# QuantResearch Internship

##### Content

## Chapter One：APT,Factors and alphas

1. APT:Clearance of Confusion about ‘Factors and Alphas’
2. Statistical Arbitrage
3. Trend Following or Forecasting
4. Evaluation of Factors: IC,IR and Others
5. Evaluation of Alphas: Sharpe,Drawdown and Others

## **Chapter Two：Machine Learning**

1. Formulaic Alpha：Genetic Algorithm and Reinforcement Learning
2. GBDT：XGBoost and LightGBM
3. RNN/LSTM
4. Transformer
5. CNN
6. All DeepLearning Structures Are Doing Feature Engineering
7. Reinforcement Learning

## **Chapter Three：Time Series**

1. ARIMA
2. GARCH
3. Kalman Filter
4. Hidden Markov Model
5. Signals Processing and Filters

## **Chapter Four: Past, Now and Future**

1. Statistical Arbitrage, High Frequency and Derivatives
2. Reinforcement Learning
3. Others

# Chapter One：APT,Factors and alphas

1. APT:Clearance of Confusion about ‘Factors and Alphas’

When I just started to do the Alphas’ or Factors’ mining works , I find it really confusing when people talking about factors or alphas what they really mean.Sometimes an alpha means the excess return as the intercept in APT, sometimes an alpha means the regression value of APT serves as a forecast to trade, and sometimes an alpha counld serve as a factor in other models to form another alpha.

Now, I have gained some knowledge about this industry so I will simply list what I know without detailed explaination cause this paper is merely a proof of work not an essay or something else.

Arbitrage pricing model is the core:

By doing regression to present return terms or future return terms,we can see different factors model such as CAPM or APT have really different meanings.

1. Statistical Arbitrage

For the models like CAPM, we are doing regression to the present return terms.

In these case,the factors are the same for all stocks, the betas of factors for different stocks are different. So the factors here serve as a background mechanism that generate the profits such as some macroeconominc reasons. And statistically speaking, factors are random variables with different means and variances.

One way to utlize the factors model is Active Equity Portfolio Management.For simplicit,if we treat the distributions of factors as normal distribution with different means and variances, and we have that different stocks have diferent betas over factors.

Now it’s pretty like Markowitz Mean-Variance model that we can allocate the portfolio to minimize the overall risk.And since stocks always have non-zero intercepts in the APT (because the market is not efficient), we can have an excess return with a minimum risk.

Another way to utlize the model that you can also allocate the portfolio to focus on the factors you want by increasing the portfolio’s exposure on the corresponding factors, in which situation you might think that factors might perform strong.

1. Trend Following or Forecasting

We can also do the regression to future return terms,for example,the next term return:

As an example, WorldQuant defines alpha as a mapping from data of stocks to values--each value for each stock. And we can according to the value to decide the positions of the stocks.

As to this situation, factors can have different values for different stocks,and betas reflect the correlations between future returns and factors. Now factors captures more specific properties of individual stock, such as price-volume factors and style factors which can contribute to the prediction of future returns.

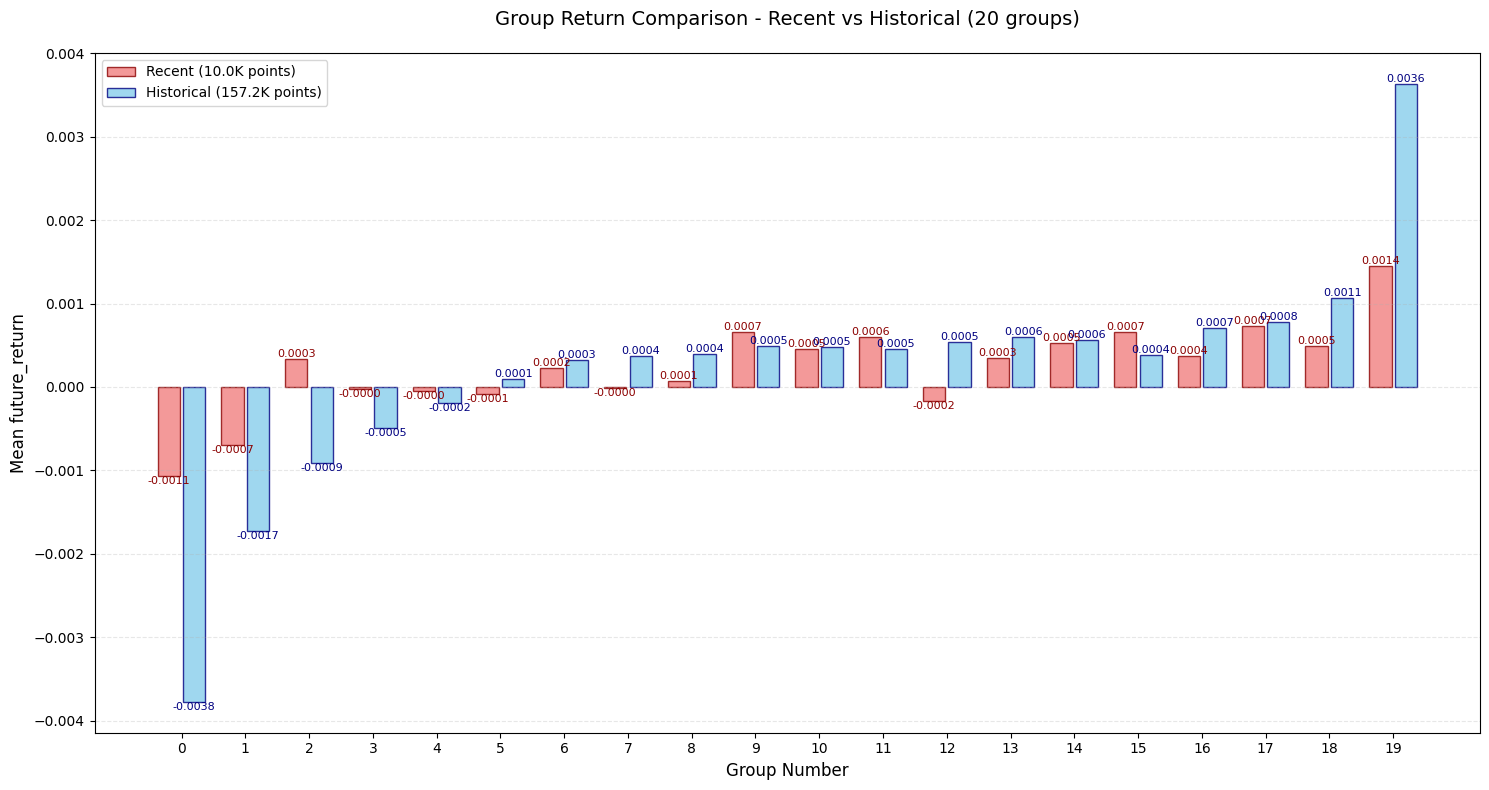
An important component of any successful investment strategy is forecasting expected returns using alpha models. So the selecting and evaluating of factors in a model is important.

1. Evaluation of Factors: IC,IR and Others

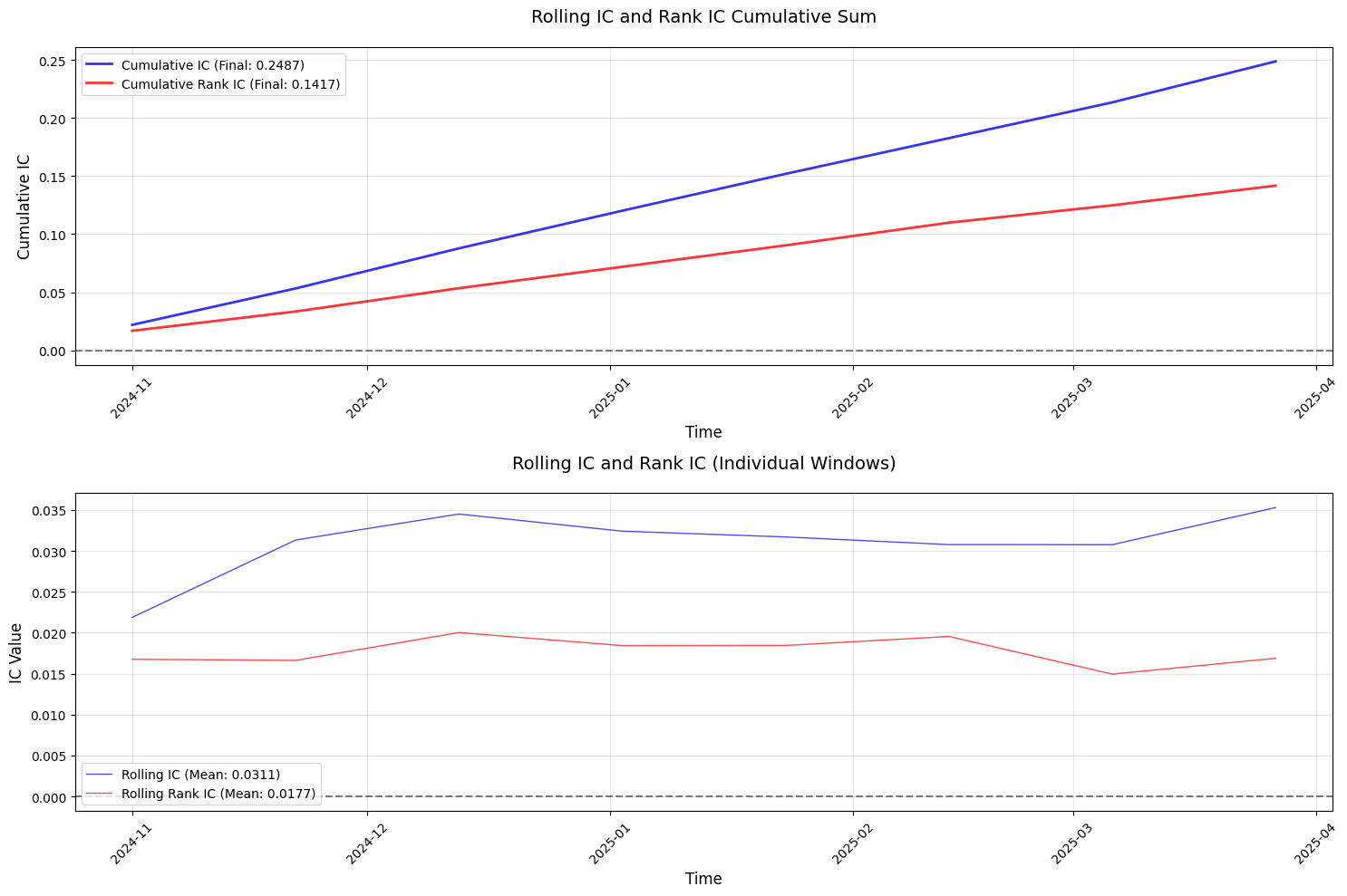
Information Coeffiecient is the Pearson correaltion coefficient for the target values and factor values.Higher IC is better. Also we have rank\_IC.

