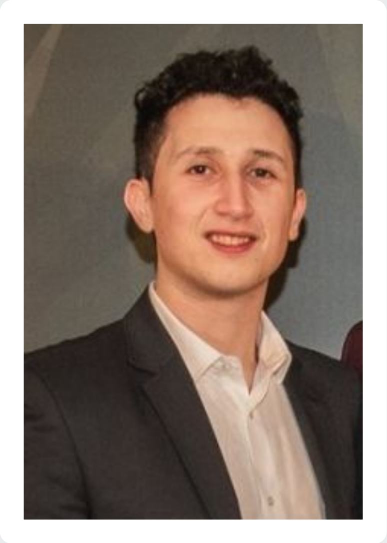
HUDSON AND THAMES

ML for Pairs Selection

Aaron Debrincat

Bio: Aaron Debrincat



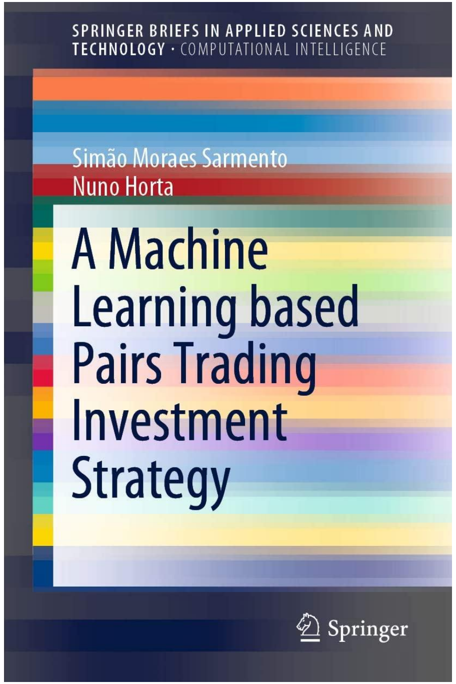
- Master’s student in AI at the University of Malta.
- Presently an Apprentice at Hudson Thames.
- **LinkedIn**: bit.ly/3nOkhHT
- **Twitter**: @debrincat_aaron
- **Github**: @AaronDeb

Overview

- Major Statistical Arbitrage Components:
  - Get Tradable Pairs.
  - Model the Spread between those Pairs.
  - Trade based on the spread model.
- Important foundational building block of your strategy.
- How do we get them? A sprinkling of ML.
- Are there parameters that need tuning? Only if you want to.

The following work is an implementation based on the research work of Sarmento & Horta

Sarmento, S.M. and Horta, N., 2020. A Machine Learning based Pairs Trading Investment Strategy.



SPRINGER BRIEFS IN APPLIED SCIENCES AND TECHNOLOGY · COMPUTATIONAL INTELLIGENCE

Simão Moraes Sarmento
Nuno Horta

A Machine Learning based Pairs Trading Investment Strategy

Springer

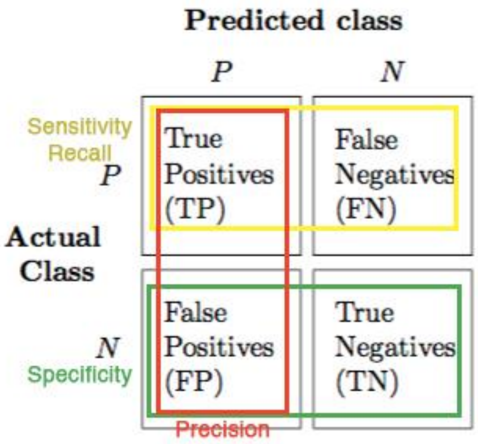
HUDSON AND THAMES

Previously Proposed Methods

- Classic Approach
  - Take in the whole asset universe.
  - Generate all possible pairwise combinations.
  - Execute statistical similarity tests.
- Similarity Test Approaches;
  - Distance Based (Sum of squared returns)
  - Cointegration Based (Engle-Granger, Johansen)
  - Correlation Based
  - Hybrid Methods

Major Issues

- **Large Computational Cost.**
- **Family Wise Error Rate (5%).**
  - Proposed fix by (Harlacher 2016) using Bonferroni Correction.
  - Results were mixed.
  - The approach turned out to be too conservative and also impeded the discovery of truly cointegrated combinations.
  - Author recommends pre partitioning the asset universe before running the combination calculations.



Predicted class

Actual Class	P	N
Sensitivity Recall P	True Positives (TP)	False Negatives (FN)
N Specificity	False Positives (FP)	True Negatives (TN)

Precision

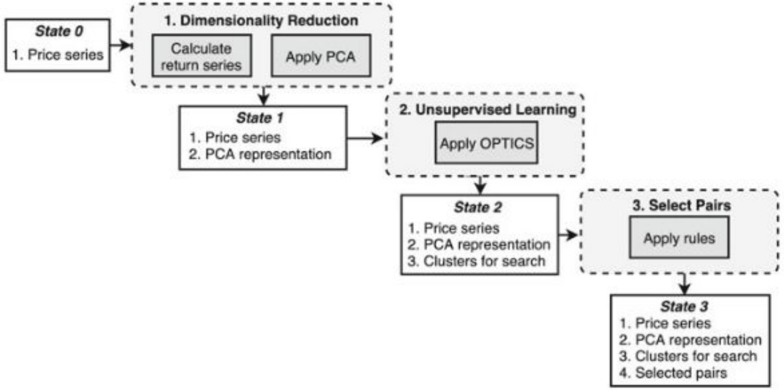
New Pairs Selection

Constraints

- No constraint on asset universe size.
- Minimize likelihood of finding spurious relationships.
- Find uncommon combinations that haven't been found by overall trading community.

