Zeyuan Pan's Week4 note

Contribute

Collated and cleaned over 20,000 rows of GMRF data from four countries, including Canada, the US, China, and the UK, and also performed a simple analysis of the obtained data.
Wrote a program to crawl the stock data on Yahoo Finance to do the groundwork for further processing of the data and backtesting of the factors
Assist teammates to complete the production of reporting slide

meeting note

the sponsor gave us lots of help about our dataset, We don't need to worry about the stock data mapping things.
And also we have to focuse on the following three parts in the next week
1.reading the whole world dataset
2.extracting the head tag into the combined dataset
3.using the risk subsection to read data

note

Something Interesting:

1. For the data given, it can be simply divided into 3 periods:2006-2016, 2017-2019, 2020-2023

1. 2006-2016: Every factor seems to be stochastic process with 0 mean.
2. 2017-2019: They seem to be stochastic process with a mean > 0
3. 2020-2023: Stochastic Processes with mean < 0
   We can’t tell this pattern is caused by datasets we use or market’s influence. We use the data of US to build these graphs and the data of US seem to be a small sample of GMRF. So it might not be able to represent some typical characteristics of the whole market. So, we want to know if it draws the same conclusion from the whole datasets. If it’s mainly caused by some systematic factors, we think we can do some research on it, for there is a high degree similarity between the pattern of these factors and S&P500.

2. The change of word_count has a very similar pattern with the avg_sent, but other factors increase faster when word_count increases and decrease faster when word_count decrease. It seems that when the number of word increase, they are more likely to delivery something like opinions and comments rather just describe some data or facts. Perhaps we should to go into a bigger picture to see what's the difference between a short report and a longer report.

So, based on these initial analysis, it might be a good start that we can test if there is some causality between these factors and the market. If they do have some relationship, we might be able to use these data to build some models to predict the market.

3. There are some abnormal phenomena. From the lazy price, the negative hits are much more than positive hits. But the graph shows that the positive hits are always more than negative hits. In addition, the change of the difference between positive_hits - negative_hits has the similar pattern with the change of positive_hits and the change of negative_hits themselves.