Information Ratio defines as the IC’s mean divided by IC’s standard deviation ,sort of a Sharpe ratio.

Stratified Testing:



Rolling window test



1. Evaluation of Alphas: Sharpe,Drawdown and Others

Sharpe ratio;

Drawdown;

Hit ratio;

Sortino ratio;

AUC,for classification models;

# Chapter Two：Machine Learning

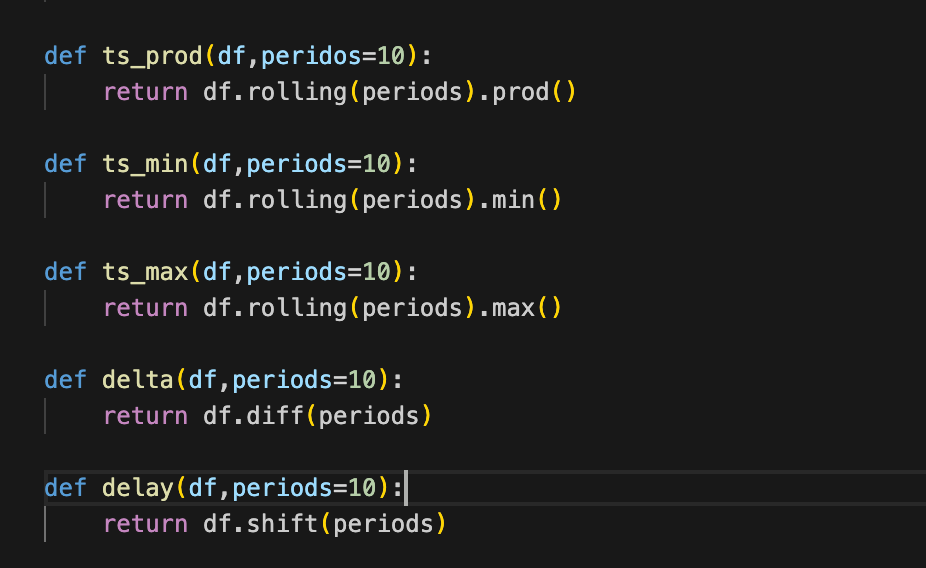
1. Formulaic Alpha：Genetic Algorithm and Reinforcement Learning

As we can see,lots of factors or alphas are functions of price-volume data from the assets, and they can be expressed as formulaic combinations of fundemental functions.And we can use maching learing methods to search for such efficient conbinations.

One of the most matural and flexible method is Genetic Algorithm(GA).Basically,it imitates the evolution process as the evolution theory in biology and I will simply demonstrate the usage of it.

Luckily,we already have ‘gplearn’ ,which is a python library implementing GA.

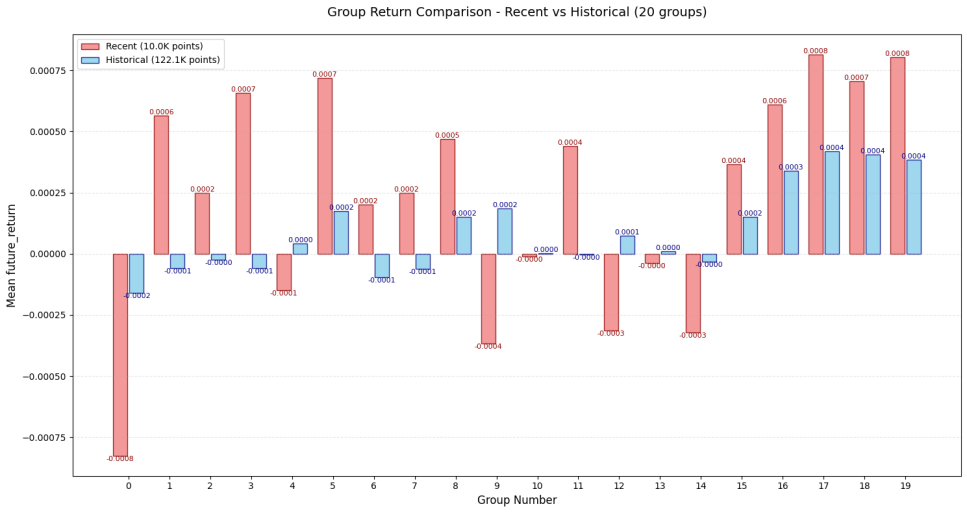
To use it,we need to determine the functions set we use.Except some basic functions as sin ,exp, etc,we can use ,for example,the functions from WorldQuant’s famous paper <101 formulaic alphas>,and we can define some customized functions accroding to our needs.For example:

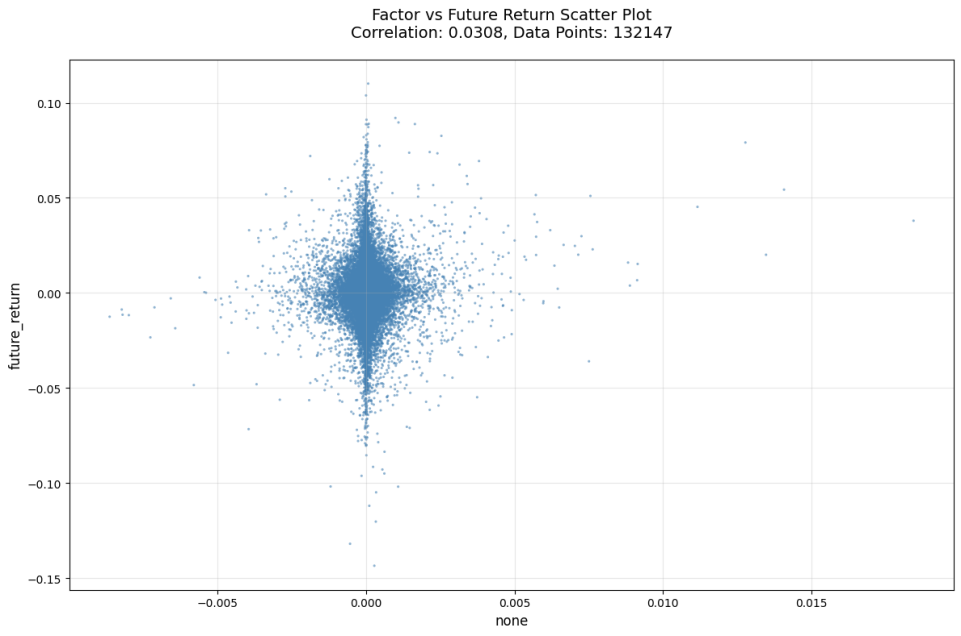


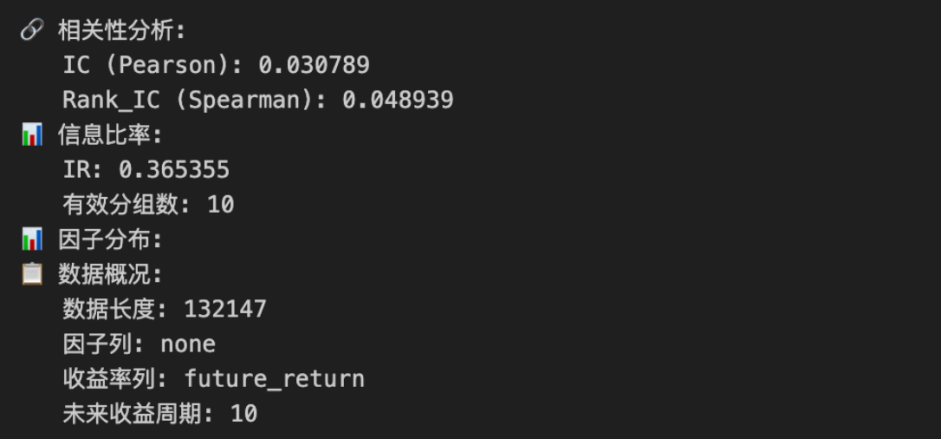
Then,with the data and functions,we can implementing GA to search for the formulaic alphas expressions.

You can see some implementation of gplearn on my Github repository:<https://github.com/wudidawangchongchongchong/Quant-Research-.git>

And I put one GA alpha performence here:







If we follow the idea that we can use the powerful computer to search the massive possible formulaic alphas, Reinforcement Learning is another promising and powerful method.

