Multi-factor Sentiment Strategy on A-stocks

Longpeng Xu

August 31, 2021

Summary

Figure 2 Each factor against HS300

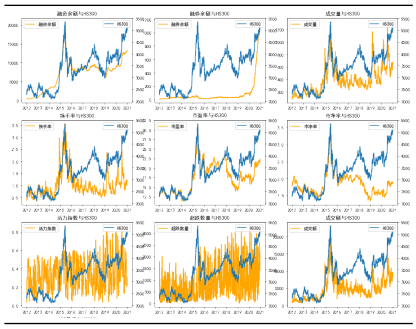


Figure 4 The cumulative return of each strategy against HS300

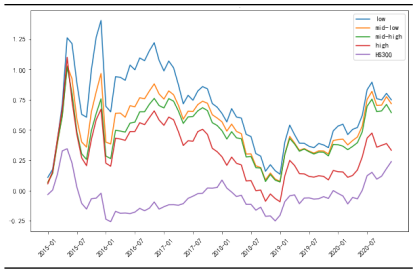
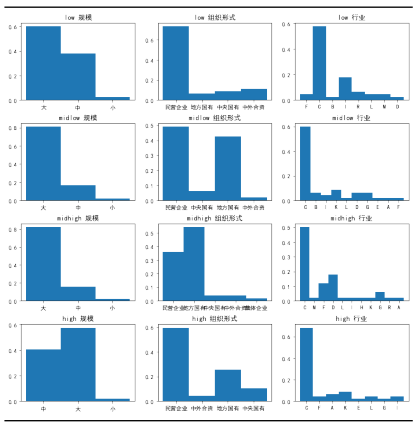


Figure 5 The size, ownership and industries the component stocks belongs to for each strategy



▶ **Market sentiment index is reasonably constructed**

Obtain the daily financing balance, securities lending balance, trading volume, turnover rate, PE ratio, PB ratio, active index, oversold quantity, trading volume, and IPO quantity of all A-stockss from 2012 to 2020. Among them, the active index and oversold quantity are calculated from the closing price of the range, MA20, the highest price and MA10. The data of the first two years of the month are cut out month by month, and PCA is used for dimensionality reduction, and the daily sentiment index each month is obtained by weighted summation of the variance explanation ratio. Figure 2 shows that except for the active index and the number of oversold, other factors show a certain extent of positive correlation with the HS300 index.

▶ **Market sentiment affects stock returns**

The model used is sentiment-CAPM time-series regression. Caculate individual stock i return $R_{i,t}$, HS300 return $R_{\text{mkt},t}$, and market sentiment S_t at time t. When regressing on a stock, slice S_t within 12 months prior to t and define its coefficient $\beta_{i,t}$ as sentiment factor.

▶ **Stratified sentiment strategy harvests promising returns**

Divide all monthly stock sentiment factors into low, medium-low, medium-high, and high sensitivity strategies, and calculate the monthly list of component stocks and average return of each strategy. Cumulative returns are further calculated based on average returns. Figure 4 shows that all four strategies outperformed the HS300, and the relative advantage of the strategies' cumulative returns did not change in the backtest range: low > medium low > medium high > high sensitivity strategy. Finally, in 2020.12, the cumulative returns of the four strategies were 74.98%, 71.76%, 64.40%, and 33.25% respectively. In addition, Figure 5 shows that the typical stocks that are frequently selected in the four strategies are large or medium-sized, private or local state-owned manufacturing companies.

1 Background research

1.1 Behavioral fiannce

Behavioral finance (BF), random walk theory, modern portfolio theory (MPT) and efficient market hypothesis (EMH) are all financial asset pricing theories emerging in the 1990s. This theory starts from individual behavior and the psychological motivations behind it, trying to explain and predict changes in asset prices and the development of financial markets. Behavioral finance is actually opposed to the efficient market hypothesis, because the former partially explains the following problems, while the latter cannot at all: (1) investors are not completely rational; (2) investors' cognitive biases are systematic rather than Offsetting each other; (3) Arbitrage restrictions, asset deviations from fundamental values cannot create arbitrage opportunities because the costs and risks of correcting mispricing are too high (including fundamental risks, noise trader risks, and various implementation costs). Robert Shiller proposed: "We should remember that stock market pricing is not a perfect science." In 2013, he believed that there was almost no way to accurately predict the direction of the stock and bond markets in the next few days or weeks, but it might be possible to predict prices for more than three years through research (Shiller, 2014). Therefore, this article's backtesting strategy of stratified sentiment based on long-term historical data still has reference value.

