Long/Short Statistical Arbitrage Strategy

Boyu Wang

March 10, 2024

1 Executive Summary

In this project, we built a framework for developing long/short statistical arbitrage strategy which consists of three modules: backtesting, single factors and factor combination. We developed 14 single factors and combined these single factors using linear regression and gradient boosting machine. The in-sample and out-of-sample performances of combined factors and single factors are listed in Table 1 and Table 2, respectively. The in-sample period is from Jan 4, 2013 to Dec 31, 2018, and the out-of-sample period is from Jan 1, 2019 to Mar 39, 2021.

Github repository: https://github.com/wangboyu15/stat_arb

Onedrive sharing folder: OneDrive Sharing Folder

	Return(%)	Volatility(%)	Sharpe Ratio	MDD(%)	IC(%)	IR
Combined Factors						
Gradient Boosting	3.75	3.45	1.09	−3.83	0.49	0.05
Linear Regression	1.75	5.05	0.35	−7.20	0.40	0.03
Single Factors						
reversal_1d	5.63	6.06	0.93	−9.01	0.66	0.04
reversal_1y	2.89	6.21	0.47	−9.44	0.01	0.00
skew_1m	1.94	3.65	0.53	−9.44	0.01	0.00
skew_1y	1.35	3.19	0.42	−7.71	0.44	0.05
kurt_1y	2.10	3.28	0.64	−5.15	0.12	0.01
illiquid_1d	1.35	4.12	0.33	−9.28	−0.41	0.03
maxRtn_22d	1.58	5.46	0.29	−11.85	0.09	0.01
salience	5.38	6.07	0.89	−10.00	0.83	0.05
vol_volume_5d	2.86	3.81	0.75	−5.86	0.28	0.03
vol_per_volume_5d	2.60	4.55	0.57	−6.09	−0.21	−0.02
rank_cross_sec_5d	5.05	5.54	0.91	−8.80	0.85	0.06
vov_5d	0.76	4.56	0.17	−11.62	−0.31	−0.02
jump_5d	1.96	5.73	0.37	−8.82	−0.01	−0.00
bollinger_5d	4.05	5.66	0.72	−14.29	0.55	0.03

Table 1: In-sample (Jan 4, 2013 to Dec 31, 2018) performance of combined factors and single factors. Return, volatility and Sharpe ratios are all annualized. MDD stands for maximum drawdown. IC stands for information coefficient, and IR stands for information ratio. All quantities are reported without accounting for transaction fee. Please refer to the Jupyter notebook for the detailed statistics that accounts for transaction fee.

	Return(%)	Volatility(%)	Sharpe Ratio	MDD(%)	IC(%)	IR
Combined Factors						
Gradient Boosting	0.68	5.01	0.14	−12.95	0.26	0.02
Linear Regression	2.89	8.56	0.34	−13.58	0.32	0.02
Single Factors						
reversal_1d	−3.60	9.12	−0.39	−17.93	−0.77	−0.03
reversal_1y	6.70	10.29	0.65	−10.92	0.20	0.01
skew_1m	2.61	5.35	0.49	−7.12	0.66	0.06
skew_1y	2.20	4.40	0.50	−6.55	−0.78	−0.08
kurt_1y	3.84	4.42	0.87	−5.01	−0.27	−0.03
illiquid_1d	−2.15	4.73	−0.45	−7.65	−0.50	−0.04
maxRtn_22d	7.59	7.93	0.96	−8.52	0.92	0.04
salience	−3.38	9.13	−0.37	−18.14	−0.62	−0.03
vol_volume_5d	−2.13	3.97	−0.54	−5.98	−0.32	−0.03
vol_per_volume_5d	0.98	5.80	0.17	−6.35	−0.33	−0.02
rank_cross_sec_5d	−5.35	9.29	−0.58	−23.88	−0.72	−0.04
vov_5d	2.60	6.25	0.40	−7.09	−0.13	−0.00
jump_5d	−0.13	8.49	−0.01	−17.18	−0.34	−0.02
bollinger_5d	−6.92	8.78	−0.79	−16.22	−0.69	−0.03

Table 2: Out-of-sample (Jan 4, 2013 to Dec 31, 2018) performance of combined factors and single factors. Return, volatility and Sharpe ratios are all annualized. MDD stands for maximum drawdown. IC stands for information coefficient, and IR stands for information ratio. All quantities are reported without accounting for transaction fee. Please refer to the Jupyter notebook for the detailed statistics that accounts for transaction fee.