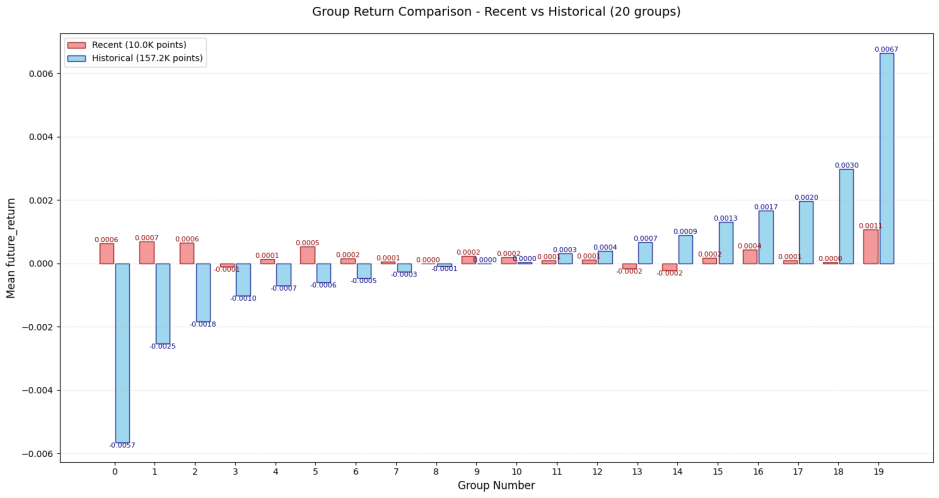
1. GBDT：XGBoost and LightGBM

Gradient Boosting Decision Tree is one kind of Emsemble Learning which focus on gradually fit the residual from the previous trees to have a overall better performence.I main used two popular GBDT models during my quant research intership:XGBoost and LightGBM.

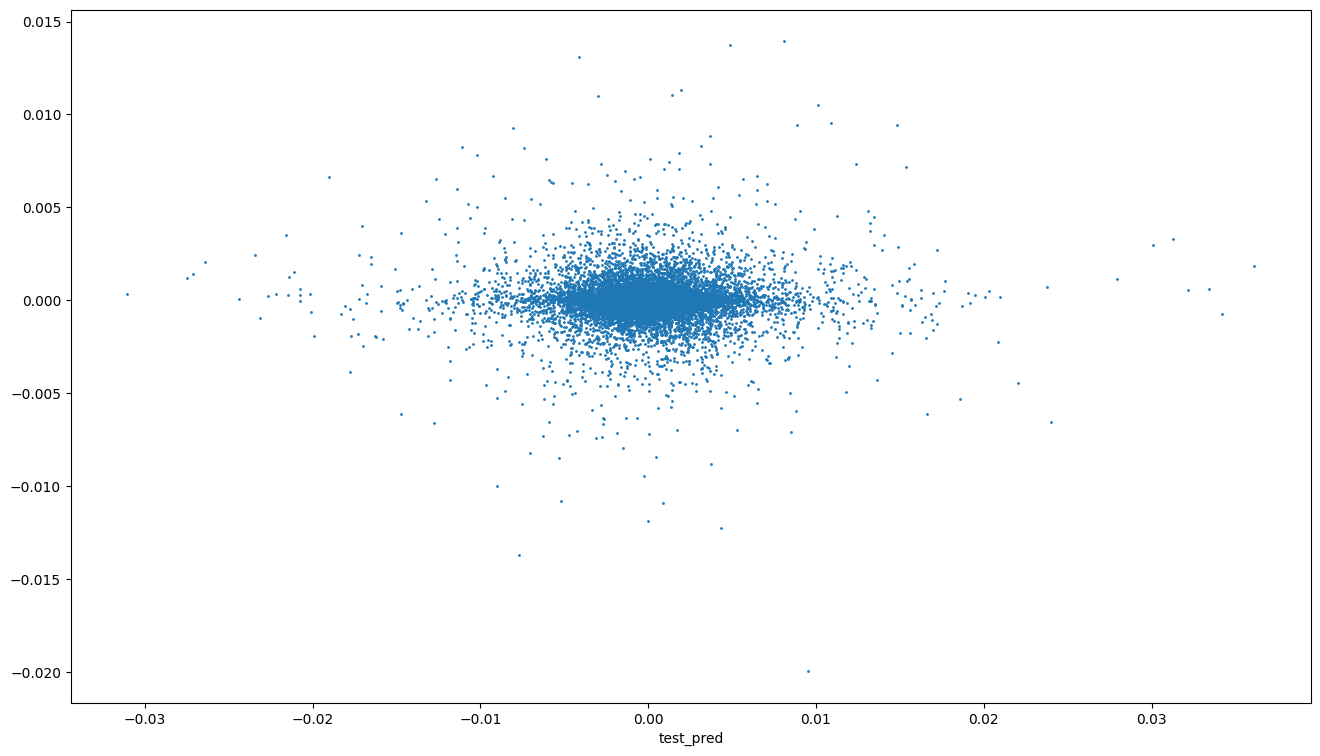
Tree models can be either used to regression tasks or classification tasks，for example:

1. Regressor to predict expect returns;
2. Classifier to predict future asset’s price movement direction:up ,down or no move;

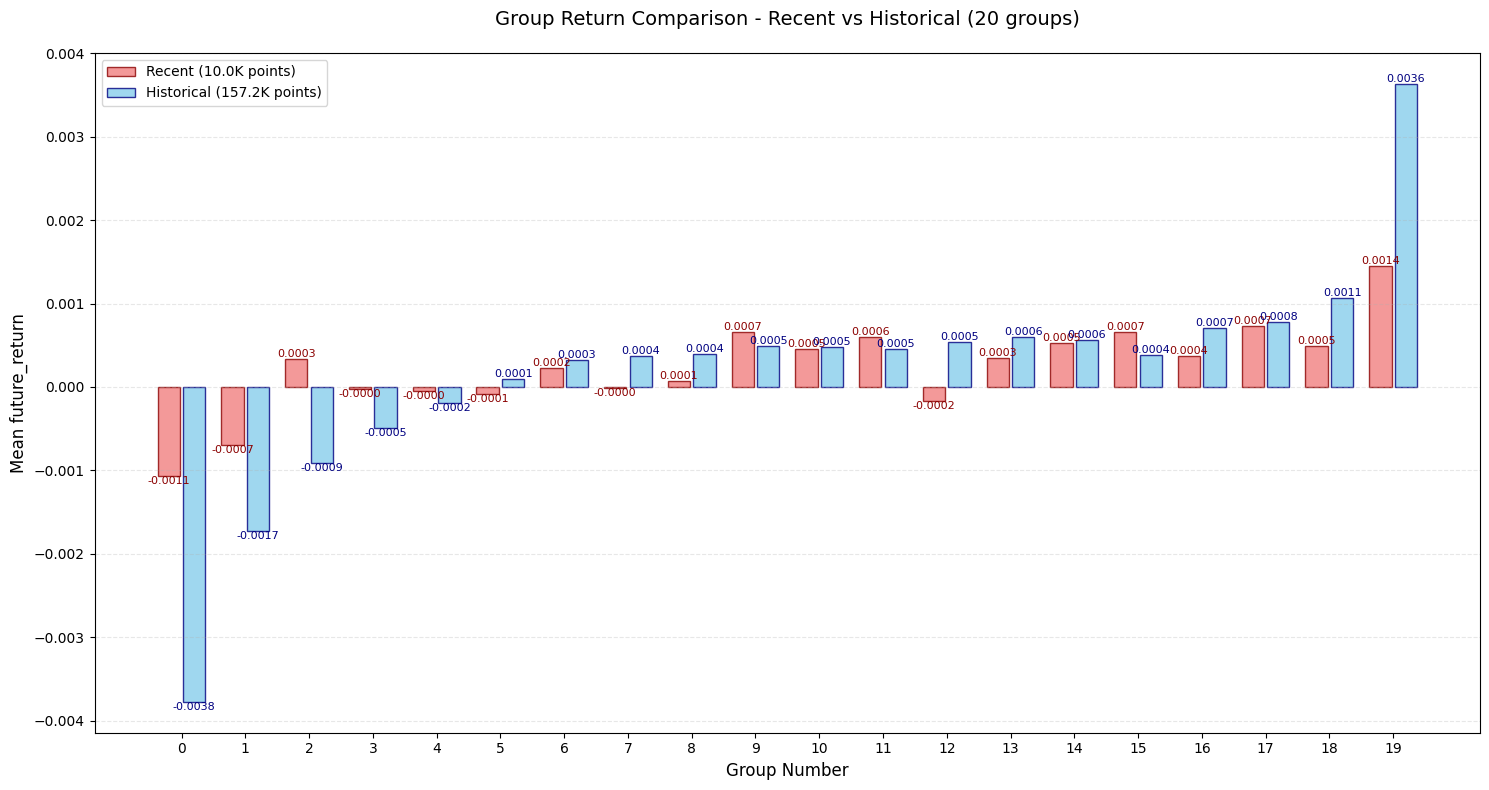
An XGBoost regressor to predict expect returns example:



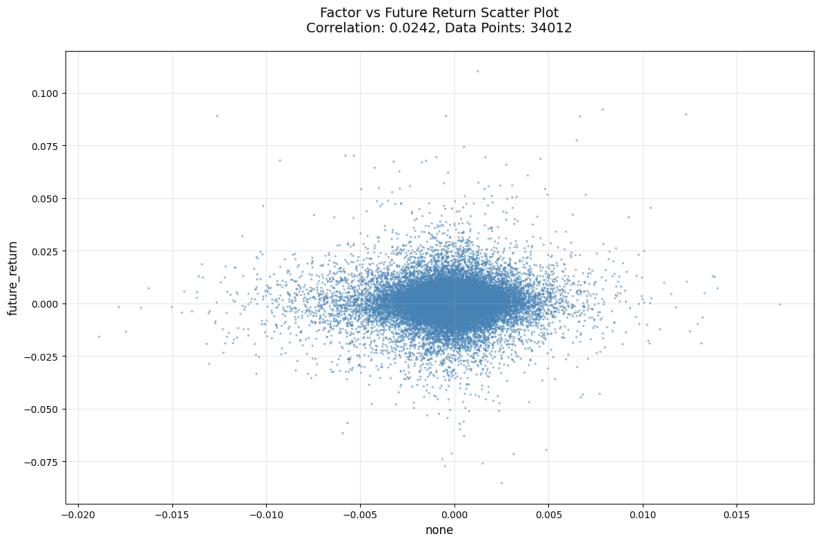
In this naive example,because I used the historical data to train the XGBoost tree,clearly it’s overfitted in the training data.If we take a look at it’s performence on the test data,the IC is insignificant.But of course we can improve the results by optimizing the hyperparameters and feature engineering ,etc.