1.2 Investors' sentiment

Investors' sentiment can be directly captured by a variety of market indicators. In order to characterize investor sentiment to the greatest extent, we focus on the selection of sentiment proxy indicators and the construction method of sentiment composite indicators.

- Lin (2013) focuses on (1) the number of oversold stocks and the degree of oversold. (2) Comparison of the market performance of oversold stocks and other stocks: analysis from the perspective of winning rate (comparison of increases and decreases) and performance (comparison of relative net worth).
- Zhang (2017) focused on the active index, B-class fund discount and premium rates, basis changes, VIX index, and the net increase in holdings of important shareholders. Among them, the active index is based on the moving average theory and is calculated: active index = the number of component stocks above the N-day moving average/the number of all component stocks in the index; the net increase in holdings of important shareholders is selected because when investors believe that the company's stock price is undervalued, they tend to increase their holdings, otherwise they reduce holdings.
- Ren & Liu (2020) followed the ideas of Baker & Wurgler (2006) and empirically added the Baker-Wurguler index to improve the effect of CAPM, FF3F and other models. The index includes components: closed-end fund discount and premium rates, NYSE stock turnover rate, number of IPOs, first-day yield of newly listed stocks, proportion of issued equity, and dividend premium. Among them, NYSE stock turnover rate = NYSE stock trading volume/range average number of shares (natural logarithm, 5-year moving average).
- Chen (2016) used Chinese fund total volume index, HS300 index, the number of new investors, stock market trading volume and closed-end fund discount rate.
- Qi (2019) focuses on the HS300 60-day strong stock proportion, HS300 eight moving average indicator, HS300 relative strength indicator, A-stock turnover rate, stock index futures premium and discount and other indicators. In addition, overall market indicators such as overall PE ratio, PB ratio, and

trading volume have also been mentioned by many researchers.

To construct composite indicators,

- Ren & Liu (2020) and Chen (2016) both used the principal component analysis to eliminate the common information contained in different sub-indicators. Furthermore, they paid attention to the time sequence relationship between different sub-indicators and market sentiment, and selected sub-indicator data from the previous period or the current period. In addition, they processed the data as follows: (1) Standardization, so that each sub-indicator has the same contribution to the sentiment index and has unit variance. (2) Adjust the division of time intervals. Because the sentiment index has different interpretations of fund returns under different market environments, it is divided into 5 different intervals according to the level of the index value. (3) Set the investor sentiment factor as a dummy variable, and use the interval quantile ranking results 1-5 of the Baker-Wurgler Sentiment Index to replace the actual numerical results to improve regression results.
- Zhang (2017) introduced and demonstrated the application of the max. diversification (MD) model in the construction of composite indicators.

In summary, the return volatility of each sub-indicator within a moving time window is first calculated, and then the MD algorithm is applied to configure the optimal weight for each sub-indicator based on these volatilities and their covariance matrices.

2 Construct market sentiment index

2.1 Active index and oversold

This report defines the active index as

$$\text{active index} = \frac{\text{the number of A-stocks with closing price} > \text{MA20}}{\text{the total number of A-stocks}}, \tag{1}$$

and defines oversold as

$$\text{oversold} = \text{the number of A-stocks with the highest price} < \text{MA10}. \tag{2}$$

The data obtained is the daily closing price, highest price, MA10 and MA20 of all A-stocks from 2012 to 2020. The code idea is relatively simple, which is about reverse melting the mentioned panel data into 1D, then merging according to the definition (the closing price is merged with MA20, the highest price is merged with MA10) and compared to calculate the corresponding indicator.

2.2 Select sentiment factors

Based on the above research and the availability of data, the following indicators are selected for the sentiment index: financing balance, securities lending balance, trading volume, turnover rate, PE ratio, PB ratio, active index, oversold, trading volume, the number of of IPOs. The data is spans daily from 2012 to 2020. The data here refers to "all A-stock (001004)" data; trading volume and turnover rate are smoothed by the 5-day moving average, and other data are directly obtained. In addition, the HS300 index is obtained. The correlation coefficient heat map between indicators and the relative trend of HS300 are shown in Figure 1 and Figure 2, respectively. It can be found that: (1) Except for active index, oversold and # IPOs, other

indicators have the same trend as the HS300 for at least a period of time; active index and oversold always fluctuate sharply without obvious trends. (2) 2015-2016 was a period of sharp rise before sharp decline of HS300. During this period, financing balance, transaction volume, turnover rate, PE ratio, PB ratio, and transaction volume also showed highly similar trends.

