Applying Machine Learning to Pairs Trading

HUDSON AND THAMES

About Me

- Quantitative Research Team Lead at Hudson & Thames
- M.Sc. in Computer Science and Econometrics at the University of Warsaw

Twitter: @IllyaBarziy

GitHub: @PanPip

LinkedIn: www.linkedin.com/in/illyabarziy/

H&T PRODUCTS

ARBITRAGELAB
BY HUDSON & THAMES

Key pairs trading statistical arbitrage algorithms.

MI•FINLAB
BY HUDSON & THAMES

Power of machine learning with interpretable and easy to use tools.

PORTFOLIOLAB
BY HUDSON & THAMES

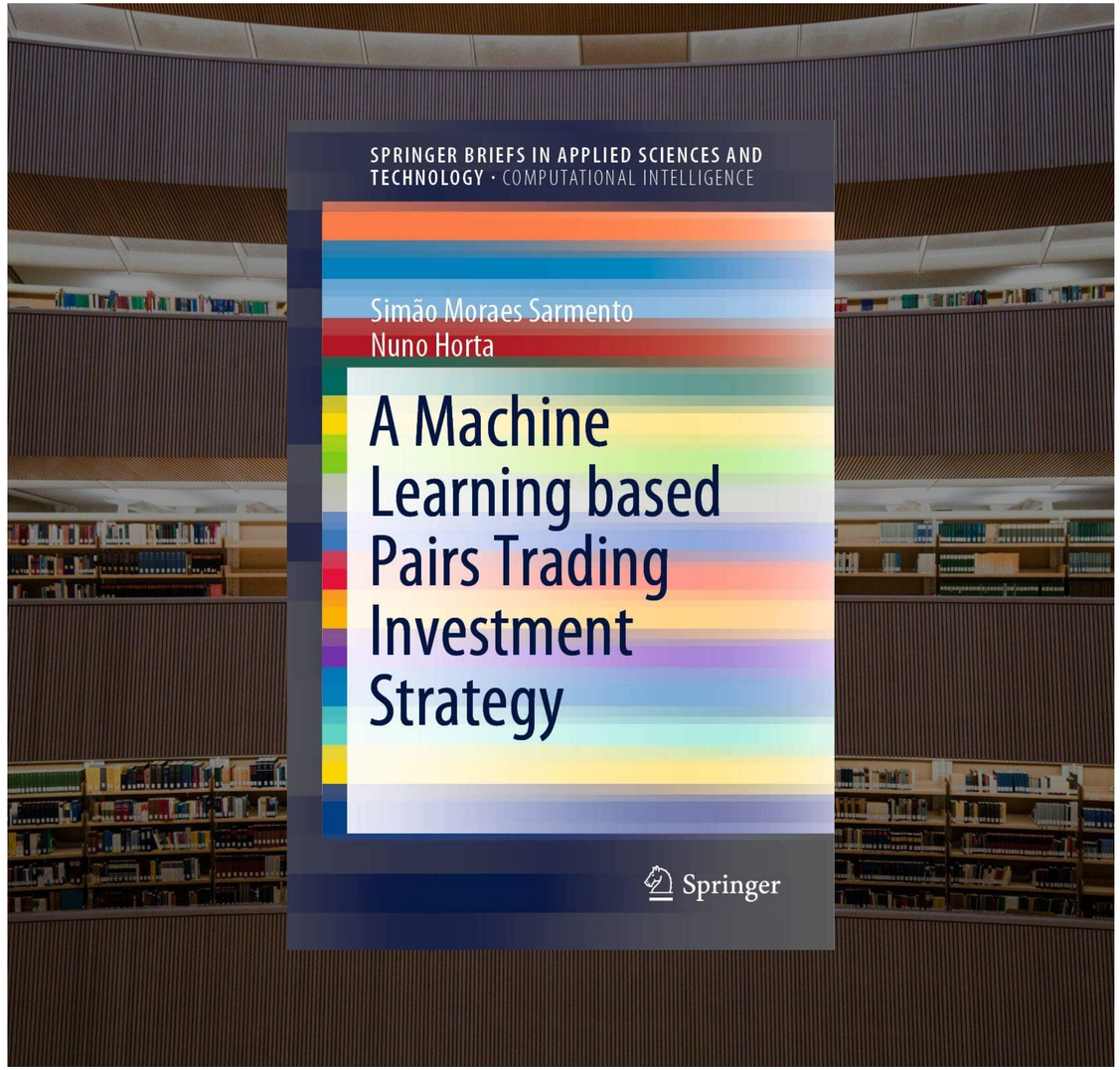
Landmark implementations regarding portfolio optimization.

CONTENTS

INTRODUCTION

A Machine Learning Based Pairs Trading Investment Strategy. (2020)

by *Sarmento, S.M.* and *Horta, N.*



Idea:

- Introduce a new Pairs Selection framework;
- Test the performance of Unsupervised Learning algorithms for asset clustering;
- Compare performance of novel and classic methods;
- Provide ML-based Pairs Trading framework.

Results: Pairs selected using novel methods show more consistent performance qualities.

01. PAIRS TRADING

The Concept of Pairs Trading

FIND A PAIR

Two securities whose prices have moved together historically

MONITOR THE SPREAD

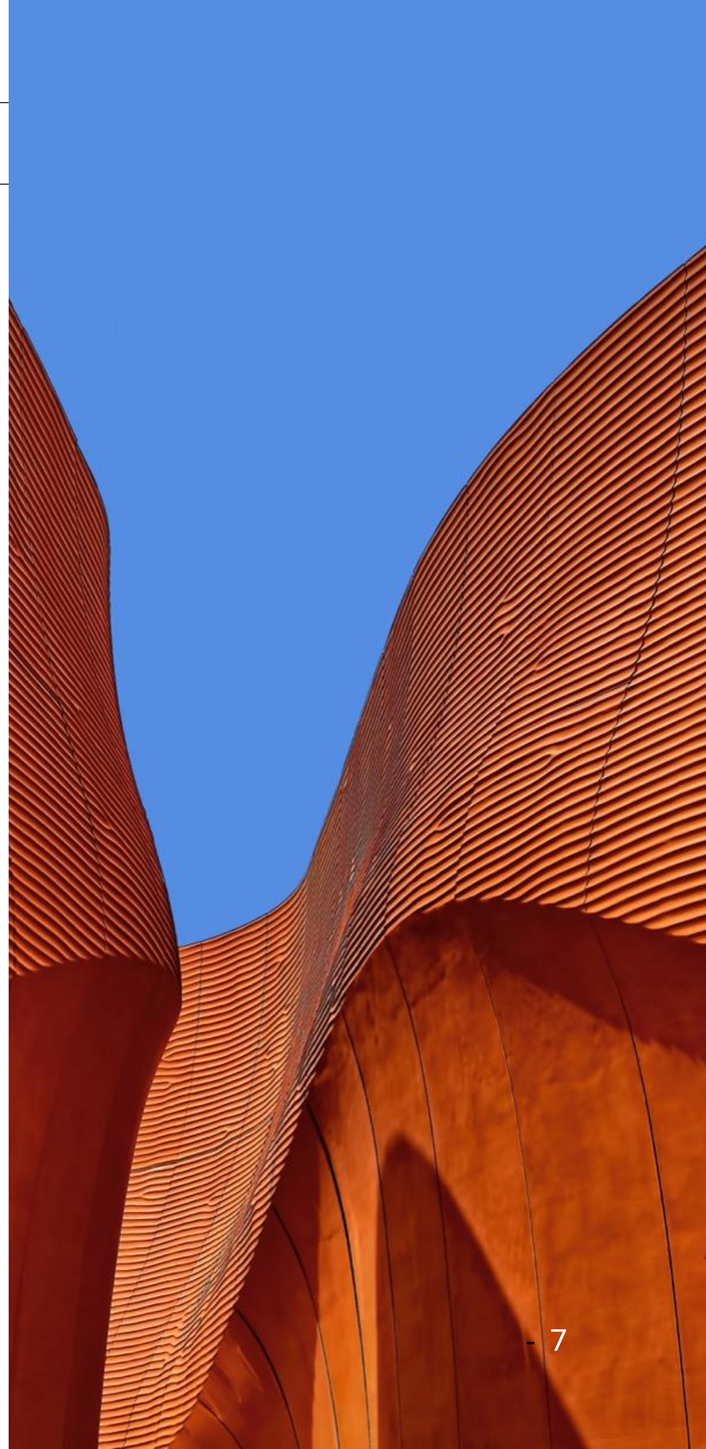
Monitor the spread between prices in a subsequent period

