Method to Compare Two Funds Issued Before and After Stock Crash

Dao Zhang

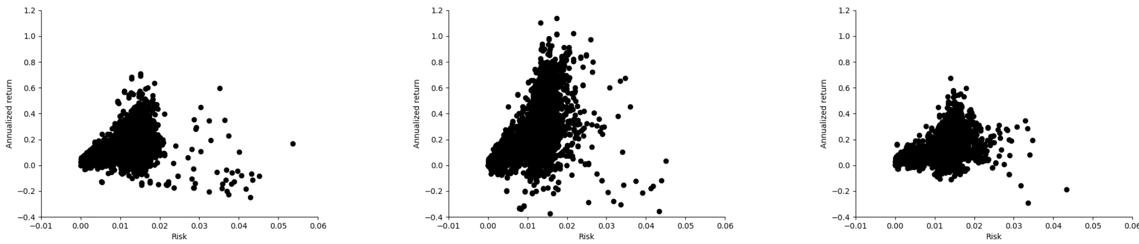
Abstract

Many people prefer to choose stock funds as their main investment variety, but it's not a easy job for us to choose one while there are about ten thousand funds open in China. One method is to compare the annualized return and risk between funds, but it's hard to compare funds with different issuing date using this method. One reason is the funds are deeply influenced by the fluctuation in overall stock market, like bull market or bear market, we have to eliminate this influence when we compare funds. In this paper, I will train a model to get the adjusted accumulative net value 3 years ago for new issues. After adjusted, the comparation between funds with different issuing date is more convincing.

1 Introduction

There are lots of investment varieties, one of the famous variety is the stock investing. But the risk and return coexist in the stock market, and it would spend the investor much time to invest and learn related knowledge. So many people prefer to choose stock funds as their investment variety, the return is higher than deposit and the risk is lower than stock investing, besides, it won't spend us too much time and don't need us to be a specialist in stock market. The portfolio manager is specialized in stock investing and would use portfolio to lower the risk, we only need to pay them a service charge.

There are about ten thousand funds open in China, it's not a easy job for us to choose one fund. People would use some metrics to compare them, like annualized return, risk, etc. I figure the annualized return and risk of all funds in Figure 1(a) and (b). The time range is 3 years and the last day I analyze is Dec 2nd, 2020, all the results in this paper are based on this.



(a) Funds issued more than 3 years

(b) Funds issued less than 3 years (Before adjusted)

(c) Funds issued less than 3 years (After adjusted)

Figure 1: Annualized return and risk

People would choose the funds with higher annulized return if the risk are same. Figure 1(a) and (b) shows we can get the highest annualized return from new issues, but some risk averters prefer old issues which survived from stock market crash, e.g. the crash happend in 2015 and 2018 in Chinese stock market. Some mistakes happen using these metrics.

One reason is the funds are deeply influenced by the fluctuation in overall stock market, for example, almost all funds would rise in a great bull market, beside, almost all funds would decrease in stock market crash. We have to eliminate this influence when we compare funds, the method is to estimate the adjusted accumulative net value 3 years ago, then we can get a adjusted annualized return and use this value to compare with other funds.

I analyze the correlation between funds and find we can't get the adjusted accumulative net value by simply counting its portfolio. But we can find an old issue whose trading strategy is same or similar to any new issue, so I can use the portfolio and value curve in this old issue to train a model to get the the adjusted accumulative net value of new issue.

2 Related work

Actually most of the paper about quantitative investment are focused on stock investment and how to manage the portfolio(Yang et al., 2020; Feng et al., 2019), there are few papers focused on funds and I didn't find any paper considering the influence by fluctuation in overall stock market when invest fund.

3 Methodology

3.1 Problems to Compare Funds with Different Issuing Date

People may use annualized return and risk to analyze one fund, in this paper the return is defined as the growth ratio of accumulative net value from j days ago to i days ago.

$$return_{i:j} = \frac{V_i - V_j}{V_j} \tag{1}$$

where V_i represent the accumulative net value in i days ago.

So the return from N days ago to today can be represented as

$$return_{0:N} = \frac{V_0 - V_N}{V_N} \tag{2}$$

The risk is defined as the standard deviation of growth ratio in one day.