Figure 1 Heatmap between factors

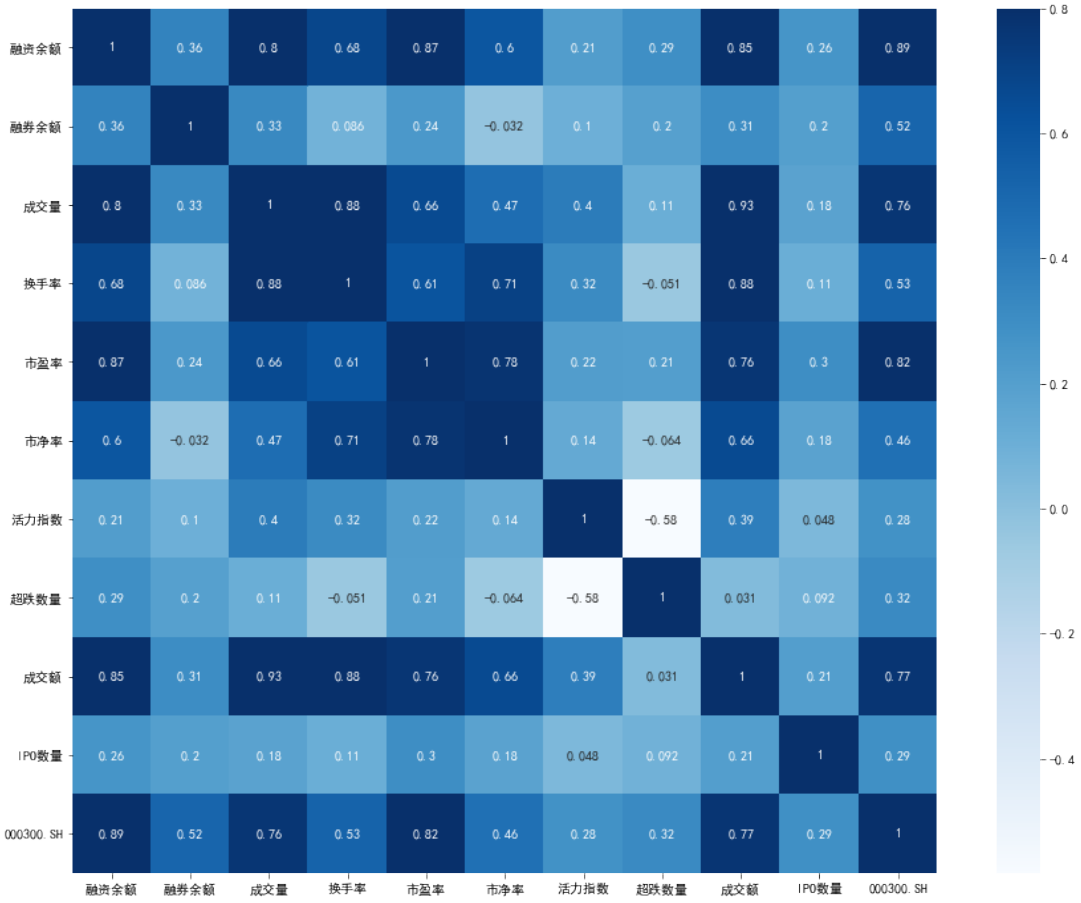
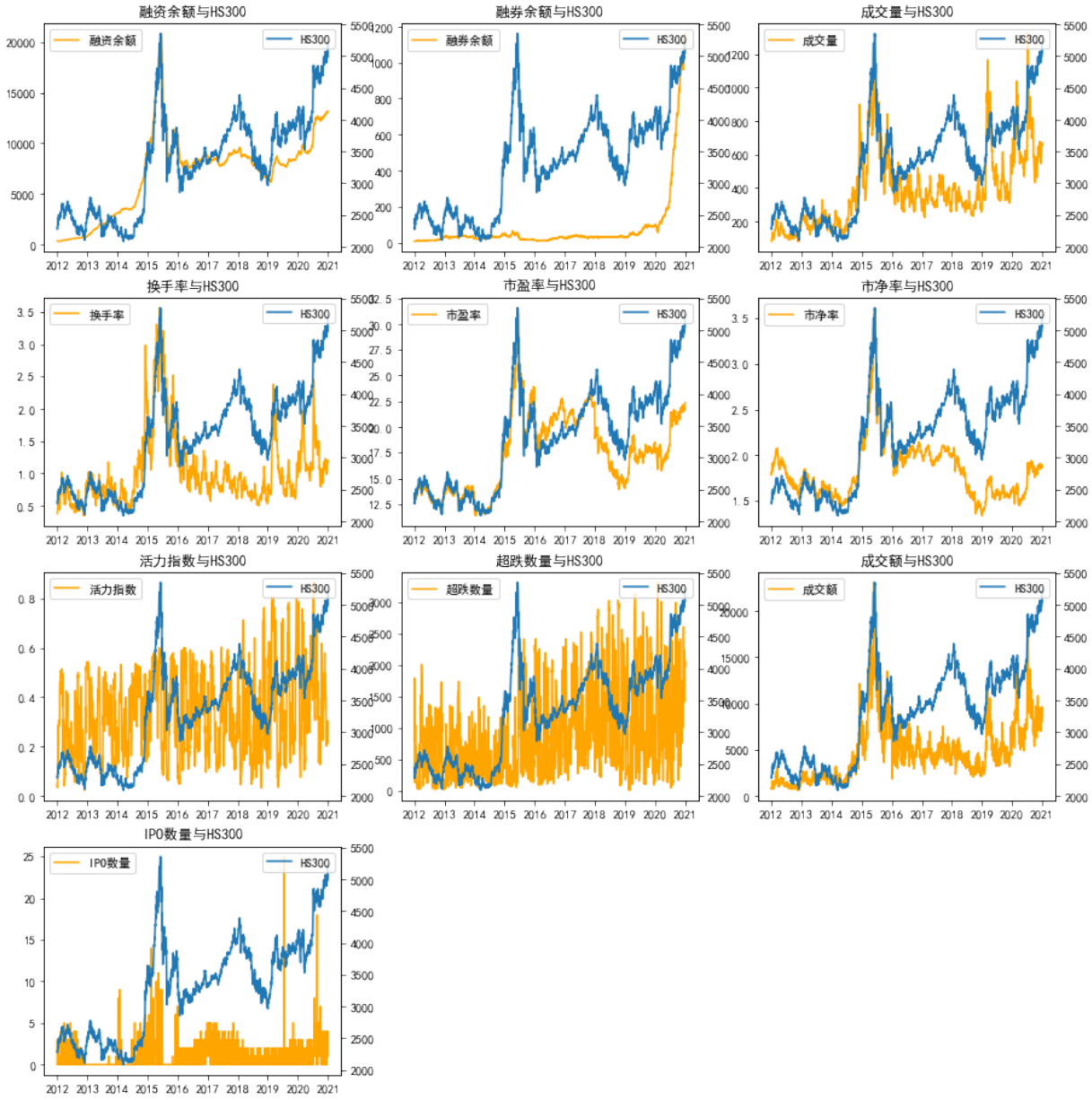