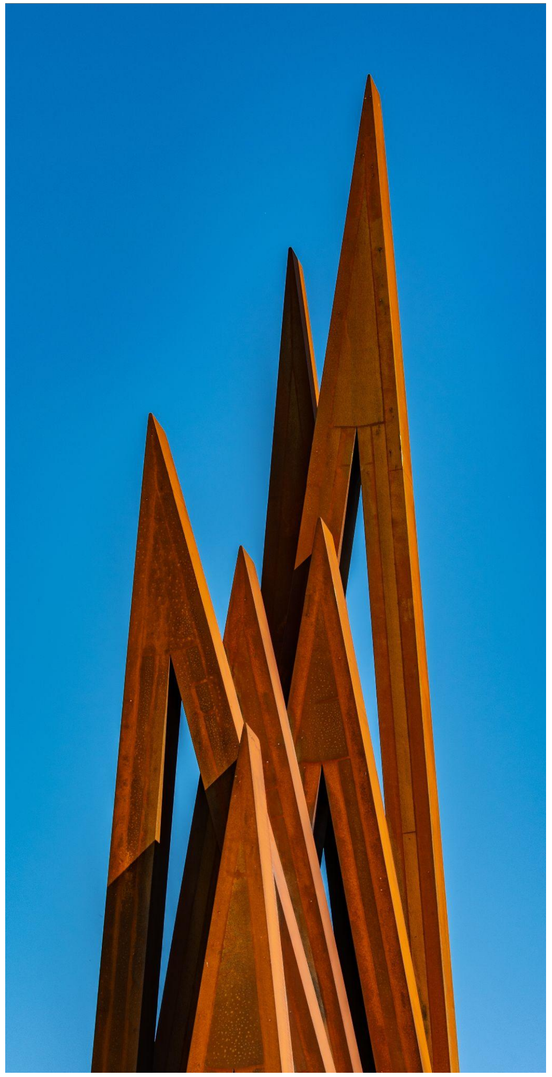
TRADE

If the prices diverge and the spread widens, short the winner and buy the loser

VARIATIONS

Multiple ways to use the concept. Possible extensions to multivariate frameworks





Types of Pairs Trading Strategies

Univariate
(one vs one)

Quasi-multivariate
(one vs many)

Multivariate
(many vs many)

Classification of Pairs Trading Strategies

- Distance Approach
- Cointegration Approach
- Time Series Approach
- Stochastic Control Approach
- Other Approaches
  - Machine Learning Approach
  - Copula Approach
  - PCA & Other Approaches

Statistical Arbitrage Pairs Trading Strategies: Review and Outlook

Christopher Krauss
Department of Statistics and Econometrics
University of Erlangen-Nürnberg, Nürnberg

Wednesday 26th August, 2015

Abstract

This survey reviews the growing literature on pairs trading frameworks, i.e., relative-value arbitrage strategies involving two or more securities. The available research is categorized into five groups: The distance approach uses nonparametric distance metrics to identify pairs trading opportunities. The cointegration approach relies on formal cointegration testing to unveil stationary spread time series. The time series approach focuses on finding optimal trading rules for mean-reverting spreads. The stochastic control approach aims at identifying optimal portfolio holdings in the legs of a pairs trade relative to other available securities. The category "other approaches" contains further relevant pairs trading frameworks with only a limited set of supporting literature. Drawing from this large set of research consisting of more than 90 papers, an in-depth assessment of each approach is performed, ultimately revealing strengths and weaknesses relevant for further research and for implementation.

Keywords: Statistical arbitrage, pairs trading, spread trading, relative-value arbitrage, mean-reversion

1

Statistical arbitrage pairs trading strategies: Review and outlook. (2017)

by *Krauss, C.*

02. PAIRS SELECTION PROBLEM

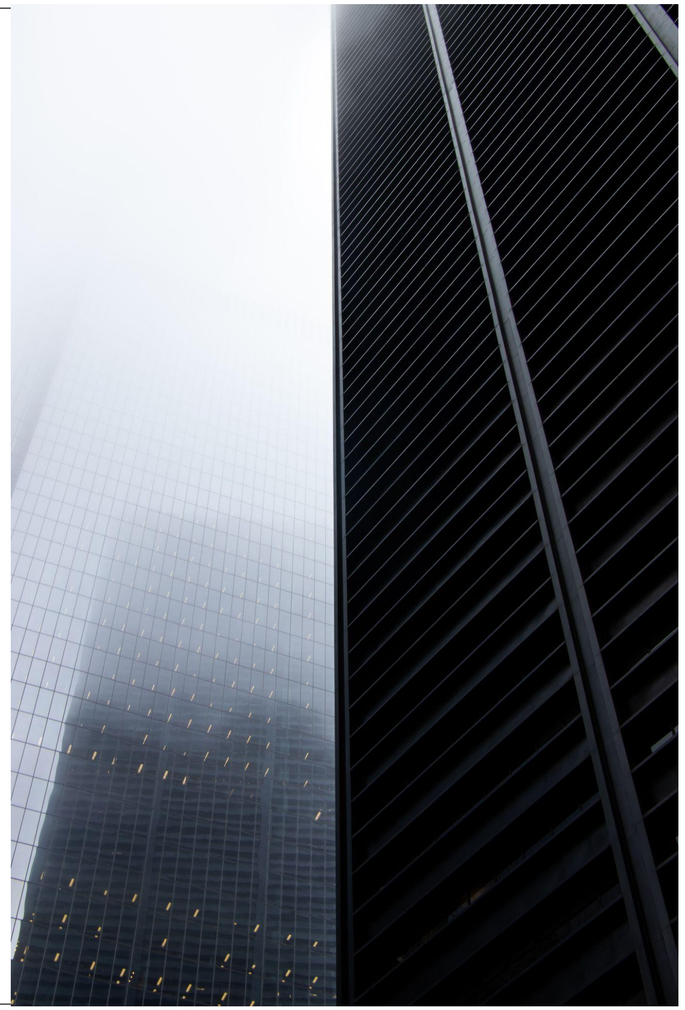
Pairs Selection Dilemma

Fewer restrictions

More restrictions

- Excessive number of combinations to explore
- More likely to find spurious relations