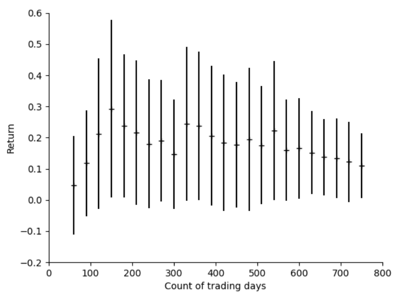
$$risk = \sqrt{\frac{\sum_0^{N-1} (return_{i:i+1} - return_{mean})^2}{N}} \tag{3}$$

The annualized return is defined as

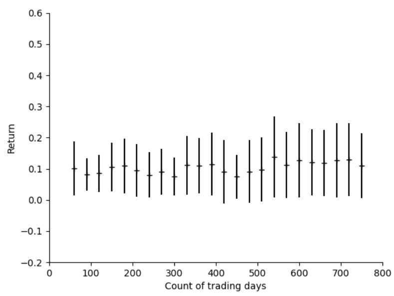
$$annualized\_return = \frac{V_0 - V_N}{V_N} \times \frac{252}{N} \tag{4}$$

as there are 252 trading days in one year.

Comparing Figure 1(a) and Figure 1(b), I find the maximum annualized return in the left one is higher than the right one, about 2 times. To verify it, I summarize the funds with issuing date, count the average and standard deviation of annualized return for every 30 trading days, the results are showed in Figure 2(a).



(a) before adjusted



(b) after adjusted

Figure 2: Distribution of average and standard deviation of annualized return

The annualized return fluctuate with the count of trading days, the funds issued about 150 day ago have the highest return, it's the day after pandamic almost disappeared in China, recovering from pandamic and the outbreak in export stimulate the economy, almost all new issues benefited from this bull market and earned a lot after that, we should eliminate this influence when compare funds.

3.2 Method to Get the Adjusted Accumulative Net Value 3 Years Ago

For a fund, we can get the accumulative net value by counting its portfolio (deposit, bond and stocks). Although it's easy to get the historical value of all stocks in 3 years, there are some difficulties to get the value by counting the historical value of its portfolio.

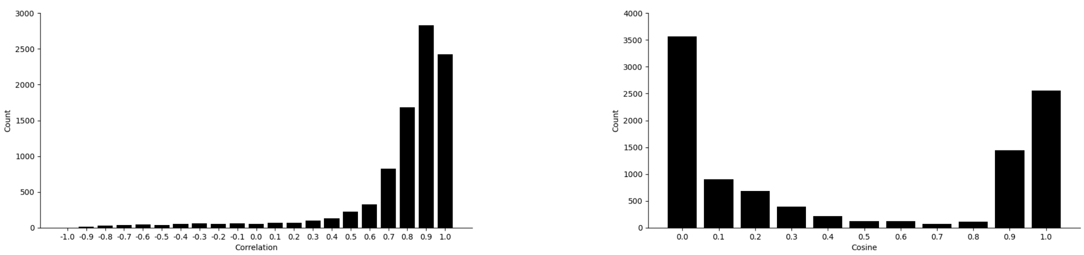
- We can't get the fortfolio between the publishing day. The portfolio managers publish the statements regularly, e.g., three months, but the portfolio would change per week, or per day.
- We can't get the report date of the portfolio in statements. Managers would publish the portfolio one week ago before the publishing date, or a month ago.
- The portfolio published by managers is not a full list, about 10% equities are unknown in some funds.

In this paper I would solve this by another perspective, the treatment are based on a hypothesis that the investing strategies are limited and portfolio managers tend to use same or similar strategy to manager their funds, there are some inferences.

- We can find one fund whose strategy is similar to any fund.
- If portfolio managers use similar strategy, the change of the portfolio would be similar.
- If portfolio managers use similar strategy, the unknown equities would be similar.

The first inference can be verified by Pearson's correlation[1] of value curve between two funds. For any fund, if we can find one fund whose Pearson's correlation is near to 1, this means we can find one fund whose strategy is similar for any fund.

As an example, I analyze the Pearson's correlation between fund *110011* and all other funds in Figure 3(a).



(a) Pearson's correlation between net value

(b) Cosine similarity between portfolio

Figure 3: Distribution of correlation between fund *110011* and all other funds

About 2300 funds' correlation with fund *110011* is 1.0. Besides, most funds have a high correlation with fund *110011*, this is consistent with our qualitative analysis, most funds are related to overall stock market in some extent.