Figure 2 Each factor against HS300 index



2.3 Construct sentiment index rollingly

The sentiment index was constructed by principal component analysis (PCA) where the cumulative explained variance (CEV) ratio was set to 85%. Taking all trading days of 2014.1 as an example, the major coding steps are as follows

(1) Slice the indicator data of 2 years prior to the month (i.e. 2012.1 to 2013.12).

(2) Train data by PCA to calculate the explained variance ratio of each indicator and CEV, and slice the principal components with a CEV hitting 85%, assuming that it is the first five columns.

(3) Select the first five columns of explained variances, and obtain daily data of the indicators. The daily sentiment index from 2012.1 to 2013.12 can be obtained by calculating the dot product of the two.

(4) Slice the last 21 (the number of the trading days in 2014.1) of these sentiment indexes as the daily sentiment index of 2014.1. The above steps are executed in a monthly cycle (so step (1) obtains data on a rolling basis), and the daily sentiment index from 2014 to 2020 can be obtained in the same way.

3 Market sentiment affects stock returns

3.1 Improved CAPM model

Traditional CAPM model depicts the influence of traditional returns by the expected excess return of investment portfolios:

$$r_i = r_f + \beta_i^P(\mathbb{E}R_P - r_f). \tag{3}$$

Introducing sentiment index to the model derives the time-series sentiment-CAPM:

$$R_{i,t} = \alpha_i + \beta_{i,t}^{\text{mkt}} \cdot R_{\text{mkt},t} + \beta_{i,t} \cdot S_t + \varepsilon_{i,t} \tag{4}$$

where $R_{i,t}$ is the return of stock i in day t. $R_{\text{mkt},t}$ is the return of HS300 in day t. S_t is the sentiment index in day t. Hence, $\beta_{i,t}$ can be regarded as "sentiment factor" since it is the regression coefficient/sensitivity against returns of individual A-stocks. This model assume that market sentiment affects stock returns, which is grounded since models like FF3F, Carhart 4-Factor and FF5F do the same, adding other factors to CAPM. In section 2.3 sentiment index has been computed, with daily return of all A-stocks based on close prices being available, it seems that regression is ready to be done and sentiment factor can be computed.

