

# Cluster-based Strategy

Anran Lu   Huanzhong Xu   Atharva Parulekar

Stanford University

May 8, 2018

## Summary

- Review of classic Statistical Arbitrage

## Summary

- Review of classic Statistical Arbitrage
- Main idea and steps of cluster-based strategy

## Summary

- Review of classic Statistical Arbitrage
- Main idea and steps of cluster-based strategy
- Current result and future work

## Review of Statistical Arbitrage



**Figure:** Stock chart of United Airline, Delta Airline and XTN

## Main Idea of Pairs Trading

**Basic model:**

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t$$

## Main Idea of Pairs Trading

### Basic model:

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t$$

- "Pair": two stocks with similar characteristics



## Main Idea of Pairs Trading

### Basic model:

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t$$

- "Pair": two stocks with similar characteristics
  - $\alpha$ : negligible compared to  $dX_t$

## Main Idea of Pairs Trading

### Basic model:

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t$$

- "Pair": two stocks with similar characteristics
  - $\alpha$ : negligible compared to  $dX_t$
  - $\beta$ : determined by regression

## Main Idea of Pairs Trading

### Basic model:

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t$$

- "Pair": two stocks with similar characteristics
  - $\alpha$ : negligible compared to  $dX_t$
  - $\beta$ : determined by regression
  - $X_t$ : key part of the portfolio

## Main Idea of Pairs Trading

### Basic model:

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t$$

- "Pair": two stocks with similar characteristics
  - $\alpha$ : negligible compared to  $dX_t$
  - $\beta$ : determined by regression
  - $X_t$ : key part of the portfolio
- "Generalized Pair": a stock and an ETF

## Determine Portfolio by $X_t$

**Cointegration:** Ornstein — Uhlenbeck modelling of  $X_t$

$$dX_t = \kappa(\mu - X_t)dt + \sigma dW_t$$

## Determine Portfolio by $X_t$

**Cointegration:** Ornstein — Uhlenbeck modelling of  $X_t$

$$dX_t = \kappa(\mu - X_t)dt + \sigma dW_t$$

- some implicit assumptions

## Determine Portfolio by $X_t$

**Cointegration:** Ornstein — Uhlenbeck modelling of  $X_t$

$$dX_t = \kappa(\mu - X_t)dt + \sigma dW_t$$

- some implicit assumptions
- $\kappa$ : mean reversion speed

## Determine Portfolio by $X_t$

**Cointegration:** Ornstein — Uhlenbeck modelling of  $X_t$

$$dX_t = \kappa(\mu - X_t)dt + \sigma dW_t$$

- some implicit assumptions
- $\kappa$ : mean reversion speed
- $X_t \sim \mathcal{N}(\mu, \frac{\sigma^2}{2\kappa})$  at equilibrium

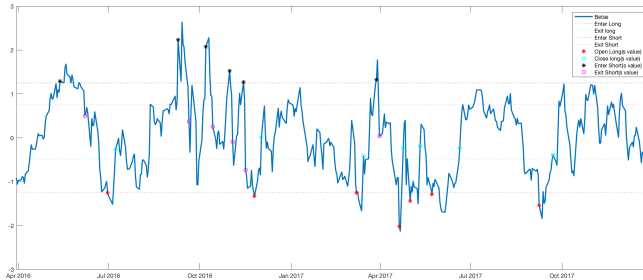


## Determine Portfolio by $X_t$

**Cointegration:** Ornstein — Uhlenbeck modelling of  $X_t$

$$dX_t = \kappa(\mu - X_t)dt + \sigma dW_t$$

- some implicit assumptions
- $\kappa$ : mean reversion speed
- $X_t \sim \mathcal{N}(\mu, \frac{\sigma^2}{2\kappa})$  at equilibrium
- $s = \frac{X_t - \mathbf{E} X_t}{\mathbf{Var} X_t}$  determines long/short position



**Figure:** Apple-XLK pair opening and closing signals



**Figure:** Stock chart of United Airline, Delta Airline and XTN

Main idea and steps of cluster-based strategy

## Motivation of our work

- Pairs with different "qualities"

## Motivation of our work

- Pairs with different "qualities"
- Dynamically adjusting pairs based on real time factors

## Motivation of our work

- Pairs with different "qualities"
- Dynamically adjusting pairs based on real time factors
- Non-stationary time series and cointegration

## Motivation of our work

- Pairs with different "qualities"
- Dynamically adjusting pairs based on real time factors
- Non-stationary time series and cointegration

**Cointegration:** Here  $u_t$  is the stationary residual obtained after regressing non-stationary  $x_t$  and  $y_t$  which also have a unit root. We can use Dicky Fuller or two step Engel Granger to determine cointegration.

$$y_t - \beta x_t - \alpha = u_t$$



## Motivation of our work

- Pairs with different "qualities"
- Dynamically adjusting pairs based on real time factors
- Non-stationary time series and cointegration

**Cointegration:** Here  $u_t$  is the stationary residual obtained after regressing non-stationary  $x_t$  and  $y_t$  which also have a unit root. We can use Dicky Fuller or two step Engel Granger to determine cointegration.

$$y_t - \beta x_t - \alpha = u_t$$

## Steps of our work

**Main idea:** We run pairs trading on each stock-ETF pair giving rise to 5 key factors per stock-ETF pair. We calculate these factors for a stock with respect to all 6 ETFs.