[1] Value of Pearson's correlation ranges from '-1' to '+1'. A value greater than '0' indicates a positive relationship between two variables. Value less than '0' indicates a negative relationship between two variables. '+1' represents these two funds are fully positive correlated.

Furthermore, I analyze the Pearson's correlation matrix between all funds and try to find the maximum correlation for each fund, it seems most of the maximum correlation are near 1.0, the first inference is verified.

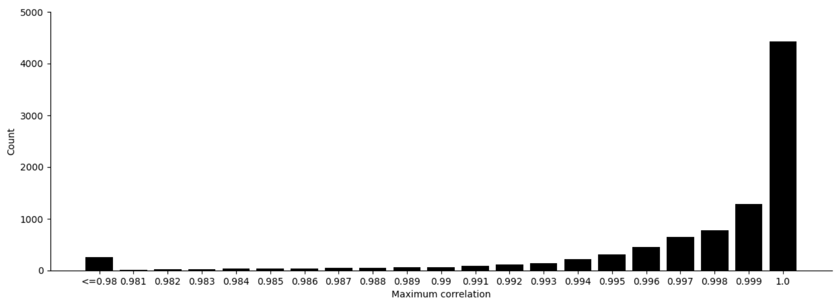


Figure 4: Distribution of maximum Pearson's correlation for all funds

Based on the analysis above, if we can find a similar old issue for a new issue, we can estimate the return in 3 years for new issue by imitating this old issue.

3.3 Find a Similar Old Issue

It's a good idea to use portfolio to find a similar fund, there are 3 parts in portfolio.

- Assets allocation. Including bond portion, cash portion, stock portion.
- Bond. Including 6715 bonds.
- Stock. Including 2855 stocks.

We can use a sparse matrix $P \in N_{9573\times M}$ to represent the portfolio for M funds, where 9573 represent the sum of number in 3 parts. In this section I use cosine similarity[2] to represent the correlation between portfolios. As an example, I analyze the cosine similarity between portfolios in fund *110011* and all other funds in Figure 3(b). It seems most of the value are located around 0 or 1. So we can find some funds with similar portfolio, but funds with similar portfolio is not certain to have similar return.

For fund *110011*, I get the cosine similarity and Pearson's correlation of all other funds and figure them in Figure 5, each point represent a fund, coordinate x represent the cosine similarity between this fund and fund *110011*, coordinate y represent the Pearson's correlation between this fund and fund *110011*.

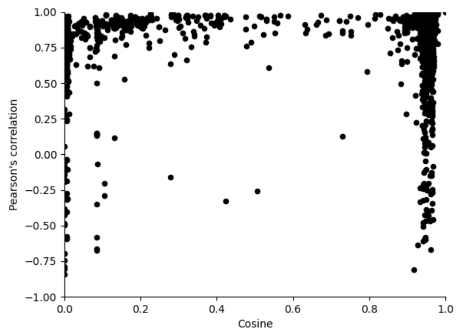


Figure 5: Cosine similarity and Pearson's correlation between fund *110011* and other funds

[2]cosine=0 represent the angle of two vectors is 90° while cosine=1 represent the angle is 0°, the smaller the angle, the higher the cosine similarity

The Pearson's correlation between these 2 metrics are only 0.135, these 2 metrics are not corrlated. It's consistent with our qualitative analysis, two funds with different portfolios may have similar return. And lots of funds with high cosine similarity have a low Pearson's correlation with fund *110011*. So we can't simply choose a fund with similar portfolio to imitate another fund, I would train a model to eliminate the influence of little change in portfolio and unknown equities.

3.4 Model Design

I formulate this problem into a regression problem and aim to estimate V_{756} for new issues, V_{756} is the return in 3 years (756 trading days). To reduce one parameter, the object of the model is to learn the adjusted factor $\alpha = V_N/V_{756}$. Using this factor, we can easily get estimated return by

$$return_{estimate} = \frac{V_0 - V^*_{756}}{V^*_{756}} = \frac{V_0 - V_N/\alpha}{V_N/\alpha} = \frac{V_0 \times \alpha - V_N}{V_N} \tag{5}$$

where V^*_{756} represent the estimated value 756 days ago.