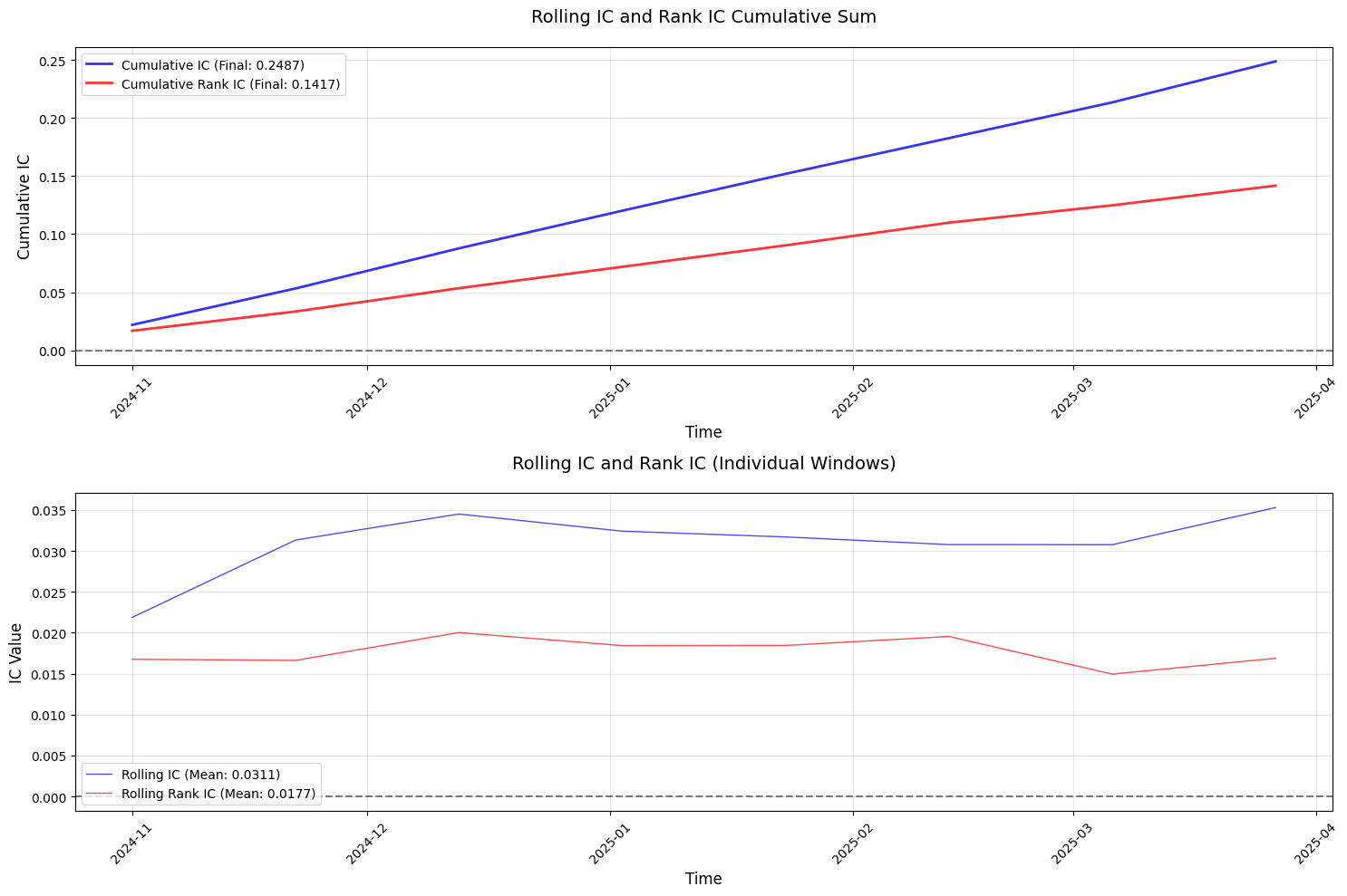


Similarly, below shows a lightGBM regressor example:

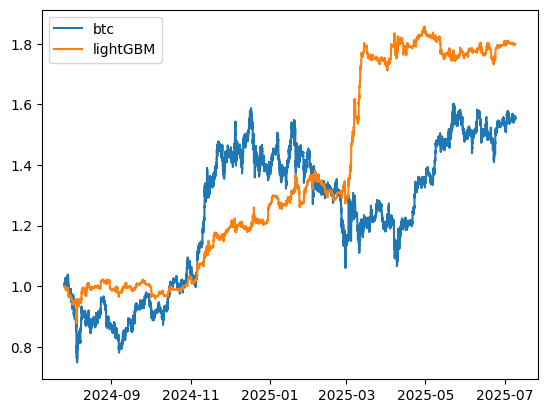


Well, although it still performs too well on the traning data,but it performs better on the testing data;





And we do a simple backtesting on the test data:we only trade the head and tail part the alphas,with the transaction fee,and single interest



Annualized Return：63.78%

Sharpe Ratio：5.6862

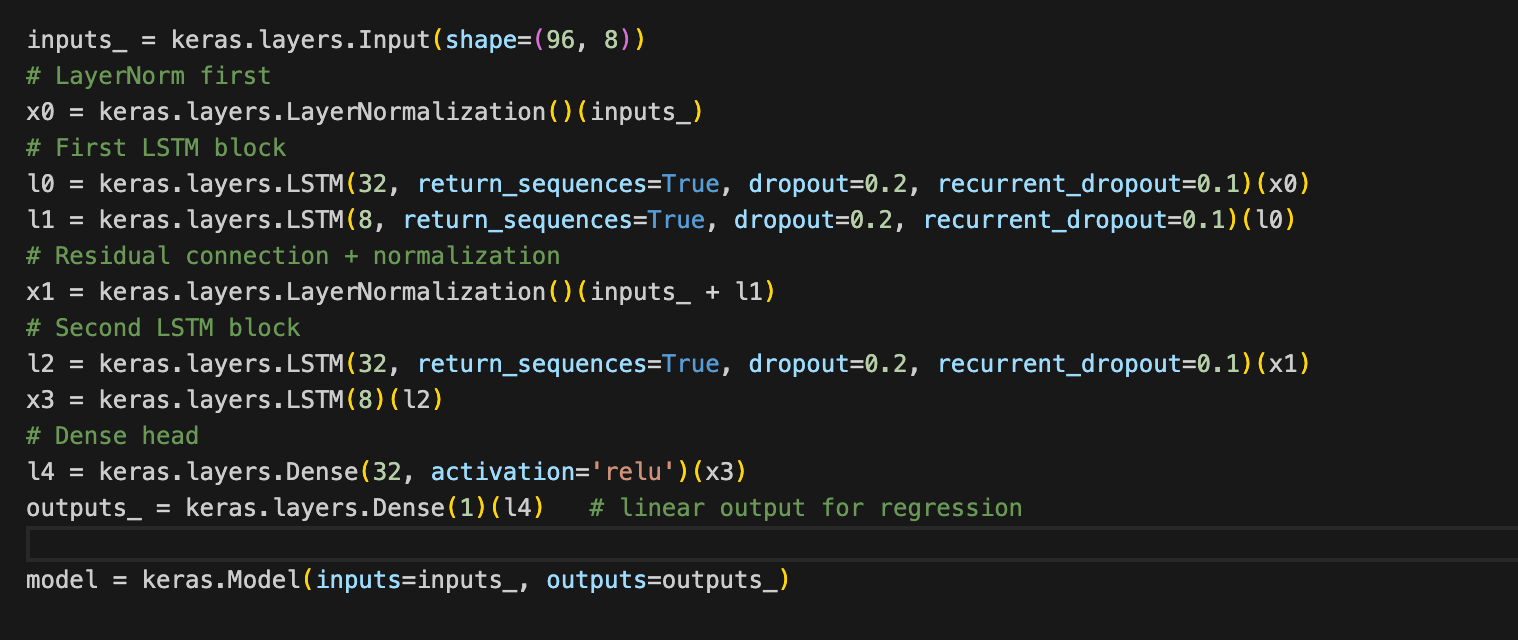
Trade Count：2997 /347 day；

MaxDrawdown：12.2611%

1. RNN/LSTM

Recurrent neural network could be viewed as one of the most popular and classic Deeping Learning models in time series related tasks region. Although now we know from Google that “Attention is All You Need”,but in some particular senarios we can still use RNN.And what’s more, in practice, I often use LSTM,but I simply use RNN to refer it.