- Less combinations to explore and use
- Resulting pairs are likely to be traded in large volumes



Stages of Pairs Selection

I. Finding Candidate Pairs

- Selecting a set of securities
- Selecting from the same sector/industry
- Using clustering methods on securities
- Using statistical or fundamental information
- Using other techniques

II. Selecting Most Promising Pairs

- Using closeness measures
- Conducting cointegration tests
- Applying rank correlation metrics
- Selecting high mean-reverting spreads
- Using other ranking methods

03. CLASSIC PAIRS SELECTION METHODS

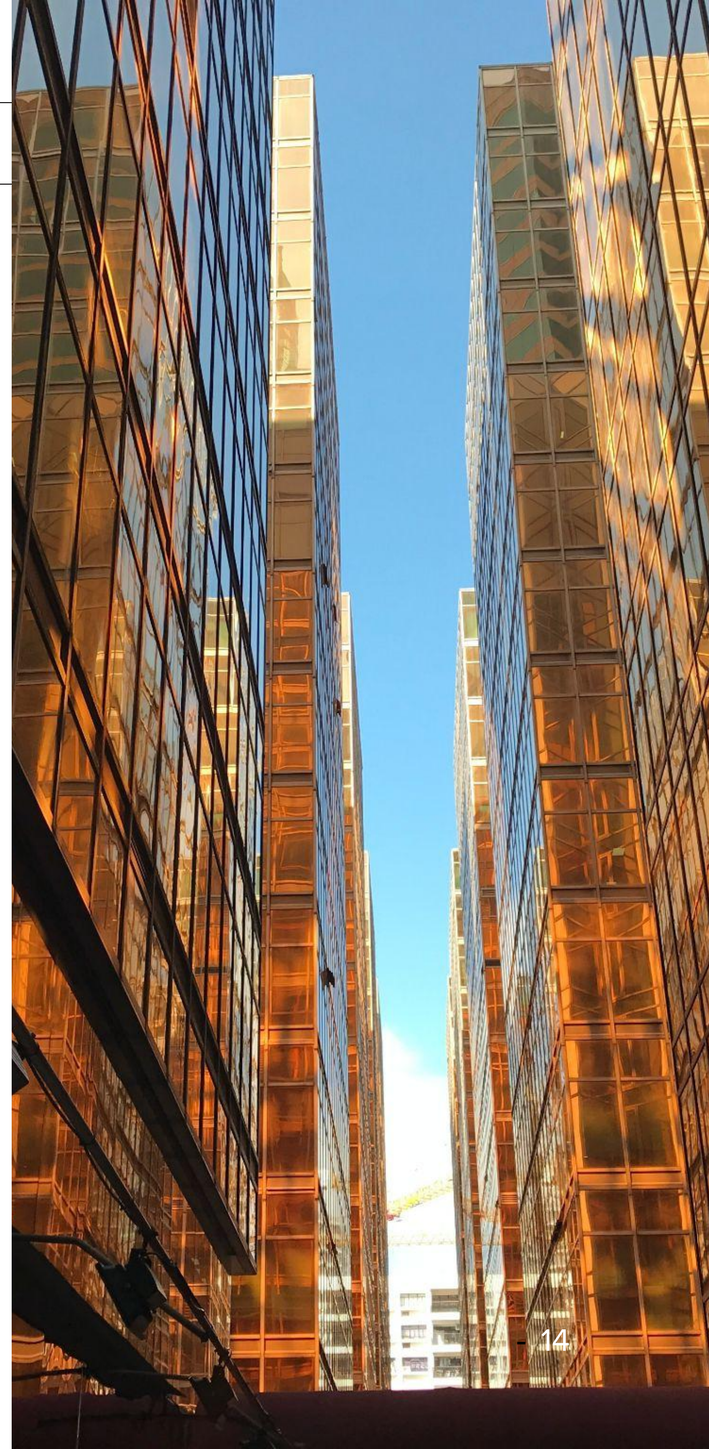
An Overview of Classic Methods

The Minimum Distance Approach

The Correlation Approach

The Cointegration Approach

Other Approaches

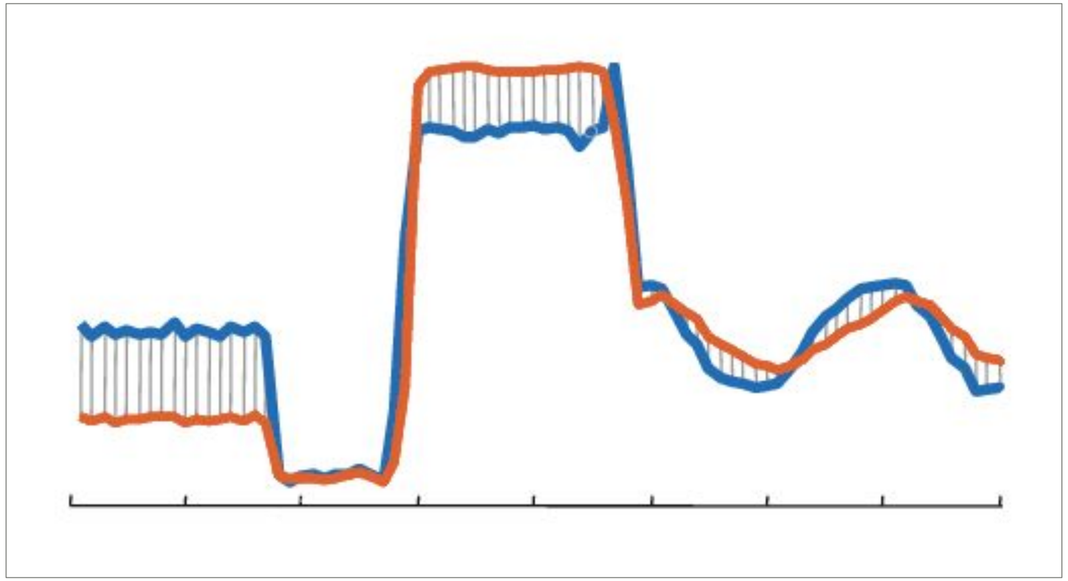


The Minimum Distance Approach

1) Construct cumulative returns indices of securities.

2) Calculate Sum of Euclidean Squared Distances (SSD) between obtained time series.

3) Rank pairs according to minimum historic SSD during the period.

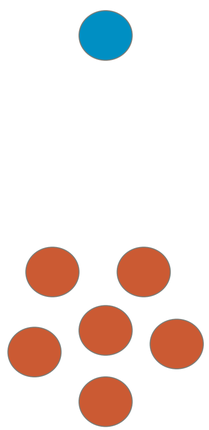


The Correlation Approach

1) For each security ***i*** calculate correlation coefficients with other securities.

2) Pick ***N*** highest correlated securities to ***i*** as its set of pairs.

