Fund Performance Attribution and Prediction

Lian Weihang 1501213456

Background:

In China, when a public fund is issued. The asset management company must give the target benchmark of the fund. However, for some reasons such as the market quotation and the fund manager's personal factors, the performance of active funds are not stable. Sometimes, the trader styles of active funds are not corresponding to benchmarks. Therefore, I think some machine learning methods can be used to study the performance attributions. Also, we can test whether the style of an active fund is stable or not and make predictions on the future performance.

Data: Time series data of net value

Benchmark index:

- HS300 Index (000300.SH)
- CSI500 Index (399905.SZ)
- CSI1000 Index (000852.SH)
- CSI Universal Bond Index (H11009.CSI)

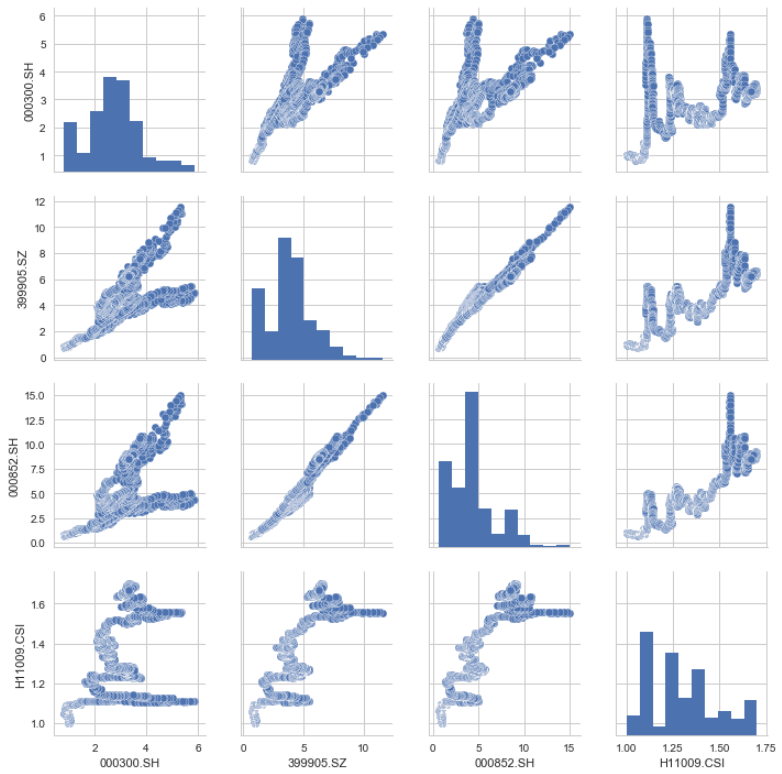
Target Active Funds (with a history of more than 10 years):

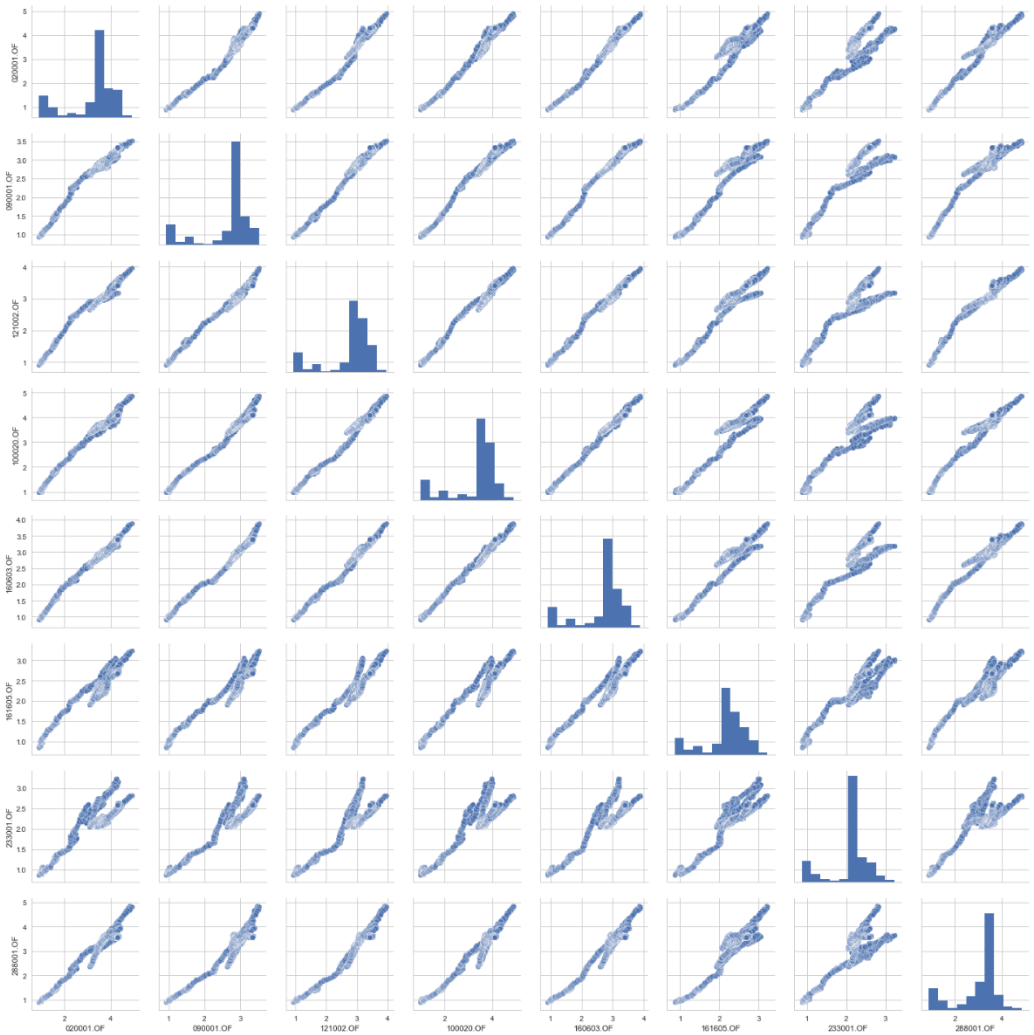
- 020001.OF (80%*000300.SH+20%*H11009.CSI)
- 090001.OF (80%*000300.SH+20%*H11009.CSI)
- 100020.OF (95%*000300.SH+5%*H11009.CSI)
- 121002.OF (75%*000300.SH+20%*H11009.CSI+5%*Risk-free rate)
- 160603.OF (70%*000300.SH+30%*H11009.CSI)
- 161605.OF (75%*000300.SH+25%*H11009.CSI)
- 233001.OF (0.55*000300.SH+45%*H11009.CSI)
- 288001.OF (0.6*000300.SH+20%*H11009.CSI+20%* Risk-free rate)