3.2 Transaction status as a filter

In fact, for $R_{i,t}$, it is proper for regression only when the stock i is not at the upper or lower limit and the trading status is normal at time t. Collect data on whether all A-share stocks have a upper limit or a lower limit, and trading status (normal trading, unlisted, etc.) from 2014.1 to 2020.12 into three data frames. According to the intersection ("No" for upper limit, "No" for lower limit and "normal trading"), merge (NOT concatenate) the data frames, taking the binary value 1 or 0. Subsequent regression steps are based on the merged data frame.

3.3 Regress rollingly for sentiment factor

Based on the finished work, for each stock in each month, I can conduct regression on $R_{i,t}$ against $R_{\text{mkt},t}$, and S_t (daily data spanning 12 months prior to the current month), which derives $\beta_{i,t}$ being the sentiment factor of the current month. In coding, it is about:

(1) Extract and merge each stock's return rate, HS300 index return, and sentiment index during 2014-2020.

(2) Extract the data of the stock with unqualified trading status ("0") in this range, and filter the merged data.

(3) Regression month by month, taking 2015.1 as an example, extract the data from 2014.1-2014.12 for simple linear regression to obtain the sentiment factors of 2015.1, and follow this rule to obtain the emotional factors of 2015.2, ..., 2020.12.

By executing the above steps in a stock cycle, I can obtain the monthly sentiment factors of all A-share stocks from 2015 to 2020.

4 Construct stock strategy from sentiment factor

4.1 Sentiment factor and average returns

The sentiment factor $\beta_{i,t}$ is the sensitivity of the sentiment index to individual stock returns, and the sensitivity of different stocks will be different in different months. The sensitivity levels can be divided into four strata: low, medium-low, medium-high, and high. Each sensitivity level can be regarded as one stock selection strategy. Monthly rebalancing is a realistic option in terms of transaction costs, so I can compute the component stocks and the average return of each strategy for each month. When calculating the average monthly return of each strategy, it is assumed that funds are invested equally in each stock, so the average return is the arithmetic mean; in addition, it is necessary to obtain the monthly return of all A-stocks from 2015.1 to 2020.12.

Therefore, in coding, backtesting the segmented sentiment strategy is about:

(1) Extract the sentiment factors of all A-stocks of a month, and use `qcut` to obtain the equi-frequency ranges from low to high.

(2) Map equi-frequency intervals to text labels, for subsequent index slicing being easier.

(3) Fill in the data frame `groups` the component stocks of each strategy every month, which is to extract the stock codes, as index in the slice, of the component stocks of different strategies.

(4) Slice the four columns of components stocks for the current month, slice the returns of all stocks in the month, calculate the average return for each column of component stocks, and append the data frame `returns`.

By realizing the above steps in a monthly iteration, I can get the list of component stocks and the average return of the four strategies every month from 2015.1 to 2020.12.

Although there is more than one coding solutions for backtesting, they are mostly cumbersome. See the code listing below:

Listing 1 Backtesting

```
def fill_groups_returns(i):
        '''<i> passes an integer in range(len(sentiment_beta_table.index))'''
        month = yyyymm[i]

        # Stratify the stocks of this month to get the equi-freqency ranges from low to high
        df_ = pd.DataFrame(sentiment_beta_table.iloc[i,:])
        index_ =  list(df_.index[np.where(np.isnan(df_))[0]])
        df_ = df_.drop(index = index_)
        cls = pd.qcut(df_[month],q = 4)
        cls_ =  list(np.sort(pd.DataFrame(pd.value_counts(cls)).index))
        cls_ = [ str(j) for j in cls_]

        # Map equi-frequency ranges to textual labels
        dis = {cls_[0]:'low', cls_[1]:'mid-low', cls_[2]:'mid-high', cls_[3]:'high'}

        # Label each stocks for this month
        df_['Equi-freq partition'] = cls
```

```
        df_['Equi-freq partition'] = df_['Equi-freq partition']. map(lambda x: str(x))
        df_['sensitivity'] = df_['Equi-freq partition']. map(dis)

        # fill up <groups>
        for m in  range(4):
                tag =  list(dis.values())[m]
                item =  list(df_[df_['sensitivity'] == tag].index)
                groups.iloc[: len(item), int(m+4*i)] = item

        # fill up <returns>
        monthly_returns_slice = monthly_returns.loc[month,:]
        groups_slice = groups[[month + ' ' + n for n in dis.values()]]
        cols_res = []
        for col in  range( len(groups_slice.columns)):
                in_col_res = []
                for row in  range( len(groups_slice.index)):
                        stock = groups_slice.iloc[row,col]
                        if  str(stock) == 'nan':
                                pass
                        else:
                                in_col_res += [monthly_returns_slice[stock]]
                cols_res += [np.mean(in_col_res)]
        returns.loc[month,:] = cols_res

for p in  range( len(yyyymm)):
        fill_groups_returns(p)
```