3) Monitor the deviation of ***i***'th security return from the equal-weighted average return of its ***N*** pair securities.



The Cointegration Approach

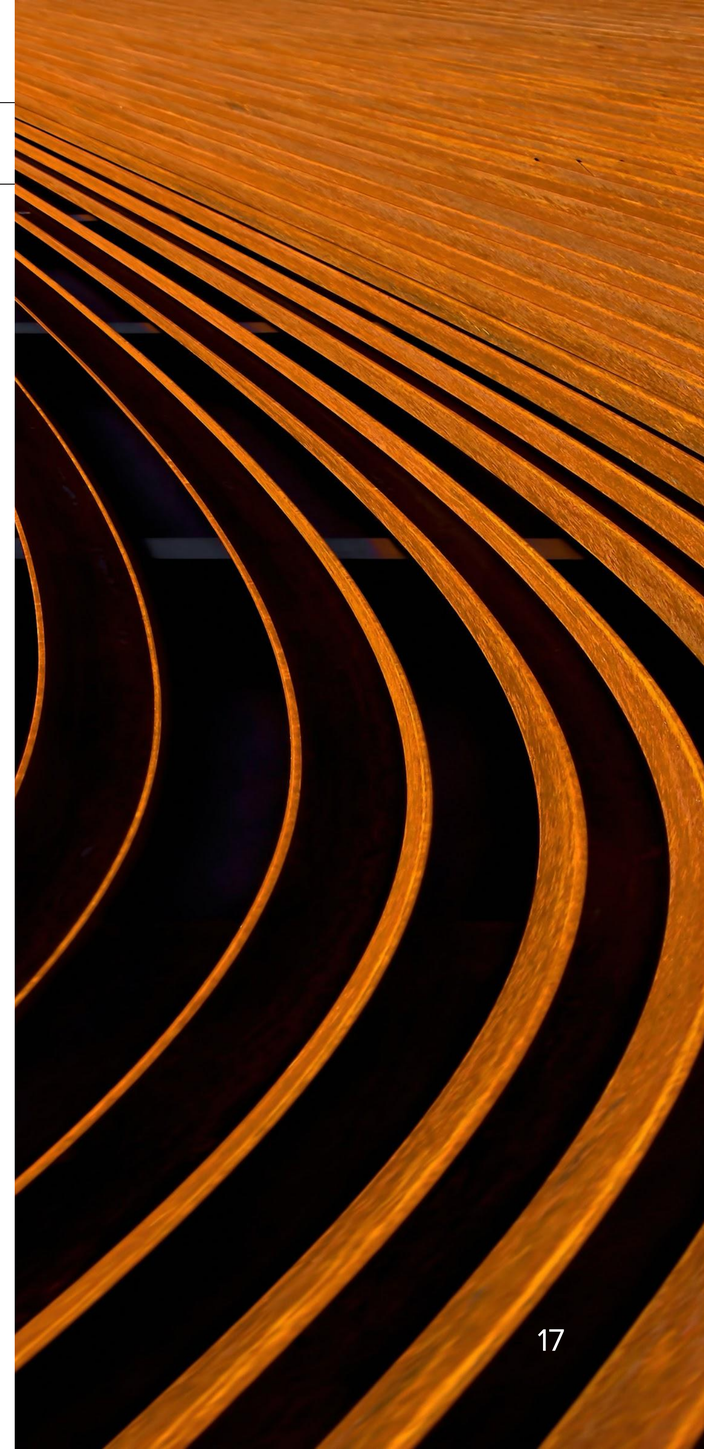
1) Preselect potentially cointegrated pairs based on statistical or fundamental information.

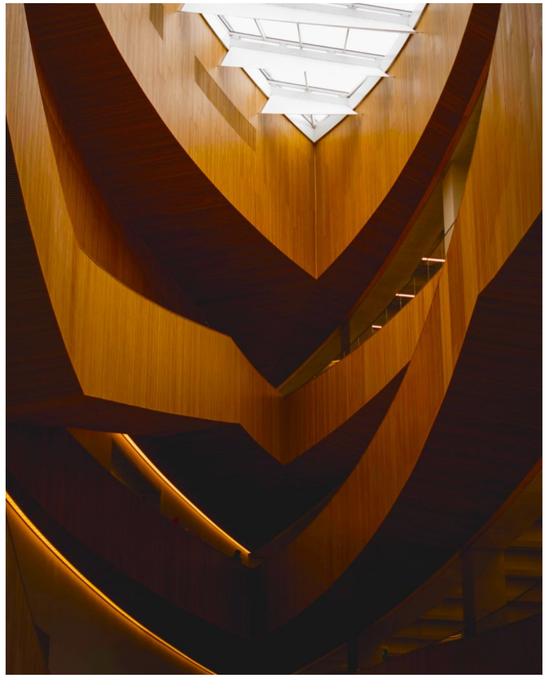
2) Perform a cointegration test of the spread or a test of a similar nature.

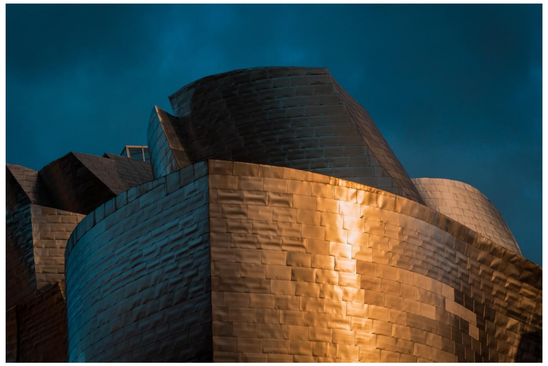
Most used Cointegration Tests:

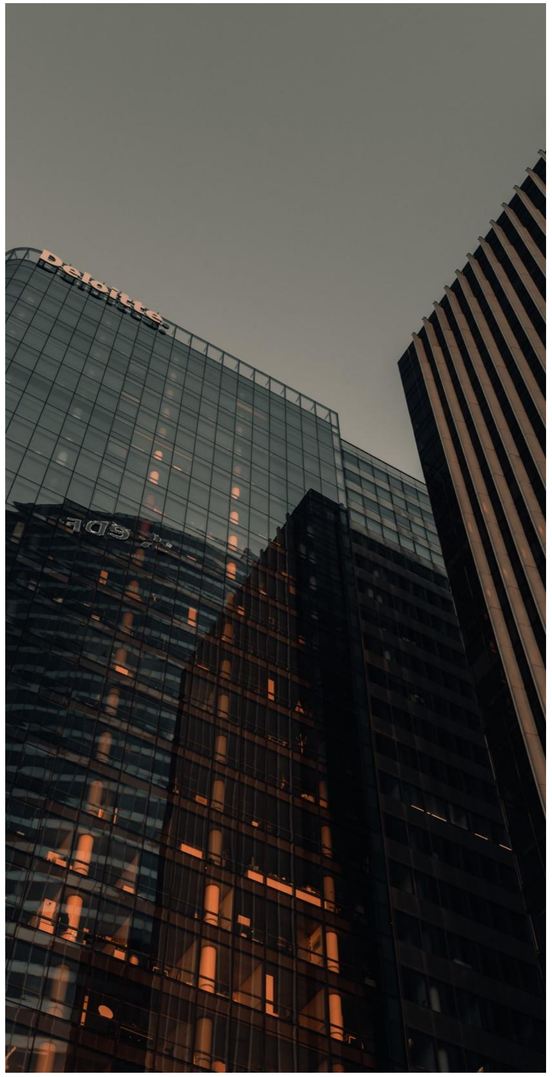
Engle-Granger Test

Johansen Test









Other Approaches

- Ranking based on the spread Hurst exponent
- Ranking based on the spread variance
- Spread number of zero crossings
- Rank correlations - Spearman's ρ, Kendall's τ
- Copula fit measures
- Distribution comparison

