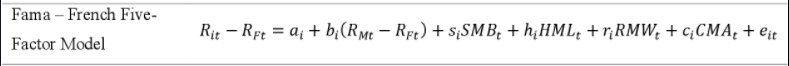
Yau Kwun Fung Kevin [kevinyaukf2000@gmail.com](mailto:kevinyaukf2000@gmail.com) 64803543

**Factor Model by Machine Learning:**

Idea: 

Factor model has long been a widely studied topic in the academia. Previously, researchers used commonly known factors like Market Cap (small minus big), risk premium, and some used profitability and operation efficiency factors like ROE/ ROA/ Sales turnover as explanatory factors.

The factor models have been proved effective and successfully captured inefficiency in the market. However, one limitation on these ratios is that they are reported once every quarter, or even every year. And most researchers used monthly return as testing and rebalancing every month. It is infeasible for hedge funds or quantitative trading as it has a too long holding period, which is very risky. For the more frequently updated ratios, they are mostly based on price and have limited explanatory power as there are much noise in price. The problem is: how to extract common risk exposure for all stocks in a timely manner?

I applied Principal component analysis to reduce the dimension and extract n PCs from the daily returns from the entire universe. The PCs are known as the common factors and systematic returns that every stock is exposed to. Then, I regress every stock on n PCs to get the betas (exposure to different risk) for each stock on each factor. Next, I predict for the expected return on the same day for every stock.

My signal is the difference between the E(R) and the actual return of that day. As I proposed the PCA factor model can provide a more accurate prediction than the market return, there will be a reversion for their difference (residual).

Also, I tested the stationarity on the residual by ADFuller test. And I found that the residual is mean-reverting and proves my hypothesis that there should be a reversion.

Some minor tweaks on the final signal:

I standardized both the predicted values of PCA+ regression and actual returns by Z-score to ensure they are comparable. And applied Tanh function to the Z-score to crop the outliers. Then, the signal will be computed as the difference:

**signal=(np.tanh(Z\_score(pred\_df.T))-np.tanh(Z\_score(ret\_df.T)))**

**🡪**

The signal will then be standardized by the mean, to form the weight for each stock. The weight will sum to 0 and absolute weight will sum to 1. It is a long-short strategy.

Also, I applied an exponential moving average on the signal to lower the turnover and reduce the transaction cost. As I am shorting the returns, it is very noisy. Therefore, taking a moving average can keep a similar return while lowering turnover.

On the backtest, I will use rolling 500 days as the training period, and the first 501-1000 days as in sample period to pick the parameters. Then I will show the summary statistics on the best model for the entire period. Statistics used are Sharpe, Skew, turnover and Average returns.

In sample period testing with No Moving Average:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of PCA factors | Sharpe | skew | turnover | ER |
| 3 | 1.397668 | -0.7589 | 1.400701 | 7.229019 |
| 5 | 1.746541 | -0.27004 | 1.417661 | 7.893705 |
| 7 | 2.092896 | 0.049456 | 1.423763 | 8.586081 |
| 10 | 2.526764 | 0.111546 | 1.431289 | 9.55812 |
| 15 | 2.653406 | 0.195944 | 1.437809 | 9.265252 |
| 20 | 2.096305 | 0.121167 | 1.446881 | 7.030289 |
| 25 | 2.410565 | 0.304187 | 1.452092 | 7.565163 |
| 30 | 2.849874 | 0.280868 | 1.453113 | 8.415704 |
| 40 | 3.390304 | 0.285854 | 1.459686 | 9.355346 |
| 50 | 3.022989 | 0.064959 | 1.464081 | 7.72458 |
| 75 | 1.965906 | 0.057674 | 1.469377 | 4.565969 |
| 100 | 2.147696 | 0.149195 | 1.477541 | 4.454622 |

In sample period testing with EMA1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of PCA factors | Sharpe | skew | turnover | ER |
| 3 | 0.976591 | -0.56685 | 0.576747 | 2.933792 |
| 5 | 1.406048 | -0.35416 | 0.581484 | 3.631588 |
| 7 | 1.49435 | -0.12132 | 0.583411 | 3.585872 |
| 10 | 1.881617 | 0.050267 | 0.585571 | 4.174098 |
| 15 | 1.900091 | -0.08329 | 0.587645 | 3.926918 |
| 20 | 1.51937 | -0.26363 | 0.589691 | 3.001822 |
| 25 | 1.792595 | -0.06186 | 0.59151 | 3.241419 |
| 30 | 2.42664 | 0.040359 | 0.592872 | 4.159529 |
| 40 | 2.832216 | 0.220308 | 0.594685 | 4.466832 |
| 50 | 2.71401 | 0.025519 | 0.5959 | 3.997756 |
| 75 | 2.011085 | 0.095507 | 0.598196 | 2.674619 |
| 100 | 2.291174 | 0.130088 | 0.60051 | 2.804452 |

🡪 In terms of the Sharpe, Skew, and expected return, I will choose 40 PCs with EMA1 on the final signal. Now I will show the results of the entire period.

Chart, line chart

Description automatically generated

Entire period:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PCA | Sharpe | skew | turnover | ER |
| 40 | 2.2403 | 0.5787 | 0.5929 | 3.4999 |