4.2 Cumulative returns

Based on the monthly average return of each strategy, the cumulative return of each strategy can easily be calculated. The cumulative returns of each strategy in 2015.1 were, from low to high, 0.1081, 0.0687, 0.0649, 0.0545, and the cumulative returns of each strategy in 2020.12 were 0.7498, 0.7176, 0.6440, 0.3325. Hence, the sentiment strategy achieved promising returns. Figures 3 and 4 show the average return and cumulative return of each strategy, with HS300 for comparison. The following findings are from the figures:

(1) Concerning average returns, the four strategies looks highly overlapping with indistinguishable differences. The amplitude of HS300 is smaller than that of the four strategies.

(2) Concerning cumulative returns, the relative relationship of the four strategies is stable: low sensitivity stocks > medium-low sensitivity stocks > medium-high sensitivity stocks > high sensitivity stocks, all of which are higher than HS300 in any backtesting interval.

(3) Before 2018.1, the advantage in net worth by the four strategies over HS300 was obvious. After that, the advantage of high-sensitivity stocks over HS300 narrowed, despite that the margin remained stable.

(4) The cumulative returns of the four strategies fluctuately declined from 2016.1 to 2019.1, before fluctuately increasing.

Figure 3 The return of each strategy against HS300

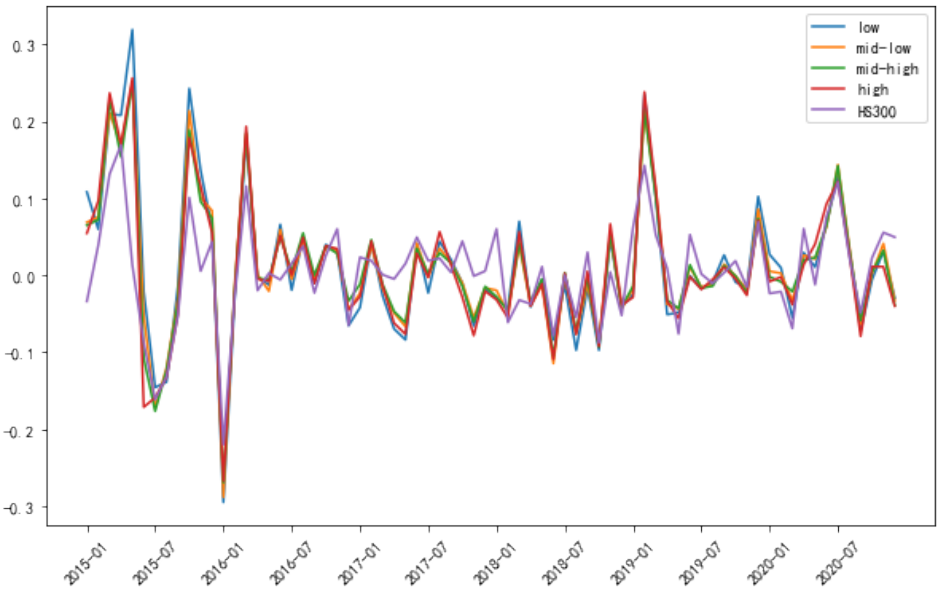
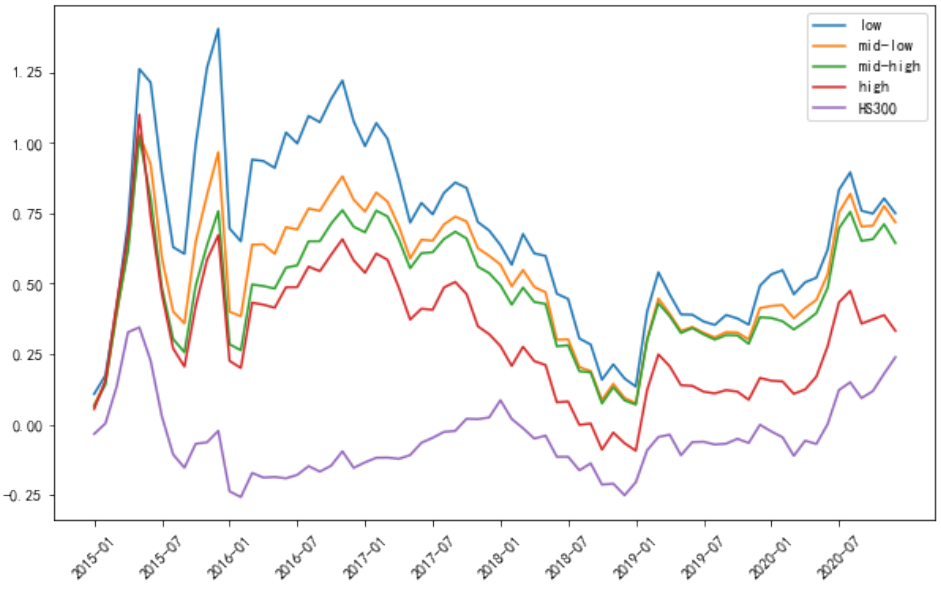