04. INTRODUCING THE ML MODEL

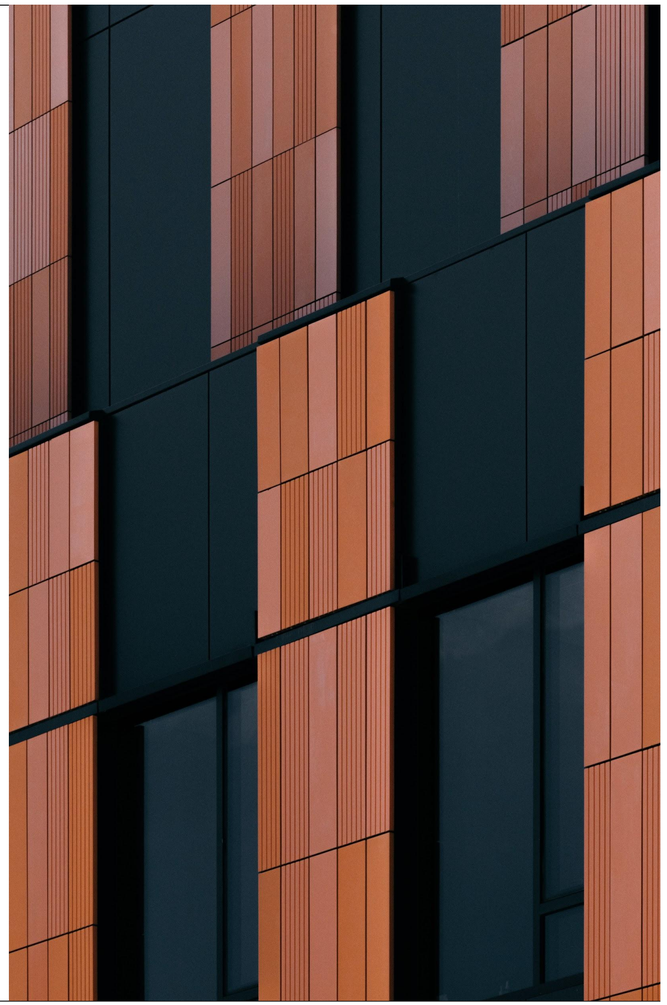
ML Pairs Selection Framework

I. Dimensionality Reduction

II. Unsupervised Learning

III. Pairs Selection

- Find a compact representation of a security
- Cluster securities to unique groups
- Select pairs from clusters with preferred properties



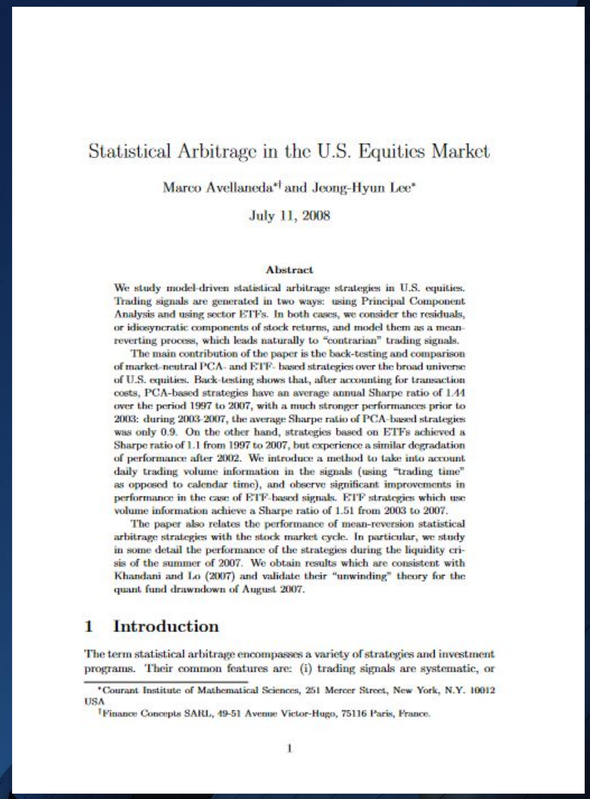
Dimensionality Reduction

1) Construct correlation matrix from normalized returns series of securities.

2) Apply Principal Component Analysis (PCA) to the correlation matrix.

3) Select ***k*** eigenvectors corresponding to the highest variance.

4) Use eigenvector values as ***k*** features for our securities.



Statistical Arbitrage in the U.S. Equities Market

Marco Avellaneda*† and Jeong-Hyun Lee*

July 11, 2008

Abstract

We study model-driven statistical arbitrage strategies in U.S. equities. Trading signals are generated in two ways: using Principal Component Analysis and using sector ETFs. In both cases, we consider the residuals, or idiosyncratic components of stock returns, and model them as a mean-reverting process, which leads naturally to "contrarian" trading signals.

The main contribution of the paper is the back-testing and comparison of market-neutral PCA- and ETF- based strategies over the broad universe of U.S. equities. Back-testing shows that, after accounting for transaction costs, PCA-based strategies have an average annual Sharpe ratio of 1.44 over the period 1997 to 2007, with a much stronger performances prior to 2003: during 2003-2007, the average Sharpe ratio of PCA based strategies was only 0.9. On the other hand, strategies based on ETFs achieved a Sharpe ratio of 1.1 from 1997 to 2007, but experience a similar degradation of performance after 2002. We introduce a method to take into account daily trading volume information in the signals (using "trading time" as opposed to calendar time), and observe significant improvements in performance in the case of ETF-based signals. ETF strategies which use volume information achieve a Sharpe ratio of 1.51 from 2003 to 2007.

The paper also relates the performance of mean-reversion statistical arbitrage strategies with the stock market cycle. In particular, we study in some detail the performance of the strategies during the liquidity crisis of the summer of 2007. We obtain results which are consistent with Khandani and Lo (2007) and validate their "unwinding" theory for the quant fund drawdown of August 2007.

1 Introduction

The term statistical arbitrage encompasses a variety of strategies and investment programs. Their common features are: (i) trading signals are systematic, or

*Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, N.Y. 10012 USA

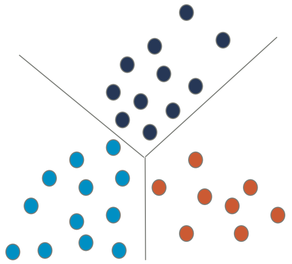
†Finance Concepts SARL, 49-51 Avenue Victor-Hugo, 75116 Paris, France.