For training dataset, I choose the funds issued more than 3 years as the training dataset. For fund j, the input of each sample contains two part, one is the portfolio $P^j \in N_{9573}$, one is $i \in \{1, 2, ..., 755\}$.

Our model is to find a map f to let

$$\alpha_{i,j} = f(P^j, i), i \in \{1, 2, ..., 755\}, j \in \{1, 2, ..., M\} \tag{6}$$

where P^j represent the portfolio vector in fund j, i represent i days ago.

So the object of training is to optimize the parameters Θ, so as to minimize the loss

$$\Theta = argmin_{\Theta} \frac{1}{755 \times M} \sum_{i=1}^{755} \sum_{j=1}^{M} loss(\alpha_{i,j}, f(P^j, i)) \tag{7}$$

Most funds only publish the largest allocation, a small fraction in portfolio matrix $P \in N_{9573\times M}$. The portfolio is sparse and there exists many missing values, GBDT is less influenced by these factors, so I use GBDT as the training model and use L1 and L2 metrics to evaluate during training. In testing, I only use the last trading day as the input i.

4 Experiments

4.1 Settings

I collect the dataset by crawler and I put the code in github[3], including crawler, data preprocessing, model training and analyzing.

There are 4340 funds issued 3 years ago, we can get 755 training samples in one fund, so there are about 3 million samples in the dataset. I split the dataset into training, evaluating and test, the split ratio is 8:1:1.

I implement the network within LightGBM framework[4]. Thanks to the higher efficiency of this framework, the experiments can be carried out on a laptop with a Inter(R) Core(TM) i7-8650U CPU, 16GB RAM and 4 cores.

As the dataset is too large to load into memory, I use csr matrix[5] to represent this sparse matrix and fill the null value and missing value as zero.

Besides, I use optuna[6] to fine tune the hyper parameters and use root mean square error as the optimization objective, you can find other hyperparameters I tuned in my project.

[3]https://github.com/wangershi/analyzeChineseFund
[4]https://github.com/microsoft/LightGBM
[5]https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html
[6]https://github.com/optuna/optuna

4.2 Results

There are 4543 funds issued less than 3 years, I get the adjusted factor for all new issues.

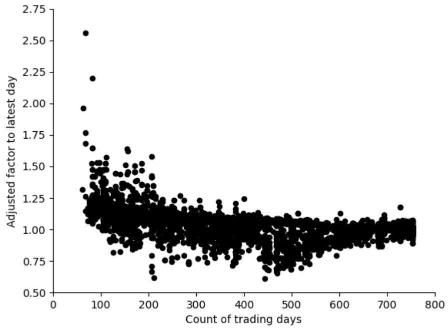


Figure 6: The relation beween adjusted factor and count of founding days

This paper aim to eliminate the influence by the stock market's fluctuation, one metric is whether we can flatten the distribution of average and standard deviation of annualized return showed in Figure 2(a). After fetching the adjusted factor, we can count the adjusted return 3 years ago, then count the annualized return.

$$annualized\_return = \frac{V_0 - V_{756}^*}{V_{756}^*} \times \frac{252}{756} = \frac{V_0 - V_N/\alpha}{V_N/\alpha} \times \frac{1}{3} = \frac{V_0 \times \alpha - V_N}{3 \times V_N} \tag{8}$$

After the return are adjusted, we can get this distribution again in Figure 2(b), it seems the distribution are flattened. Besides, we can see the annualized return and risk of new issues after adjusted in Figure 1(c), the same conclusion can be concluded when comparing it to Figure 1(b).

I count the standard deviation of the average annualized return in Figure 2, it shows the standard deviation drop from 0.0520 to 0.0175, the model is successful to flatten the distribution of average annualized return in some extent.

5 Conclusion

In this paper I analyze the return and risk of all funds in China, and find it's hard to compare funds with different issuing date using annualized return and risk because of the fluctuation in overall stock market. I use the Pearson's correlation and cosine similarity to analyze funds and find we can estimate the return in 3 years for new issues by imitating some old issues. After training a model to get the adjusted factor, the model is successful to flatten the distribution of average annualized return, the comparation between funds with different issuing date is more convincing. But it's still a long way to compare funds with different issuing date, we need to tune the model or design a better model, and conside the change of portfolio in a long inversting period, etc.

References

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. Enhancing stock movement prediction with adversarial training.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. 2020. Qlib: An ai-oriented quantitative investment platform.