Figure 4 The cumulative return of each strategy against HS300



4.3 Features of component stocks

Merge the lists of low, medium-low, medium-high, and high sensitivity component stocks across all months, extract the 50 most frequent stocks for each strategy. Then, import the characteristics of these stocks, including size (large, medium, and small companies), organizational form (such as private enterprises), industry (such as wholesale and retail). Visualizing these characteristics results in Figure 5 where the vertical axis is relative frequency. For the rightmost subplots of industry distribution in Figure 5, Table 1 maps the encoded industries to their full names, defined by China Securities Regulatory Commission (CSRC). It can be found that:

(1) Comparing sizes, low and high strategies are dominated by large and medium-sized enterprises, while medium-low and medium-high strategies are dominated by large enterprises.

(2) Comparing organizational forms, low and high strategies are dominated by private enterprises. The medium-low and medium-high strategies are dominated by private and local state-owned enterprises.

(3) Comparing industries, the manufacturing industry is dominant among all strategies.

Figure 5 The size, ownership and industries the component stocks belongs to for each strategy

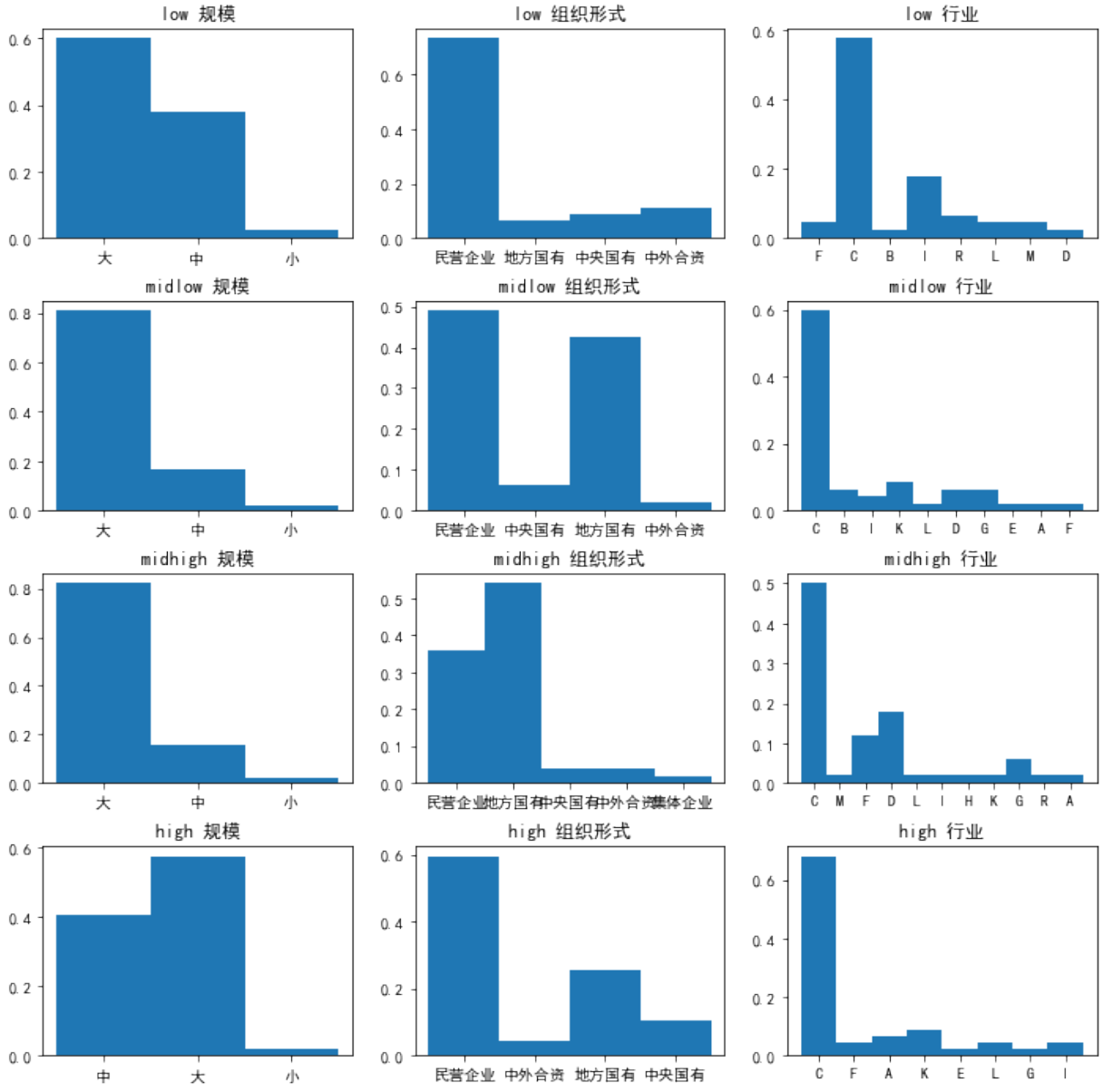


Table 1 Names of industry category by CSRC (excerpt)

Letter	**Industry category by CSRC**
A	Agriculture, forestry, animal husbandry, fishery
B	Mining
C	Manufacturing
D	Electricity, heat, gas and water production and supply
E	Construction
F	Wholesale and retail trades
G	Transportation, warehousing and postal services
H	Accommodation and catering industry
I	Information transmission, software and IT services
K	Real Estate
L	Lease and business services
M	Scientific research and technical services industry
R	Culture, sports and entertainment

5 Conclusions