1

Statistical arbitrage in the US equities market. (2008)

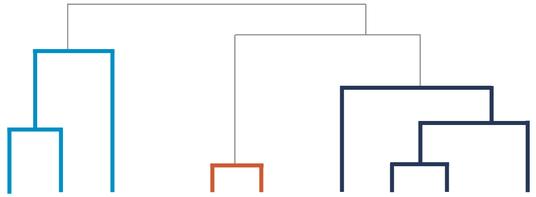
by *Avellaneda, M.* and *Lee, J.H.*

Clustering Algorithms



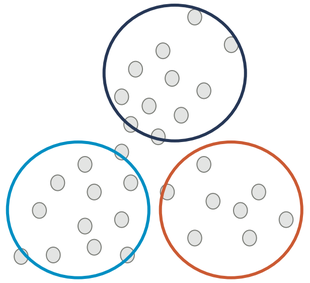
Partitioning Clustering

- Not great at handling noisy data
- Clusters are forced to be convex
- Need to specify the number of clusters in advance



Hierarchical Clustering

- Allows group selection with preferred granularity
- Need to create an automatic termination criterion



Density-based Clustering

- Robust to outliers
- Arbitrary shapes of clusters allowed
- No need to specify the number of clusters in advance

Unsupervised Learning

DBSCAN

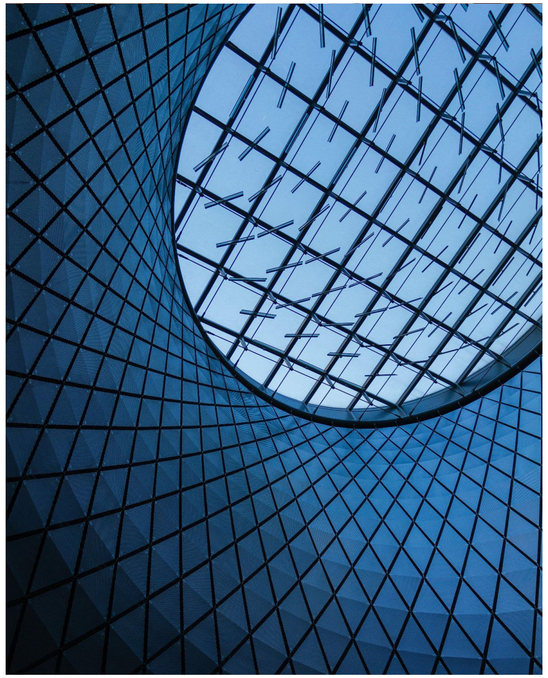
(Density-Based Spatial Clustering of Applications with Noise)

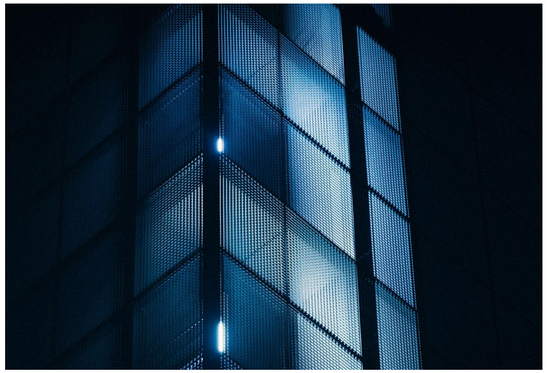
- A density-based clustering algorithm.
- Based on the concept of density-reachability.
- Dependant on the parameter of distance - ***eps***.
- Assumes clusters are evenly dense.

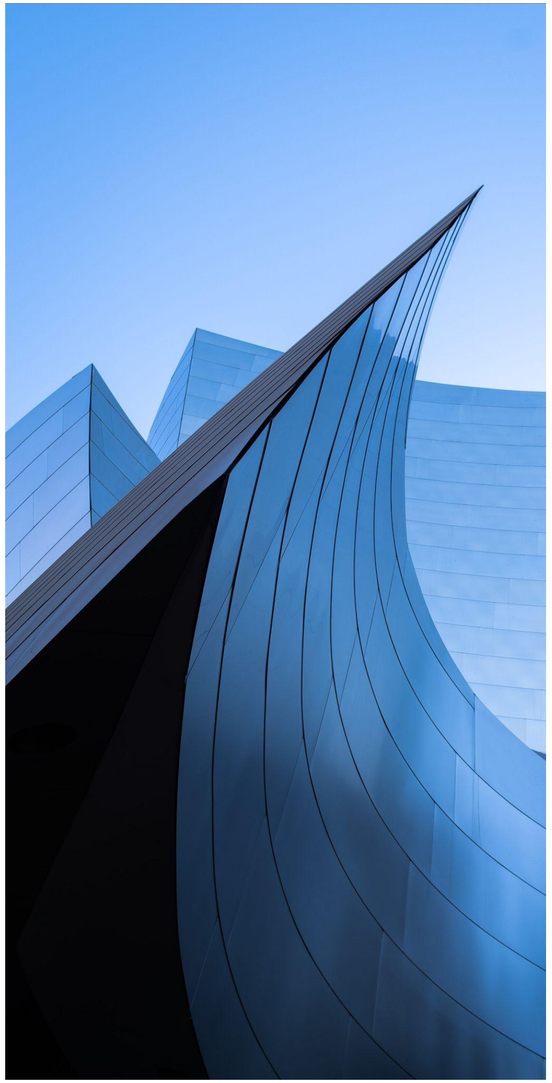
OPTICS

(Ordering Points To Identify the Clustering Structure)

- A density-based clustering algorithm.
- Based on the concept of reachability-distance.
- Automatically selects the ***eps*** parameter.
- Addresses the issue of clustering data of varying density.







ARODS (Absolute Rules Of Disqualification)

The original framework takes the following criteria into account:

- The elements in a pair are cointegrated.
- Strong mean-reversion revealed by the Hurst exponent of the spread.
- Convergence of the spread in convenient periods.
- Spread reverts to its mean with enough frequency.

