Guided Tour of Machine Learning in Finance

Week 3-1-2-1: Unsupervised Learning

Principal Component Analysis (PCA)

Igor Halperin

NYU Tandon School of Engineering, 2017

Stock returns

Analyze returns of 30 stocks from the DJI index:

$$r_i(t) = \log\left(\frac{S_i(t)}{S_i(t-\Delta t)}\right), \quad i = 1,\ldots,N_{stocks}$$

$$r_i(t) = \alpha_i + \beta_i r_M(t) + \varepsilon_{it}$$

Here:

$r_i(t)$: the log-return for stock i

$r_M(t)$: the log-return for the "market" M (the DJI index)

β_i : the "beta" - measure of correlation with the market

ε_{it} : the "noise" term (residual) $\Leftrightarrow \mathbb{E}[\varepsilon_{it}] = 0$

Δt : the time step (1m, 1h, 1d, 1w, 1m, 1y, etc.)

This is the regression equation of the Capital Asset Pricing Model (CAPM, see e.g. Cochrane, "Asset Pricing" 2005).

Eigen-portfolios as a UL taks

Trading signals for statistical arbitrage from the ML perspective

In statistical arbitrage strategies:

- Trading signals are systematic or rule-based
- Trading book has hero beta with the market (i.e. it is market-neutral)
- Use diversification to produce low volatility portfolios uncorrelated with the market

$$\hat{r}_i(t) = \hat{\alpha}_i + \hat{\beta}_i\, r_M(t)$$

$$\hat{\varepsilon}_i(t) = r_i(t) - \hat{r}_i(t) = r_i(t) - \hat{\alpha}_i - \hat{\beta}_i\, r_M(t)$$

Eigen-portfolios as a UL taks

Trading signals for statistical arbitrage from the ML perspective

In statistical arbitrage strategies:

- Trading signals are systematic or rule-based
- Trading book has hero beta with the market (i.e. it is market-neutral)
- Use diversification to produce low volatility portfolios uncorrelated with the market
  - See Avellaneda and Lee (2008) for for details
  - Multi-factor model: individual returns $R_i \ (i = 1,..,N)$ are driven by m factors $F_j \ (j = 1,..,m)$ and idiosyncratic components $\tilde{R}_i$ with $\mathbb{E}\left[\tilde{R}_i\right] = 0$

$$R_i = \alpha_i + \sum_{j=1}^{m} \beta_{ij} F_j + \tilde{R}_i$$

Eigen-portfolios as a UL taks

Trading signals for statistical arbitrage from the ML perspective

In statistical arbitrage strategies:

- Trading signals are systematic or rule-based
- Trading book has hero beta with the market (i.e. it is market-neutral)
- Use diversification to produce low volatility portfolios uncorrelated with the market
- See Avellaneda and Lee (2008) for for details
- Multi-factor model: individual returns $R_i \ (i = 1,..,N)$ are driven by m factors $F_j \ (j = 1,..,m)$ and idiosyncratic components $\tilde{R}_i$ with $\mathbb{E}\left[\tilde{R}_i\right] = 0$

$$R_i = \alpha_i + \sum_{j=1}^{m} \beta_{ij} F_j + \tilde{R}_i$$

- Factors F_j can be thought of as returns of "benchmark" portfolios representing **systematic factors** (e.g. industry and geography factors)

Control question

Select all correct answers

1. In the factor model, all individual returns are only correlated with “market” M with correlations β_i, but remain uncorrelated between themselves.
2. Correlation of two assets i and j in the factor model is given by $\beta_i \beta_j$.
3. Statistical arbitrage deals with “factor investing” to find factors F_j with highest returns.
4. Statistical arbitrage use residuals of stock returns “unexplained” by factors to extract trading signals.

Correct answer: 2, 4