## Steps of our work

**Main idea:** We run pairs trading on each stock-ETF pair giving rise to 5 key factors per stock-ETF pair. We calculate these factors for a stock with respect to all 6 ETFs.

### Key factors for pair:

- Residual  $s_t$ . (Non decorrelated  $s_t = X_t - m$ )
- Mean Reversion Time  $1/\kappa$
- Cointegration factors

## Steps of our work

**Main idea:** We run pairs trading on each stock-ETF pair giving rise to 5 key factors per stock-ETF pair. We calculate these factors for a stock with respect to all 6 ETFs.

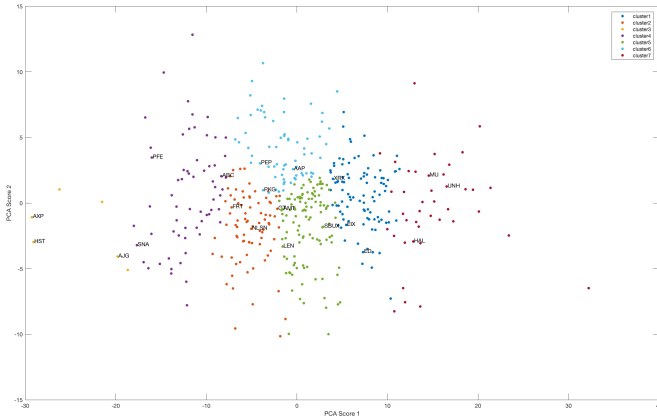
### Key factors for pair:

- Residual  $s_t$ . (Non decorrelated  $s_t = X_t - m$ )
- Mean Reversion Time  $1/\kappa$
- Cointegration factors

Apply K-means clustering based on the key factors for a stock with respect to all ETFs (30 dim vector).

## Current Result and Future Work

# Visualization of Current Result



**Figure:** PCA scores based on K-means clustering (500 stocks)

## Future work

- General performance estimation and efficiency improvement

## Future work

- General performance estimation and efficiency improvement
- Trade-off between cointegration and residual



## Future work

- General performance estimation and efficiency improvement
- Trade-off between cointegration and residual
- Alternative Models

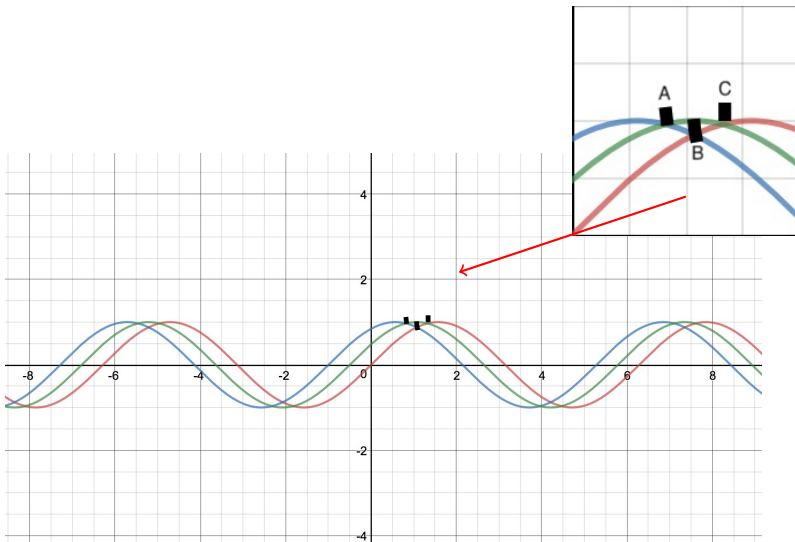
## Future work

- General performance estimation and efficiency improvement
- Trade-off between cointegration and residual
- Alternative Models
  - Mean-reverting models to better describe  $X_t$
  - EM for clustering

## Future work

- General performance estimation and efficiency improvement
- Trade-off between cointegration and residual
- Alternative Models
  - Mean-reverting models to better describe  $X_t$
  - EM for clustering
- Trade-off between trading frequency and transaction cost

# Selection of positions



Thank you!