To mitigate the risk of overfitting and look-ahead bias, we made sure that we only used data in the in-sample period to develop single factors and train models. The out-of-sample data is only used to test the final model. The testing period is chosen such that the length of the training period is roughly 70% of the whole dataset, and the testing period also covers significant and representative economic scenarios, including Covid-19 and the beginning of the Fed's monetary tightening policy. This helps inspect the robustness of the strategy.

2 Backtest Framework

We consider daily strategy; that is, we rebalance the portfolio on each trading day. We are interested in long/short statistical arbitrage strategy which long and short a portion of assets in the universe simultaneously. At the beginning of each trading day, we have a target weight for each asset, denoted as $w_t = (w_{t,1}, \dots, w_{t,N})$, where N is the number of assets (which is 225 in our project). The target weight is suggested by the predictive signal, which satisfies, for any $t = 1, 2, \dots,$

$$\mathbf{1}^\top w_t = 0, \quad \|w_t\|_1 = 1,$$

Denote $r_t = (r_{t,1}, \dots, r_{t,N})$ the vector of returns, where $r_{t,i}$ is the return of asset i at day t, and the portfolio gross return is given by

$$r_t^{\text{gross}} = w_t^\top r_t.$$

The turnover is give $\Delta w_t = w_t - w_{t-1}$, and the portfolio net return is given by

$$r_t^{\text{net}} = r_t^{\text{gross}} - c\|\Delta w_t\|_1,$$

where c is the transaction fee.

3 Single Factors

We developed 14 single factors, which are listed below. Please refer to the Jupyter notebook Single_Factors_Calculation.ipynb for the detailed calculation.

1. reversal_1d: short-term (1 day) reversal factor.
2. reversal_1y: long-term (1 year) reversal factor.
3. skew_1m: short-term (1 month) risk premium factor for bearing higher moment (skewness) risk.
4. skew_1y: long-term (1 year) risk premium factor for bearing higher moment (skewness) risk.
5. kurt_1y: long-term (1 year) risk premium factor for bearing higher moment (kurtosis) risk.
6. illiquid_1d: short-term (1 day) illiquidity factor.
7. maxRtn_22d: a factor based on theory of behavioral finance.
8. salience: a factor based on theory of behavioral finance.
9. vol_volume_5d: volatility of trading volume
10. vol_per_volume_5d: volatility generated by per unite of training volume.
11. rank_cross_sec_5d: the short-term cross-sectional rank
12. vov_5d : volatility of volatility factor
13. jump_5d: jump risk factor
14. bollinger_5d: Bollinger high band

4 Factor Combination

We adopted two approaches to combine the single factors. The first approach uses linear regression, and the second approaches uses gradient boosting machine.

Please refer to the Jupyter notebook Factor_Combination.ipynb for the detailed implementation.

Data preprocessing. We standardize each factor in cross-section, and fill the missing values using "ffill" with a maximum period of 252 days.

Rolling training and prediction . We re-train each model every 1 year, and do this in a rolling basis. In each training, we use all the historical data that was available by then. For example, on 2016-01-01, we will training the model using all the data before 2015-12-31 and do the prediction from 2016-01-01 to 2016-12-31.

5 Conclusions

We observed significant regime shift, in the sense that the out-of-sample performance of most of the single factors and the combined factors deteriorated. The linear regression model is robust in the sense that the in-sample performance and out-of-sample performance do no differ significantly, while the out-of-sample performance of gradient boosting machine method is far worse than in-sample performance.

Possible reasons include: (1) The single factors lose most of the predictive power that was observed in sample. This might due to change in the economic condition, especially in the face of a pandemic period. We observe sharp plunges around Mar 2020 for all the factors, which is the time of the meltdown of the US equity market and many other major markets in the world. (2) The factor combination model is subject to overfitting risk. We do observe that the linear model is more robust than tree-based models. The overfitting risk is especially prominent when the the number of data sample is small, which is the case for mid-to-low frequency strategies. There are possible solutions to this problem, including developing more features, using statistical methods such as bootstrapping to generate more data, using synthetic data, etc.

6 Future Research

There are several directions for future research. First, it is possible to develop more predictive signals. Second, we may try different factor combination model. Third, we can investigate portfolio optimization method. It is also possible to investigate end-to-end approaches, meaning we directly optimize for the portfolio weight, instead of first combing predictive signals and then optimize portfolios.