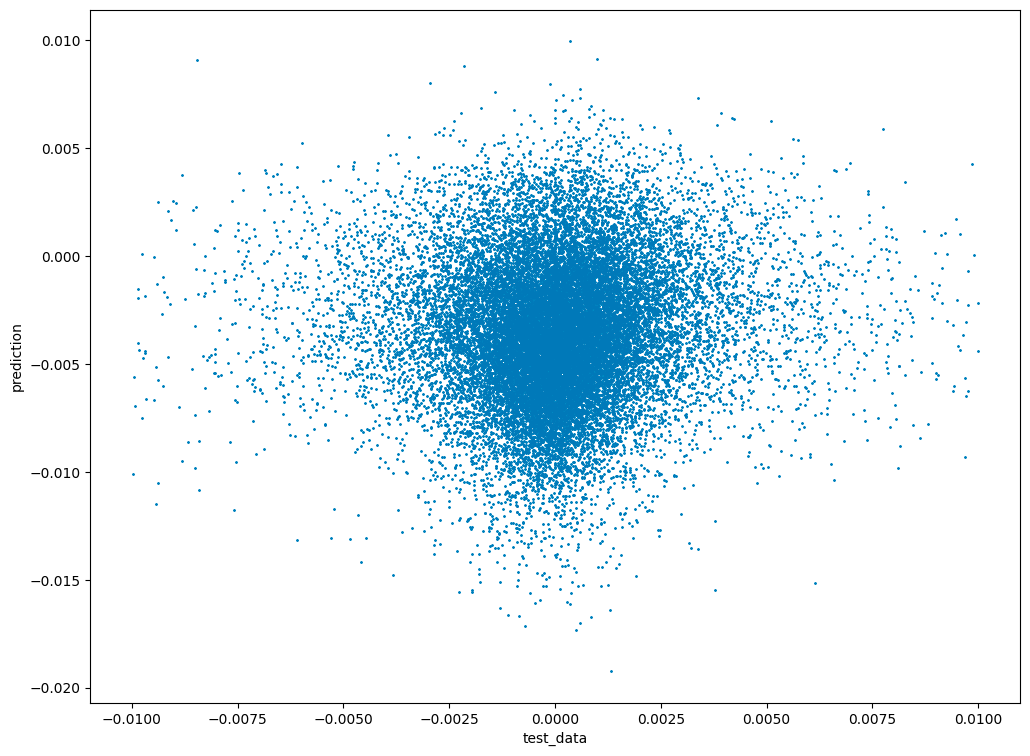
A tpyical RNN structure:



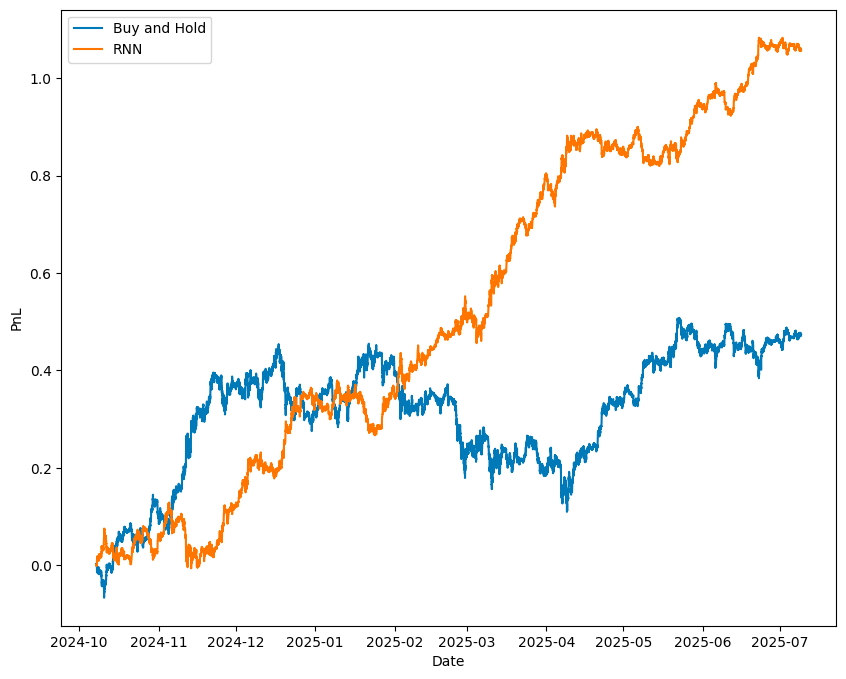
Besides LSTM units and FFN ,I also used LayerNormalization and skip connection which are pretty standard procedures.

Performance on testing data:

IC=0.056



Simple Backtesting:

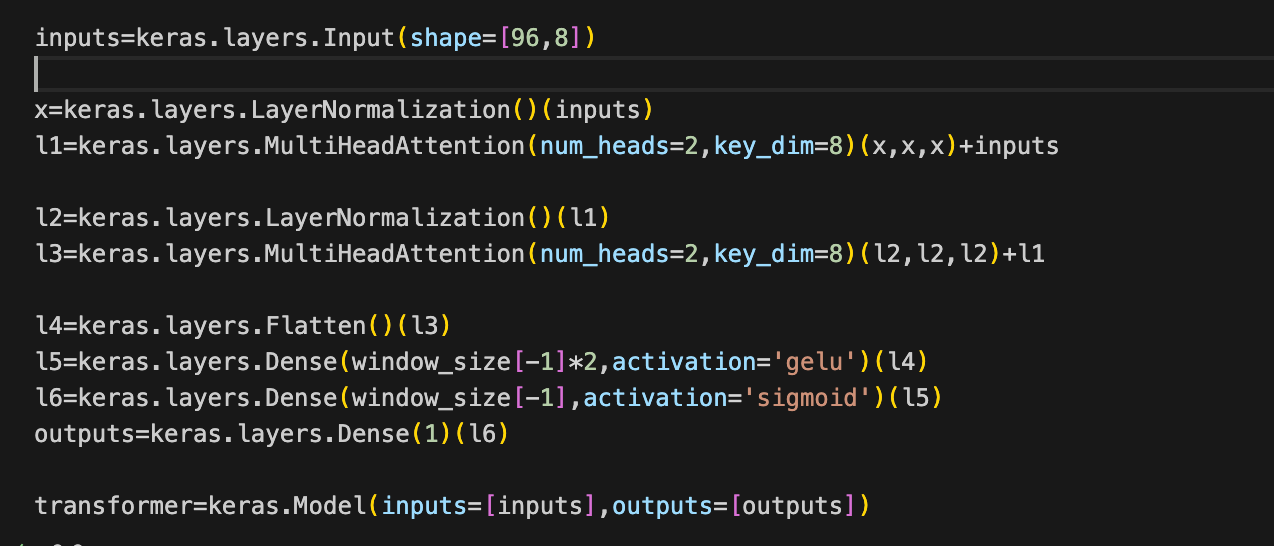


1. Transformer

Transformer now stands at the core of deep learning. With the attention mechanism,it can handle long sequences effectively and efficiently.

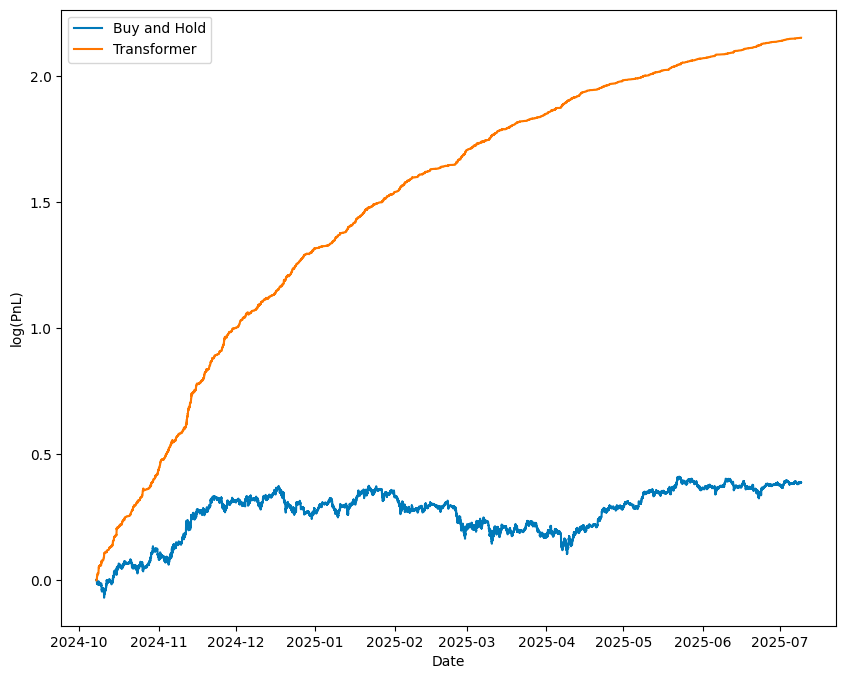
We can use a transformer to predict next term’s return rate.

A simple transformer architecture with two self-attention layers:



And again,we use layer normalization and skip connection as standard procedures.