Pairs Selection Criteria

X and Y cointegrated? → Yes

Determining a strong equilibrium relationship

H < 0.5 ? → Yes

Extra test of the spread mean-reversion

1 day < hl < 1 year ? → Yes

If the half-life of mean reversion is in line with the trading period

mean-cross ≥ 12 ? → Yes

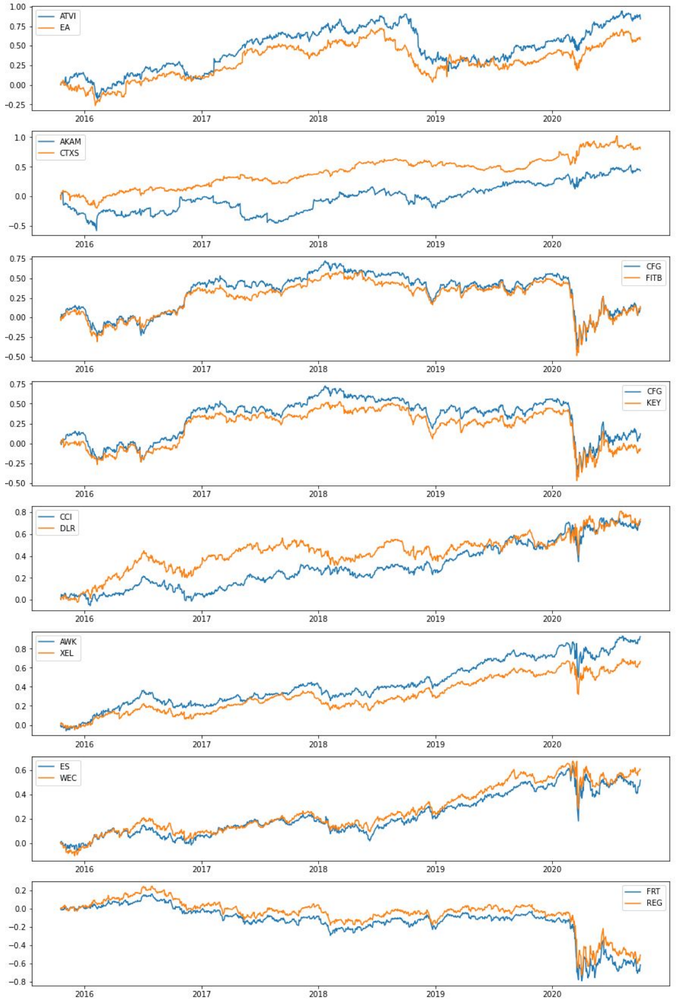
Pick pairs which spread crosses the mean with good frequency

Eligible pair

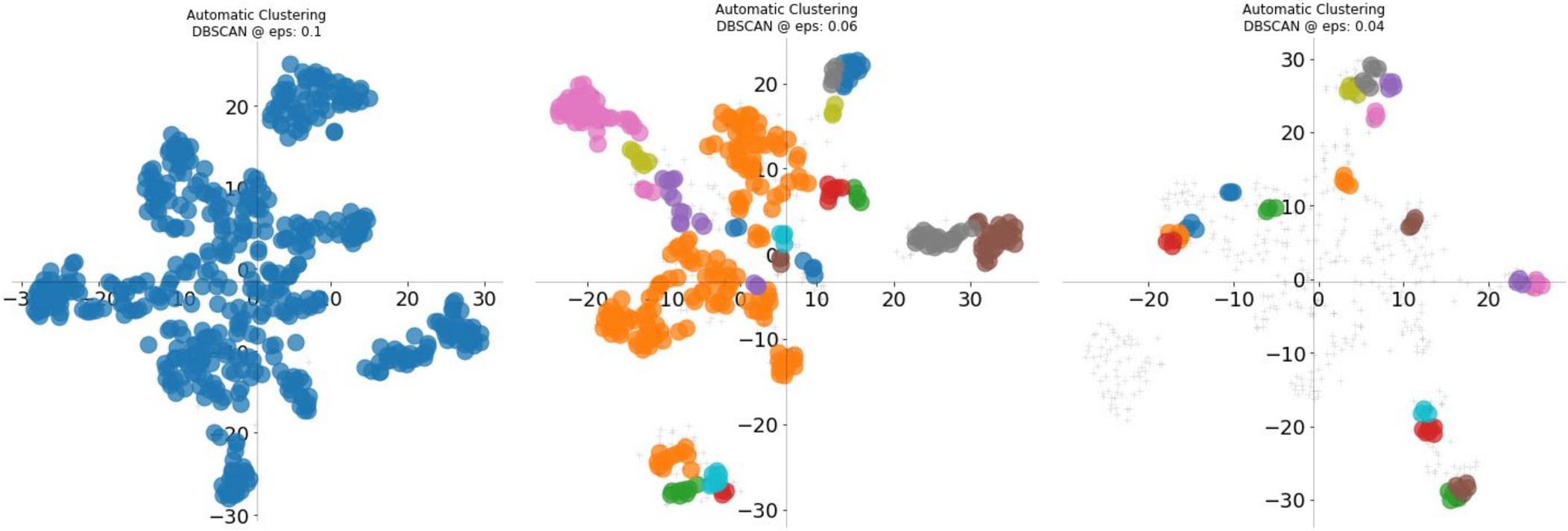
05. MODEL RESULTS

SHOW ME THE CODE

```python
# Importing packages
import pandas as pd
import numpy as np
from arbitragelab.ml_approach import OPTICSDBSCANPairsClustering
from arbitragelab.pairs_selection import CointegrationPairsSelector

# Getting the dataframe with time series of asset returns
data = pd.read_csv('X_FILE_PATH.csv', index_col=0, parse_dates = [0])

pairs_clusterer = OPTICSDBSCANPairsClustering(data)

# Price data is reduced to its component features using PCA
pairs_clusterer.dimensionality_reduction_by_components(5)

# Clustering is performed over feature vector
clustered_pairs = pairs_clusterer.cluster_using_optics({'min_samples': 3})

# Generated Pairs are processed through the rules mentioned above
pairs_selector = CointegrationPairsSelector(prices_df=data,
                                            pairs_to_filter=clustered_pairs)
filtered_pairs = pairs_selector.select_pairs()

# Generate a Panel of information of the selected pairs
final_pairs_info = pairs_selector.describe_extra()

# Import the ticker sector info csv
sectoral_info = pd.read_csv('X_FILE_PATH.csv')

# Generate a sector/industry relationship Panel of each pair
pairs_selector.describe_pairs_sectoral_info(final_pairs_info['leg_1'],
                                            final_pairs_info['leg_2'],
                                            sectoral_info)
```