Method (Steps):

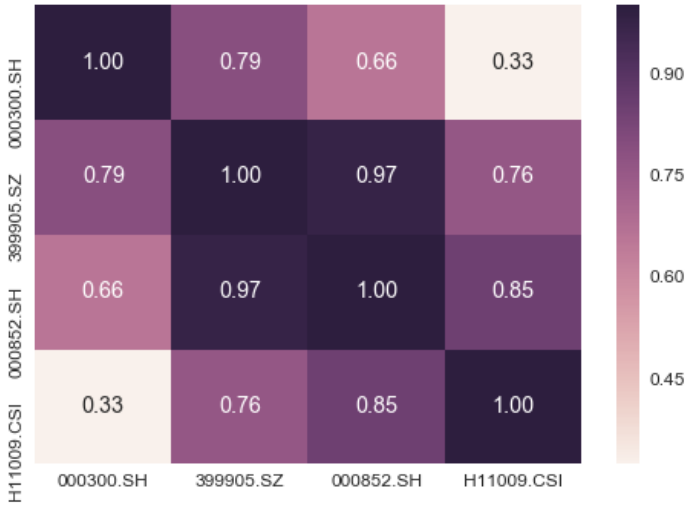
1. Preprocess the time series data. Intercept the data from 2004-12-31 to 2016-12-31 and divide them into two parts. The first several years' (from 2005 to 2015) data will be used as sample data and the last year's data will be used as lab data (testing data). Then, divide each data array by the first value of the array such that all the first values = 1.

2. Respectively create a scatterplot matrix that allows to visualize the pair-wise correlations between the benchmark indexes and between target active funds. It can be seen that in general, both benchmark indexes and target active funds are positively related with each other.


000300.SH
399905.SZ
000852.SH
H11009.CSI


020001.OF
090001.OF
121002.OF
100020.OF
160603.OF
161605.OF
233001.OF
288001.OF

3. Respectively use seaborn's heatmap function to plot the correlation matrix array for both benchmark indexes and active funds as follows:



4. Do principle component analysis(PCA) and cross-validation(CV) on each active fund and calculate the variance explained ratio corresponding to the number of components. Obviously, when $\text{n\_components} = 2$, the first 2 eigenvectors can contribute more than 99% of the whole variance in all cases of active funds. However, the PCA scores and CV scores are all negative, so the attempt to compress data with PCA is failed.

5. Respectively do ordinary least square(OLS) regression on each active fund of benchmark indexes as follows:

$$y = X\beta + \epsilon = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,$$

where α represents the weight of cash, β_i represents the coefficient (weight) of a benchmark index.

According to the attribution rule, $\alpha + \beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$, so the regression model can be written as:

$$y - 1 = \beta_1(x_1 - 1) + \beta_2(x_2 - 1) + \beta_3(x_3 - 1) + \beta_4(x_4 - 1).$$

Then set $(y - 1)$ and $(x_i - 1)$ as new variables to replace the original ones. In the Python program, the parameter fit_intercept of LinearRegression() should be set as False.

Just like the information mentioned above, the data before 2016 is regarded as the training set and data of 2016 is the test set. In this regression, the training set begins from 2008 for a better effect. Get coefficients from the regression on training data and predict on test data. Calculate the mean squared error and R square. The result is not that significant. Then, calculate a series of statistical data including mean, standard deviation, maximum and minimum of both predicted and actual data. Define

$$\text{bias ratio} = \frac{\text{pred}-\text{test}}{\text{test}} \times 100\%.$$

It can be seen that most bias ratio values of mean, maximum and minimum are less than 5% while bias ratios of standard deviation are much larger, which explains the insignificance of the results. Then, define predicted accuracy as follows: If bias ratio on a trading day is less than 5%, then mark the prediction as an "accurate" prediction,

$$\text{predicted accuracy} = \frac{\text{num of "accurate" predictions}}{\text{num of trading days}}$$

It can be seen in the ipython notebook that 5 out of 8 predicted accuracy results is larger than 80%, 2 out of 8 results are between 50% and 80%. Only the last active fund has a relatively inaccurate prediction.

6. Respectively fit a robust regression model using RANSAC and similarly repeat the procedure in Step 5. The training set starts from 2007 for a better effect and the result is displayed in ipython notebook.

Conclusion:

This paper gives out two ways (models in Step 5 and 6) of performance attribution and prediction. The models can help us to predict the trend of an active fund according to benchmark indexes and test whether the style is stable or not. Furthermore, we can try to replicate the active fund with just index fund and save an administration cost.