Project: Tracking Investment Styles of Public Funds
EY(China)

Mengjin ZHAO\textsuperscript{1}, Xingxin YIN\textsuperscript{1}, Zichao ZHANG\textsuperscript{1}, Zixuan WANG\textsuperscript{1}, Chen LI\textsuperscript{1}, Shan LI\textsuperscript{1}, Jingnan JU\textsuperscript{1}, Shuang GU\textsuperscript{2}, Yufeng SHI\textsuperscript{3}, Lei PAN\textsuperscript{3}, Huaxiong HUANG\textsuperscript{4}, Yiqing LIN\textsuperscript{4}, Shige PENG\textsuperscript{4}

\textsuperscript{1}Students \textsuperscript{2}EY(China) \textsuperscript{3}Professors \textsuperscript{4}Special Contributors

2019/05/31, the Fields Institute, Toronto
1 Data & Deliverables: In Retrospect
   - Data
   - Aims: A Short Introduction

2 Introductions of our Approaches: Attempts & Results
   - Methods on Processing Changing Points
   - Statistic Tuple Scheme
   - Wasserstein Metric Scheme
   - Fourier/Wavelet Analysis Scheme
   - ML Scheme
   - Regression Method on Predicting the Stock Total Position

3 Summary & Outlook
Data

Data (2014-2018) will be provided for this project:

- Reference Indices (Daily): 15 Style Indices / 28 Industrial Indices / 8 Market Indices
- Net value data of funds (Daily): NAV and adjusted NAV data from 300 public funds
- Periodic reports (Quarterly): Report of public funds including total position of stocks and detailed shareholding (top 10 in 1st and 3rd quarter, full list in 2nd and 4th quarter)
- Bond yields (Daily): Shibor and China 1-Year Bond Yield data.
Aims: A Short Introduction

The deliverables of this project are:

1) Use algorithm to track investment styles of funds during a quarter.

2) Identify if funds change their target industries during a quarter.
   - Statistics characteristics (high-order moments/correlation analysis)
   - Wasserstein distance of distributions
   - Fourier/Wavelets techniques
   - Detection of special changing points

3) Predict fund stock ratio at the end of each quarter.
   - ML Approach
   - Regression Approach
Changing Point

- Basic idea: When the funds’ managers change their positions on stocks, regime switching may happen at that time. We need to find a suitable length of the window to process it, in order to get accurate results of the coming approaches.

- Problem we faced: Actually, daily returns of NAV are not normal distributed, especially when the sample size is large. When the sample size is small, change-point detection methods are not so accurate.
Statistic Tuple Scheme

- Idea: The similarity between a fund and an industrial/style index gives a statistical similarity.
- Consider normalized high moments \((\mu, \sigma^2, s, k)\) of empirical distribution.
- If the characteristic tuple \((\mu, \sigma^2, s, k)\) of a fund is close enough to some index (industrial/style), we can conclude that the fund in this period is tracking the style or targeting the industrial index indicated by the tuple.
- As a result, during the sliding band days, the empirical tuples of funds and indices with their distance which we can calculate can identify the style of the funds.
Statistic Tuple Scheme: A Figure
Here in this period, Negative-Earning Stock is Dominating. The fund is tend to behave like Negative-Earning Index.

Here the power of Small-Earning Index is shrinking.
Wasserstein Metric Scheme

- Idea: Measure the distance between distributions straightly rather than correlations and moment statistics.
- Wasserstein metric can be a good idea. It is a distance function defined between probability distributions on a given metric space.
- The Wasserstein distance between two probability measures $\mu$ and $\nu$ is defined as

$$W(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R} \times \mathbb{R}} |x - y| \, d\gamma(x, y)$$

where $\Gamma(\mu, \nu)$ denotes the collection of all measures on $\mathbb{R} \times \mathbb{R}$ with marginals $\mu$ and $\nu$ on the first and second factors respectively.
Wasserstein Distance between funds and CAP Styles with Regime Switching
Fourier/Wavelet Analysis Scheme

- Here we consider the spectrum as a characteristic.
- By Fourier transformation, we can get the spectrum as the characteristic,
  \[ F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} dt \]
- For wavelet analysis, the spectrum is parameterized by \((s, \tau)\),
  \[ C(s, \tau) = \int_{-\infty}^{\infty} f(x)\Psi_{s,\tau}(x) dx \]
  with the \(\tau\)-shifted and \(s\)-scaled kernal
  \[ \Psi_{s,\tau}(x) = \frac{1}{\sqrt{s}}\Psi\left(\frac{x - \tau}{s}\right). \]
Figure: FFT does not work well as we expected.
Figure: 2-D Spectrum of Wavelet Analysis
Accuracy of Wavelet on Predicting Industries: 19.7%
ML Scheme: Description I

- We assume that the fund’s stork ratio ($Y$ - output) is related to the market, related to the major investment industry and the historical data of the fund itself ($X$ - input).