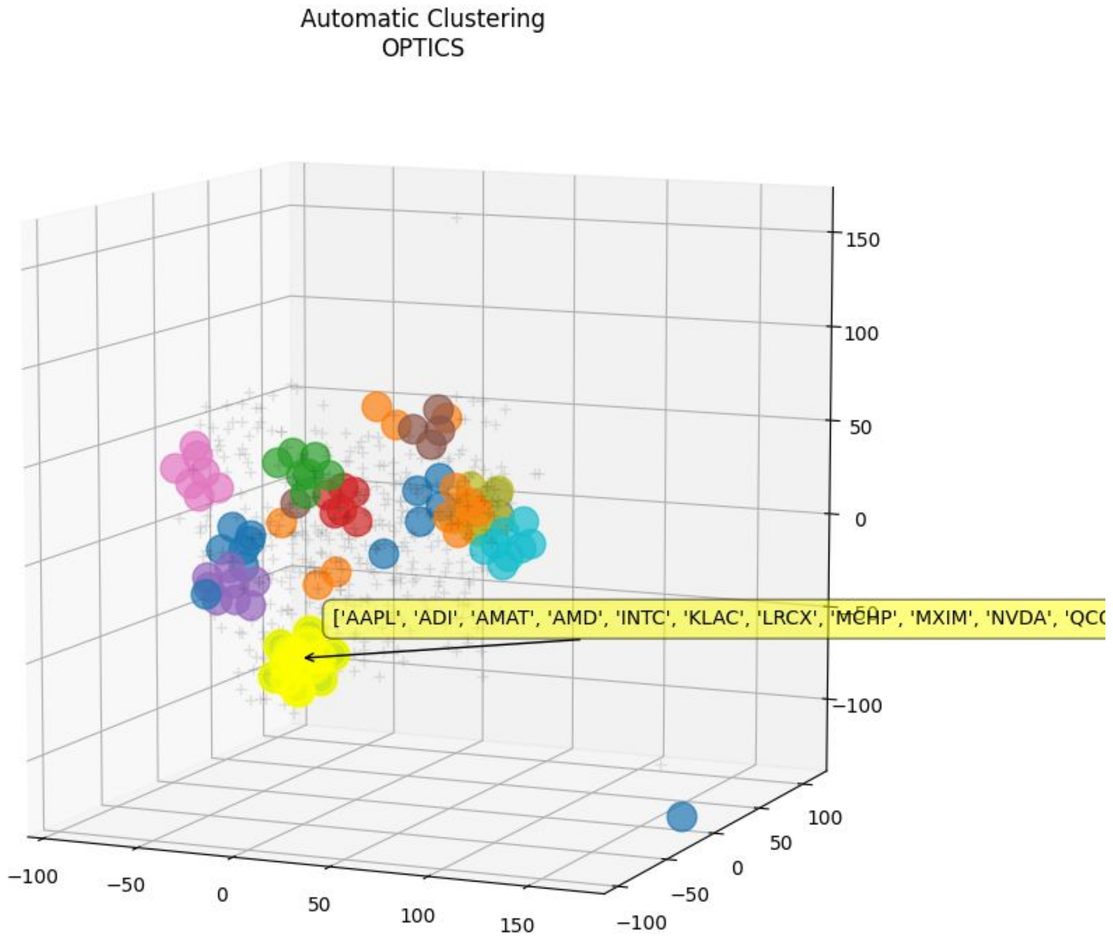



DBSCAN CLUSTERING



Parameter ***eps*** decreasing

OPTICS CLUSTERING



Automatic ***eps*** selection

PAIRS SELECTION

Final set of pairs

	leg_1	leg_2	coint_t	pvalue	hedge_ratio	hurst_exponent	half_life	crossovers
0	AJG	ICE	-5.147984	0.000011	0.832406	0.280300	1.178147	True
1	AJG	MMC	-3.984768	0.001492	0.892408	0.333155	1.325405	True
2	ICE	MMC	-4.314326	0.000420	1.072514	0.353929	3.312397	True
3	MMC	WLTW	-4.173535	0.000730	1.758648	0.330393	2.353220	True
4	CHTR	TMUS	-3.834987	0.002569	0.166618	0.350558	2.954012	False
5	EW	FISV	-3.775258	0.003171	1.247778	0.342441	1.774284	True
6	EW	GPN	-4.614576	0.000121	2.381694	0.388706	10.865253	False

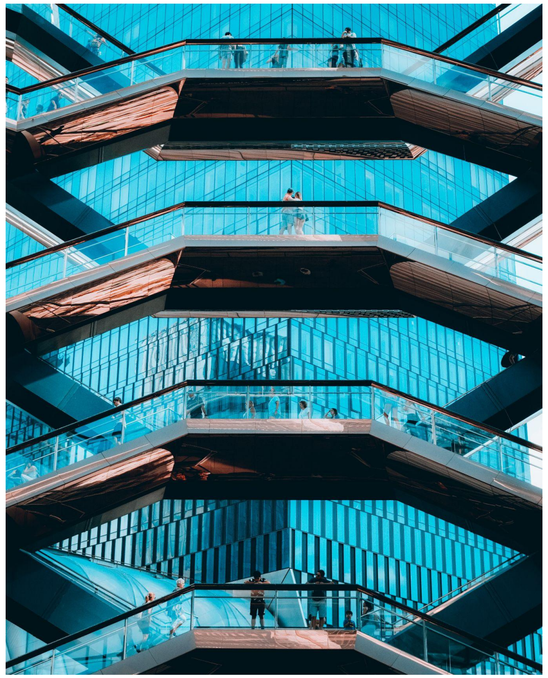
Pairs selection statistics

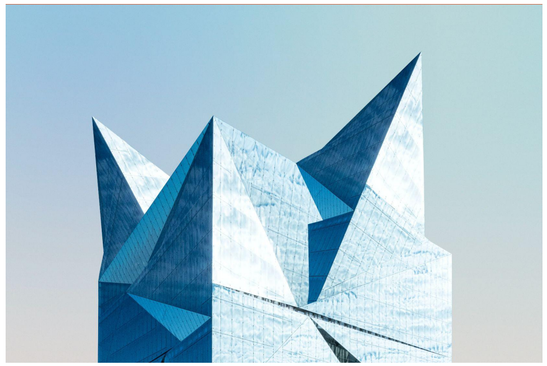
0	No. of Clusters	46
1	Total Pair Combinations	631
2	Pairs passing Coint Test	53
3	Pairs passing Hurst threshold	53
4	Pairs passing Half-Life threshold	32
5	Final Set of Pairs	23

Pairs Industry/Sector

	Leg 1 Ticker	Industry	Sector	Leg 2 Ticker	Industry	Sector
0	AJG	Insurance Brokers	Financial Services	ICE	Financial Data & Stock Exchanges	Financial Services
1	AJG	Insurance Brokers	Financial Services	MMC	Insurance Brokers	Financial Services
2	ICE	Financial Data & Stock Exchanges	Financial Services	MMC	Insurance Brokers	Financial Services
3	MMC	Insurance Brokers	Financial Services	WLTW	Insurance Brokers	Financial Services
4	CHTR	Entertainment	Communication Services	TMUS	Telecom Services	Communication Services
5	EW	Medical Devices	Healthcare	FISV	Information Technology Services	Technology
6	EW	Medical Devices	Healthcare	GPN	Specialty Business Services	Industrials

06. DISCUSSION





Extra Details

- Securities to use.
- Data preparation.
- Methods for testing the quality of selected pairs.
- Future research topics.

07. REFERENCES

REFERENCES

- **Sarmento, S.M. and Horta, N., (2020).** A Machine Learning Based Pairs Trading Investment Strategy. Springer. [Available here]

- **Gatev, E., Goetzmann, W.N. and Rouwenhorst, K.G., (2006).** Pairs trading: Performance of a relative-value arbitrage rule. The Review of Financial Studies, 19(3), pp.797-827. [Available here]

- **Krauss, C., (2017).** Statistical arbitrage pairs trading strategies: Review and outlook. Journal of Economic Surveys, 31(2), pp.513-545. [Available here]

- **Chen, H., Chen, S., Chen, Z. and Li, F., (2019).** Empirical investigation of an equity pairs trading strategy. Management Science, 65(1), pp.370-389. [Available here]

REFERENCES

- **Vidyamurthy, G., (2004).** Pairs Trading: quantitative methods and analysis (Vol. 217). John Wiley & Sons. [Available here]

- **Avellaneda, M. and Lee, J.-H. (2010).** Statistical arbitrage in the US equities market.Quantitative Finance, 10(7):761–782. [Available here]

- **Ester, M., Kriegel, H.P., Sander, J. and Xu, X., (1996).** A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231). [Available here]

- **Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., (1999).** OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record, 28(2), pp.49-60. [Available here]

THANK YOU!

Does anyone have any questions?

@IllyaBarziy

@PanPip

www.linkedin.com/in/illyabarziy/

HUDSON AND THAMES