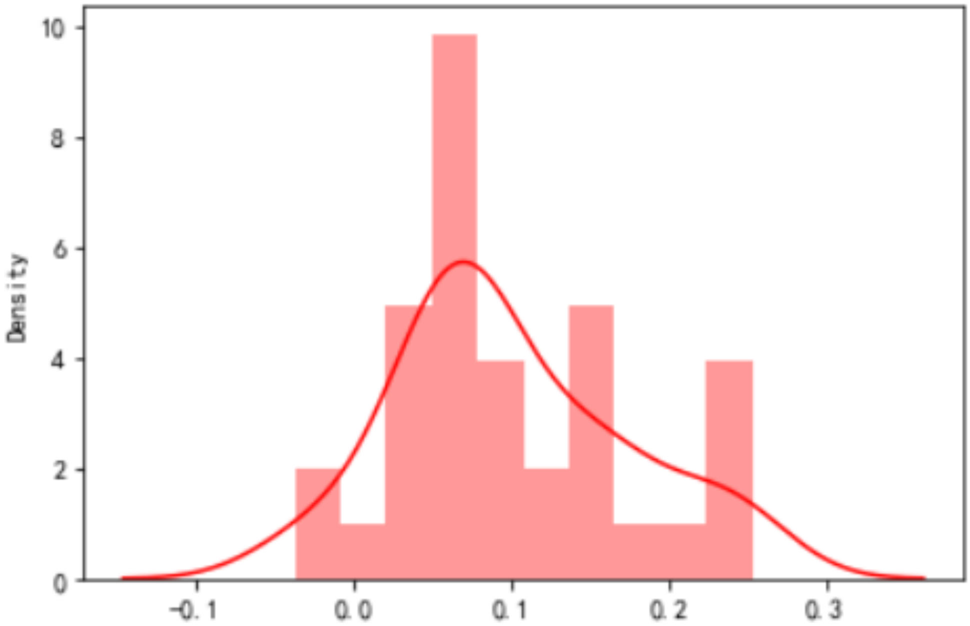


Figure 10. Cumulative Income of Pure Long Strategy (Sorted by Factor Value from High to Low)

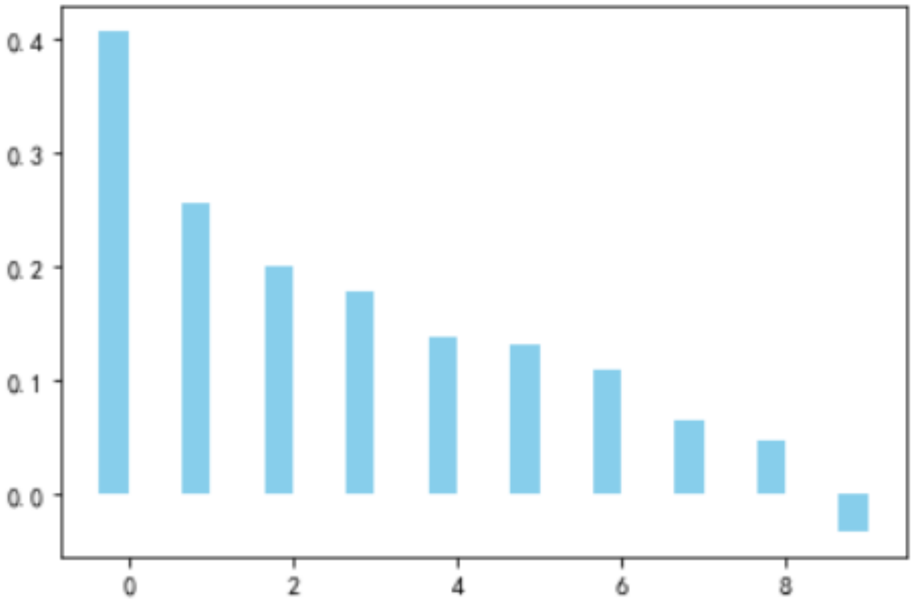
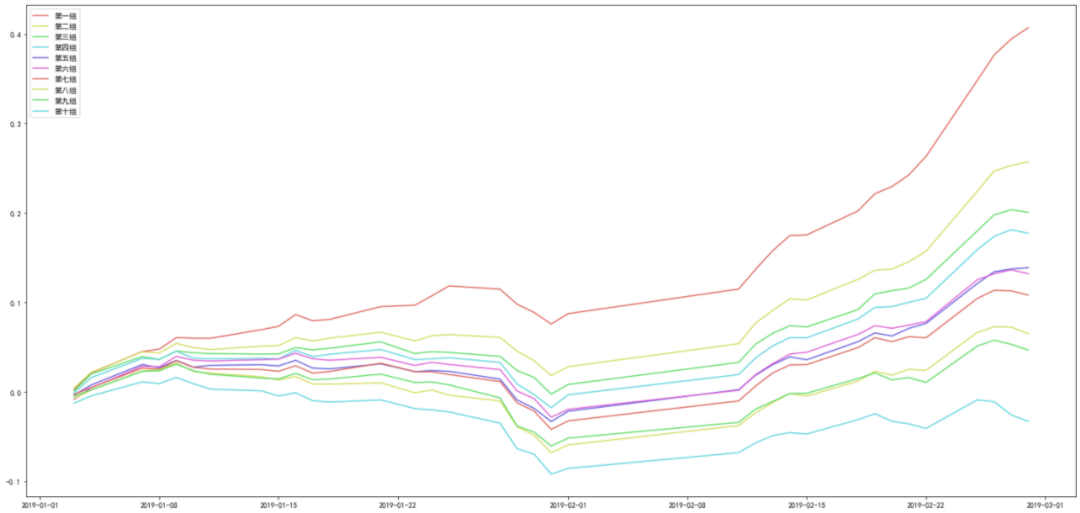


Figure 11. Cumulative Returns Grouped by Factors in Two Months


第一组
第二组
第三组
第四组
第五组
第六组
第七组
第八组
第九组
第十组