- 36-dim Input Variables: The mean, variance, skewness and kurtosis ($\mu, \sigma^2, s, k$) of
  
  - the funds’ return sub the return of the market for the past 30/20/10 days (30/20/10 F_C)
  - the industries return sub the return of the market for the past 30/20/10 days (30/20/10 I_C)
  - the funds return sub the return of the industry for the past 30/20/10 days (30/20/10 F_I)
ML Scheme: Description II

- Taking the reporting-day data, we get 6000 samples (300 funds × 20 reporting days). Foregoing 4800 for training, and the remanent for testing.
- Models: Random Forest & XGboost
ML Results: Random Forest

<table>
<thead>
<tr>
<th>Variable</th>
<th>Importance</th>
<th>Variable</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>skew30F_I</td>
<td>0.04</td>
<td>skew30F_I</td>
<td>0.05</td>
</tr>
<tr>
<td>kurt30F_C</td>
<td>0.04</td>
<td>mean30F_I</td>
<td>0.04</td>
</tr>
<tr>
<td>skew20F_I</td>
<td>0.04</td>
<td>skew10F_C</td>
<td>0.04</td>
</tr>
<tr>
<td>std20F_C</td>
<td>0.04</td>
<td>kurt10F_C</td>
<td>0.04</td>
</tr>
<tr>
<td>skew10F_I</td>
<td>0.04</td>
<td>std30F_I</td>
<td>0.03</td>
</tr>
<tr>
<td>kurt10F_I</td>
<td>0.04</td>
<td>kurt30F_I</td>
<td>0.03</td>
</tr>
<tr>
<td>mean10F_C</td>
<td>0.04</td>
<td>mean30F_C</td>
<td>0.03</td>
</tr>
<tr>
<td>std10F_C</td>
<td>0.04</td>
<td>std30F_C</td>
<td>0.03</td>
</tr>
<tr>
<td>skew10F_C</td>
<td>0.04</td>
<td>skew30F_C</td>
<td>0.03</td>
</tr>
<tr>
<td>kurt10F_C</td>
<td>0.04</td>
<td>kurt30F_C</td>
<td>0.03</td>
</tr>
</tbody>
</table>
ML Results: XGboost - Good Prediction and Right Trend
ML Results: XGboost - Good Prediction but opposite Trend
ML Results: XGboost - not-so-Good Prediction and bad Trend
Regression Method on Predicting the Stock Total Position

- The regression equation is shown below:

\[ R_{f,t} = \sum_i \gamma_i R_{i,t} + \epsilon \]

where \( R_{f,t} \) is the daily return of some fund \( f \) at time \( t \), \( R_{i,t} \) is the daily rate of return of the INDUSTRIAL INDICES at time \( t \), \( \gamma_i \) is the regression coefficient to be fitted, and the \( \epsilon \) is the residual item.

- We believe that \( \gamma_i \) represents the proportion of stocks invested by the fund in industry \( i \), so \( \sum_i \gamma_i \) is the proportion of stocks held by the fund in the total assets.
Regression Method on Predicting the Stock Total Position

In order to avoid the influence of collinearity, Lasso regression is adopted. The loss function of Lasso regression is:

$$J(\gamma) = ||Y - \gamma^T X||_2^2 + \lambda ||\gamma||_1$$

The coefficients of some industries can sometimes be compressed into 0, deriving a set of ”best regression effect” of industry group as explaining variables.
Accuracy of Lasso on Predicting Stock Ratio: 13.1%
Data provided is not digged much deeply. Approaches are relied more on closed values rather than other values such as open, high, low prices.

More methods and characteristics in ML fields can be taken into consideration.

The method on implied risk aversion function derived from a Mean-Variance framework has not been realized.

Statistics / p.d.f. / Spectrum methods can be treated more carefully. High-order comoments / metrics / wavelets can be considered.
Thanks/FAQ

Thanks for listening and questions are welcome!