Investor sentiment can affect stock returns, which is reflected by many indicators. This project selects 10 sentiment indicators, including active index and oversold which are calculated by close prices, highest prices, MA20, and MA10. Other indicators besides these two show some positive correlation with HS300. By integrating these indicators into the sentiment index via PCA, the CAPM model can be improved and the sentiment factors can be obtained using linear regression on a rolling basis. Stratifying sentiment factors into four levels, the stock selection strategy based on the A-stock multi-factor sentiment achieves a cumulative return that outperforms HS300 throughout the entire backtesting interval, with the obvious pattern: the lower the sensitivity of the strategy, the higher the cumulative return it gains. This pattern holds throughout the backtest interval. However, even the highest sensitivity strategy has a final cumulative return of 33%, and the lowest sensitivity strategy has a final cumulative return of 75%.

Further improvements to this project are but not limited to: (1) If possible, include more indicators, like the relative strength indicator, and considering deleting homogeneous indicators from the kept ones; (2) The average returns calculated by arithmetic mean are far from reality, which may require modifying the assumptions. For example, assigning higher weights to stocks with higher market capitalization.

6 Bibliography

Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance (New York)*, 61(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Chen, Y. (2016). Behavioral Finance: Timing and Stock Picking from the Emotional Side. *Northeast Securities: Securities research reports/strategy reports.*

Lin, X. (2013). Quantification of Market Sentiment Indicators Part 2: Timing of Oversold Stocks. *Guosen Securities: Transactional data mining series.*

Qi, Y. (2019). *Reconstruct the emotional system and detect market temperature — Market Sentiment Series Report 2.* Everbright Securities: Financial engineering

Ren, T., & Liu, Y. (2020). *Can investor sentiment explain excess returns of equity funds? Zhuopu Series Report No. 17.* China Merchants Securities: Financial engineering.

Shiller, R. J. (2014). Speculative Asset Prices. The American Economic Review, 104(6), 1486–1517. https://doi.org/10.1257/aer.104.6.1486

Zhang, C. (2017). A-stock multi-dimensional sentiment indicator set and position management. *GF Securities: Financial Engineering Special Reports.*