A simple backtesting shows a really strong performance:

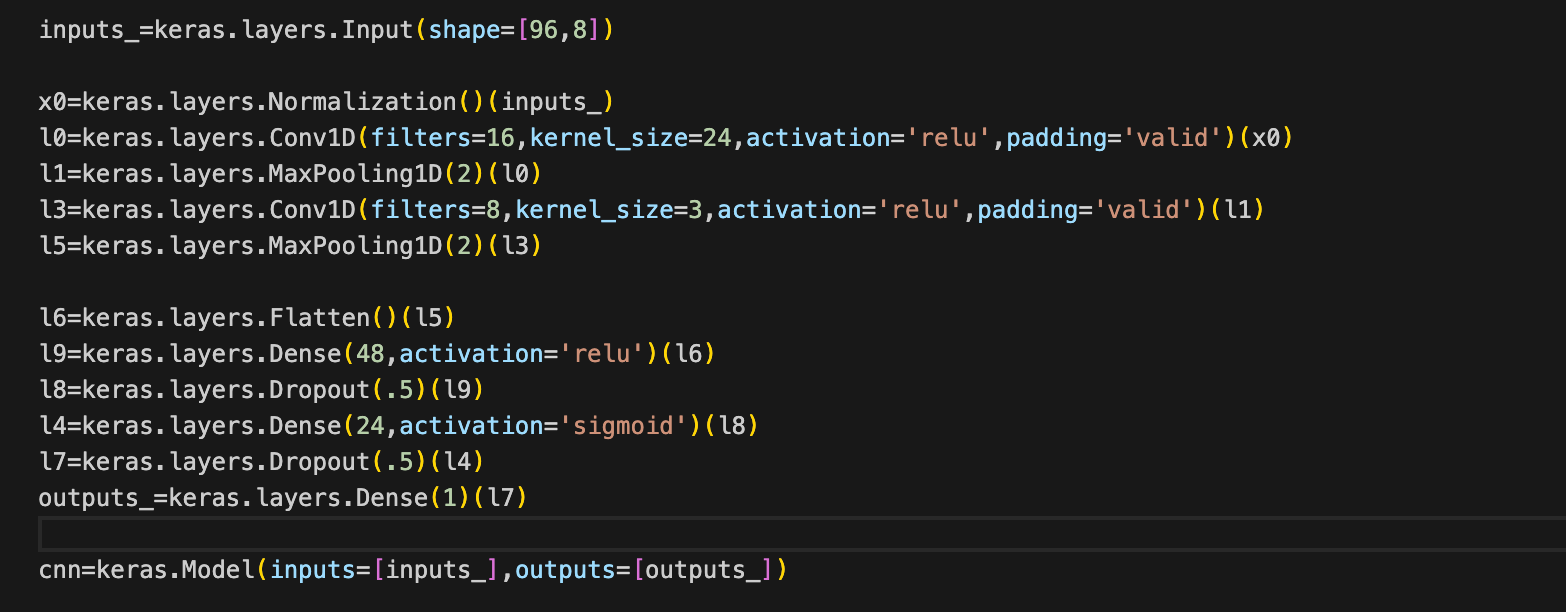


1. CNN

Traditionally,CNN is designed to computer vision problems such as ImageNet.

CNN architecture:

Conv1D-MaxPooling1D-Conv1D-Conv1D-MaxPooling1D-FFN





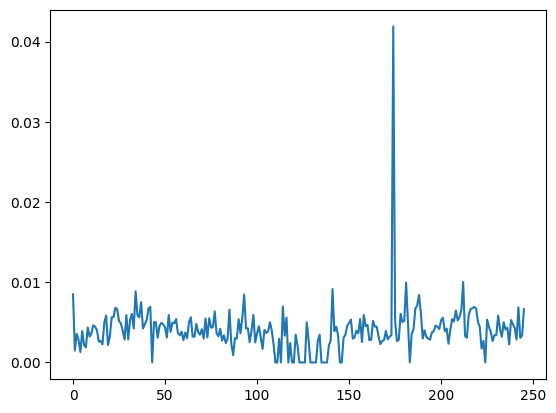
1. All DeepLearning Structures Are Doing Feature Engineering

Feature Engineering

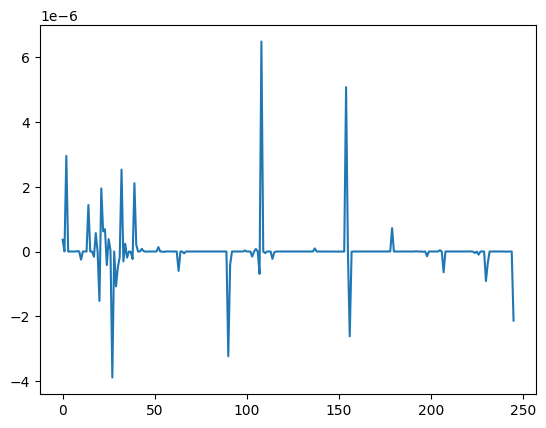
Imputation

Feature transformation

Feature selection



lasso.coef\_（alpha=1）:



1. Reinforcement Learning

I was trying to implement an auto alpha mining alogrithm by RL, but for the limited time and resources this work didn’t progress too much but I’m still interested in it and learning related subjects and hoping could implement it someday!

# Chapter Three：Time Series

1. ARIMA

ARIMA stands for Autoregressive Integrated Moving Average.It has three components.

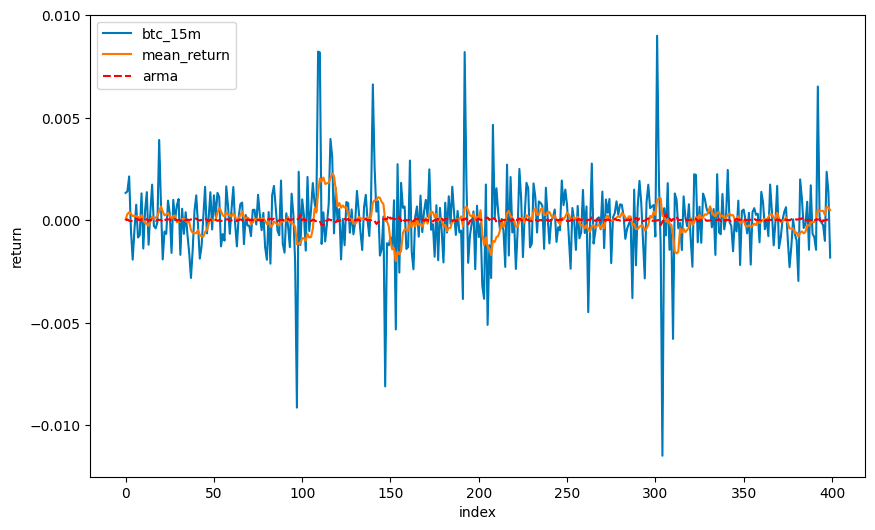
First we take a look into AR(p) model,which does regression to the past p terms and with a noise term:

where B is the Backshift operator.

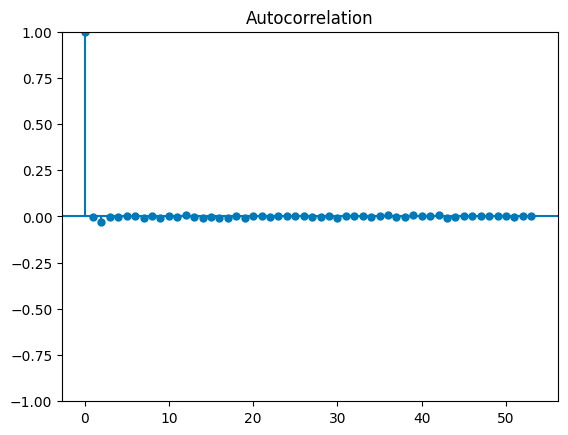
Second, we can take into consideration of more noise terms :

Together,we have a ARMA(p,q) model:

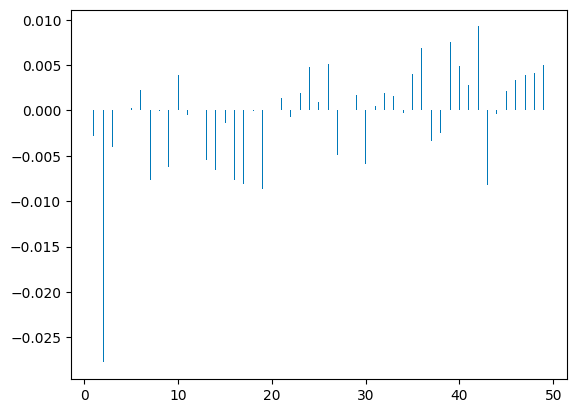
For a return rate series of an asset,I show a ARMA(5,2) model:



Well,we can see ARMA itself alone doesn’t perform well.It because financial data often have low signals-to-noises ratio which makes their autocorrelations are really low and ARMA can not extract information effectively.



We drop the first one to see the autocorrelations more clearly:

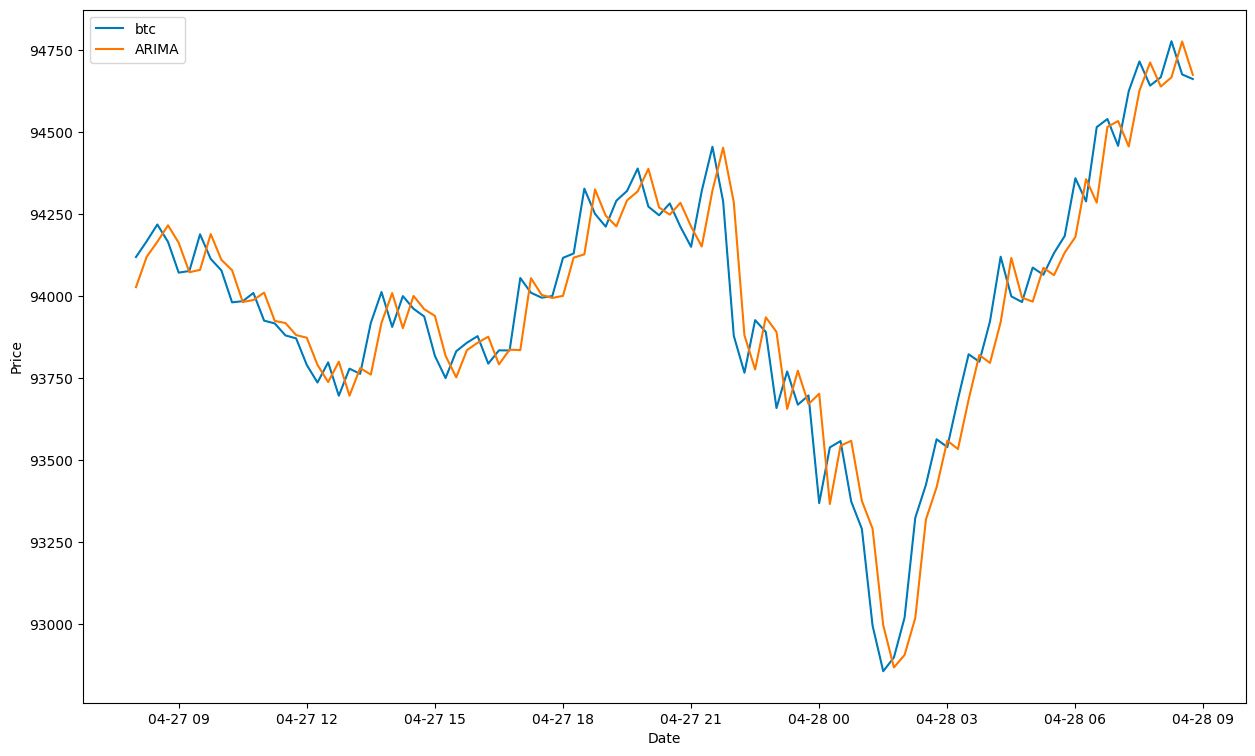


What’s more,if we want to apply the ARMA model to non-stationary series,we can add integrated terms:

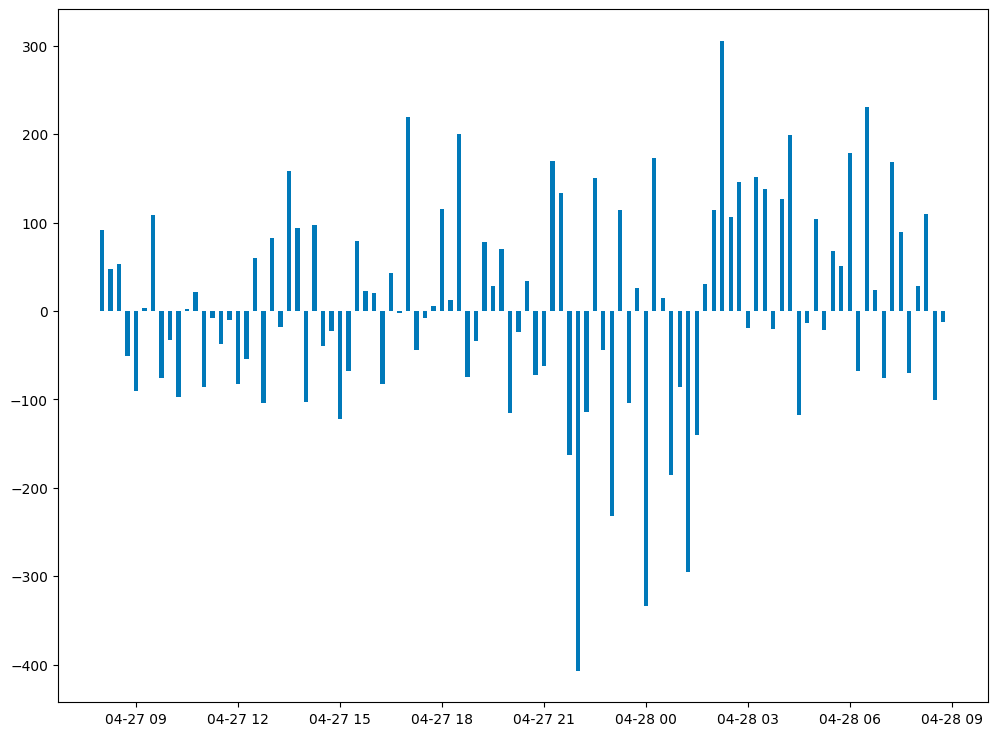
Finally,put all three parts together we have the formal expression of ARIMA:

where we use backshift operator to replace difference operator.

An example period of ARIMA(5,1,1) to fit asset’s price:



the residual of ARIMA is :



1. GARCH

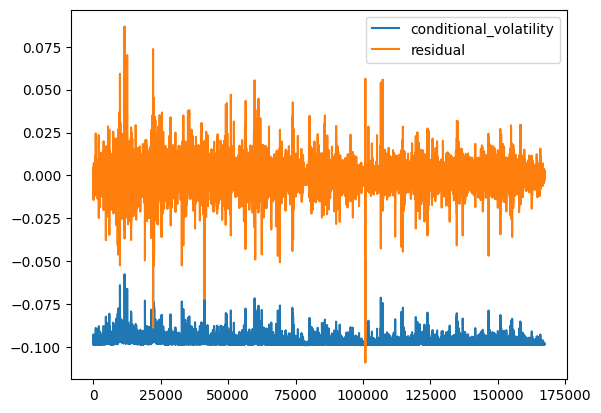
GARCH (Generalised Atuoregressive Conditional Heteroskedastic) model does the pretty same thing ARIMA does,but instead of doing regression to the series itself ,it doing regression to the variance of the series.

Time series :

with white noise and :

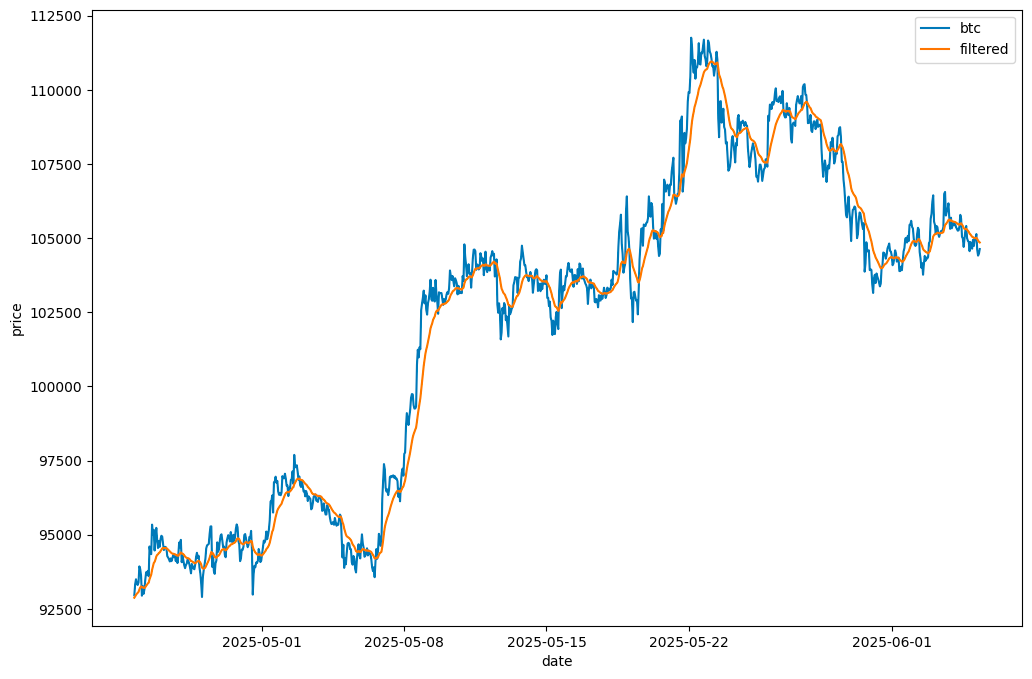
As we know ,ARIMA model can not reflect the heteroskedasticity of a time series.So after fitting by ARIMA,we can apply GARCH to the residual to model the variance.

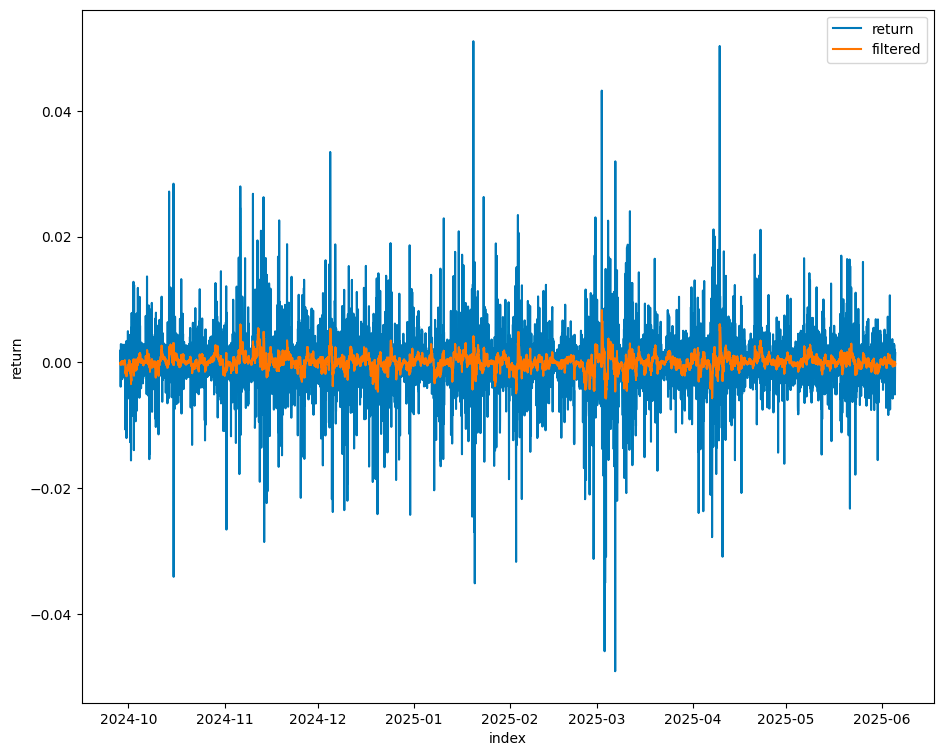
Here we use GARCH(1,1) to fit the residual,and it can fit some of the heteroskedasticity:



1. Kalman Filter

相对而言，经过卡尔曼滤波后的收益率更能反应市场真实状态；可以预估在短期市场中效果更好；使用卡尔曼滤波对收益率序列去噪：



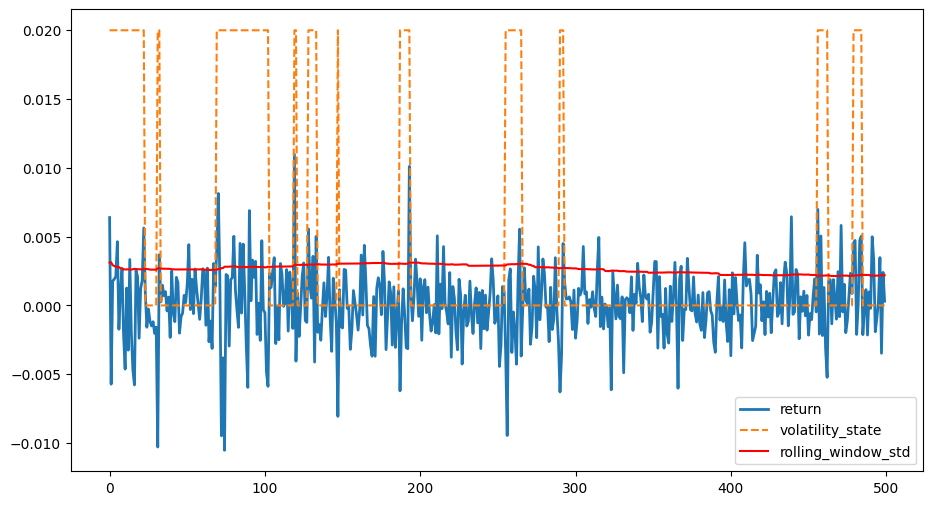


Kalman filter has various applications. We can use it

1. Hidden Markov Model

btc 收益率序列是异方差的，HMM可以有效的检验波动率状态；

假设市场有两种状态，波动率低与波动率高；



使用HMM能够快速识别出市场状态，远比使用std敏感；

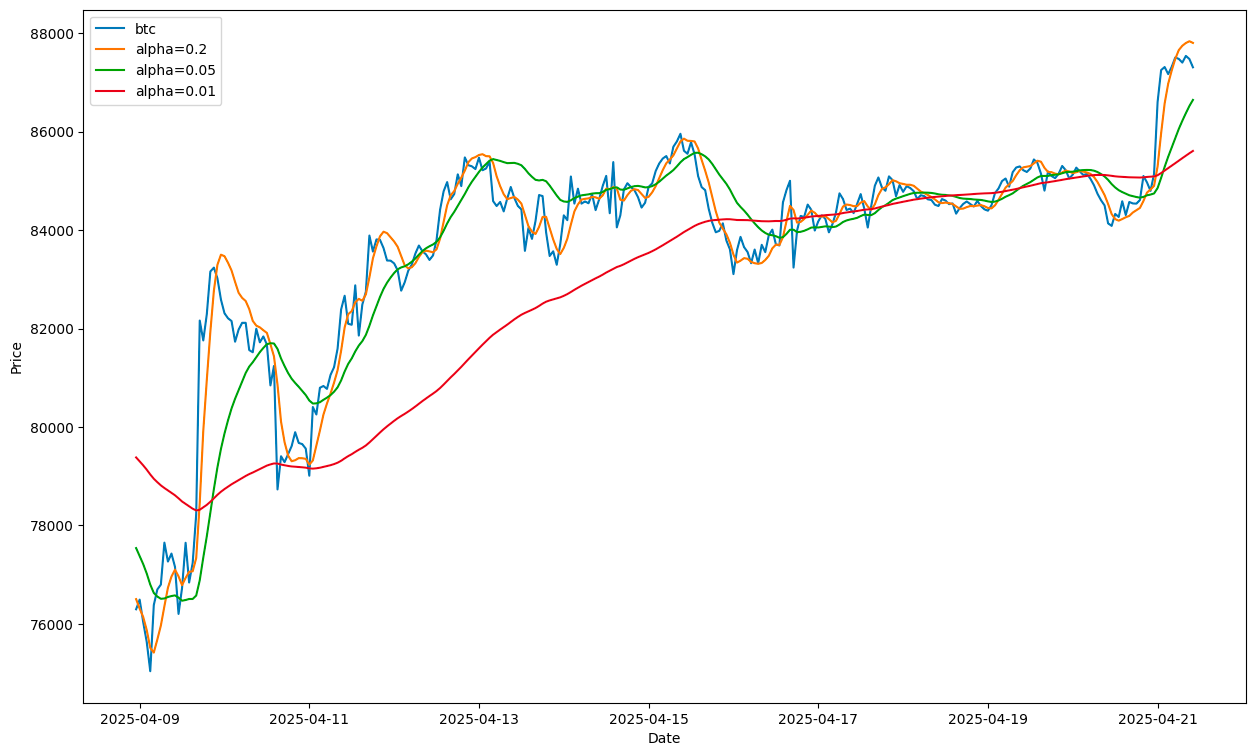
1. Signals Processing and Filters

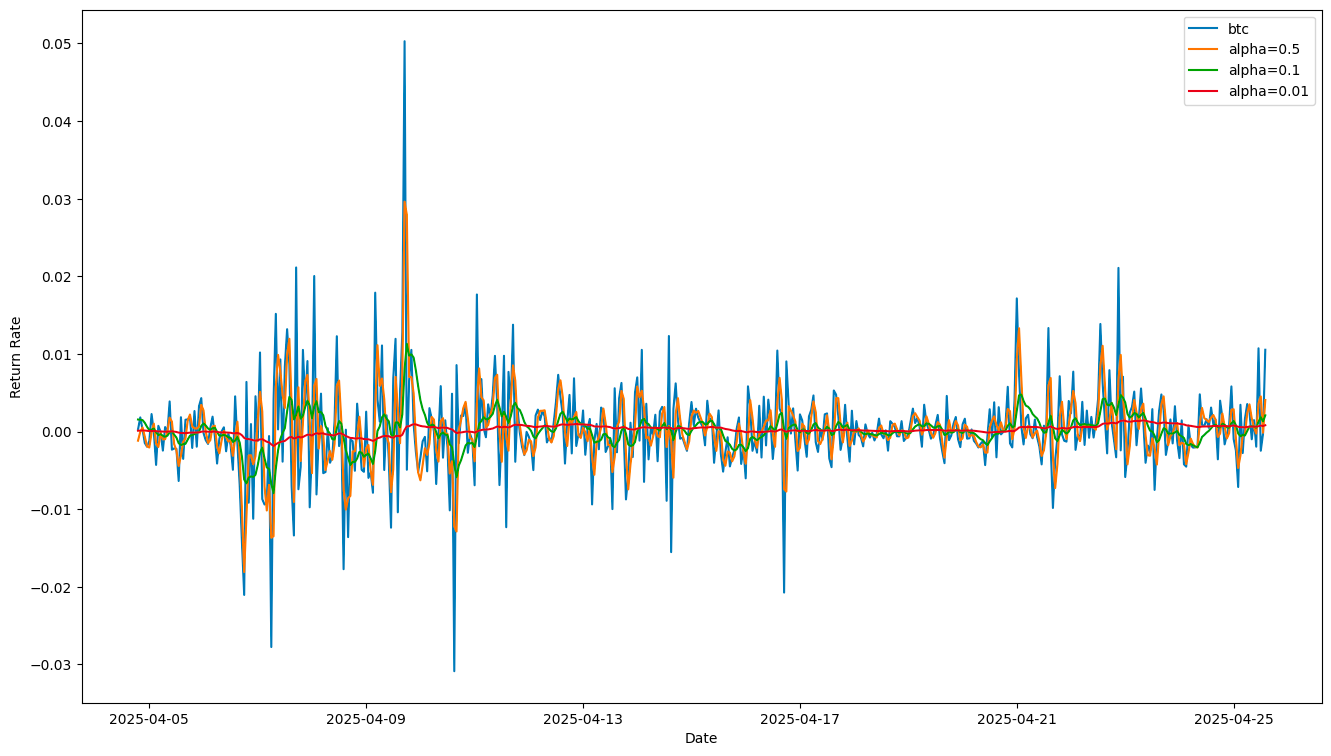
FFT

Laplace transform

1. transform

Filters





# Chapter Four: Past, Now and Future

1. Statistical Arbitrage, High Frequency and Derivatives

This summer internship is quite short and I could experience only a smart part of this industry.But I’m still curious to learn more and ambitious to do more in this job.

I think the real power in quantitative trading is statistical arbitrage. To this purpose ,multiple instruments and multiple markets will be better. That’s why I’m really interested in derivatives and which is whaht I want to do in future.

And just because the nature of arbitrage, the trading progress is intended to be faster and faster.So I’m also interested in high frequency trading and high frequncy factors.

I wish I could do some research in these fields in future.And for now I’m learning and perfectionizing related skills and knowledge such as martingale, Monte Carlo simulation and so on.

1. Reinforcement Learning

Deep Learning and AI are extremely powerful in large data driven business such as quantum trading.And I think among deep learning subjects, reinforcement learning is especially important.

I’m actively learning RL and implementing them from different levels and subjects--from Atari games to Robots,from auto alphas mining to auto trading .RL is what I hope I could continue to do in future’s work.

1. Others

Quantum Trading has a really large and intellgent community.In the studing progress, I’m benefitted a lot from various github projects. And some of are really useful and I’m trying to comprehend and master them such as Qlib and TradingAgent. I think them will definitely benefit a lot in future.