Solution *hint (Unsupervised Learning)*

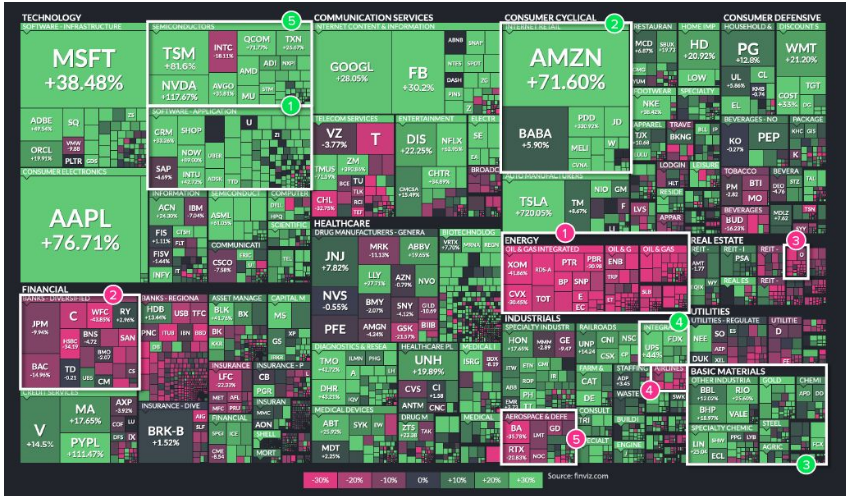
- Use PCA to find a compact representation.
- Use a clustering algorithm to separate into distinct groups.
- Use Absolute Rules of Disqualification (ARODs) to select the right pairs.



Ways to Categorize Assets

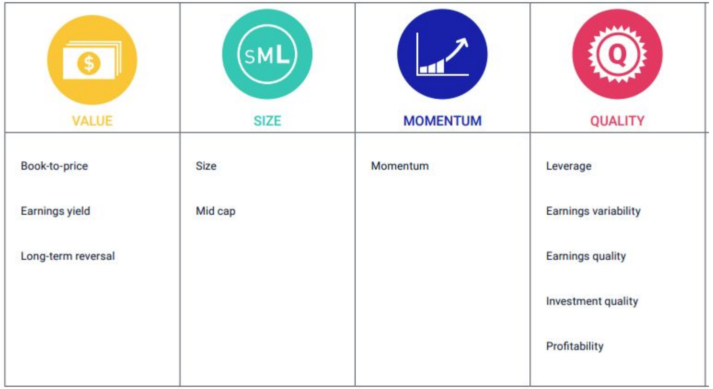
Classical Way

- Economic Category
- MSCI Market Classification



Factor Way

- Barra Risk Model

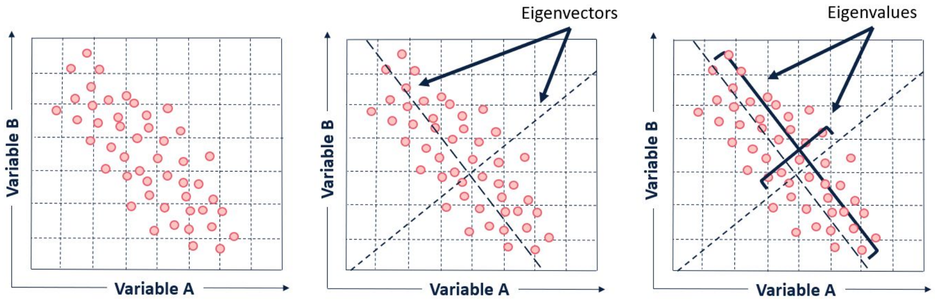


VALUE	SIZE	MOMENTUM	QUALITY
Book-to-price	Size	Momentum	Leverage
Earnings yield	Mid cap		Earnings variability
Long-term reversal			Earnings quality
			Investment quality
			Profitability

HUDSON
AND THAMES

Dimensionality Reduction

- Prepare input dataset, in this case will be stock returns.
- Use PCA to reduce the asset universe into principal components.
- Take into consideration curse of dimensionality.
- Definite cap on the number of components that can be selected.
- At the same time we don't want to leave information on the table.



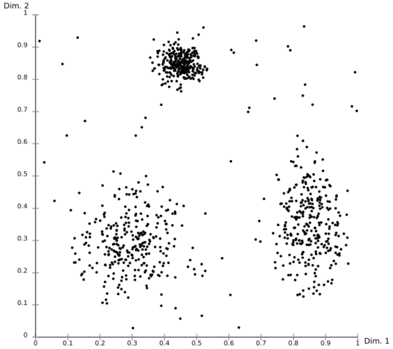
Clustering

DBSCAN

- We can easily detect clusters of points because typically the density of points within each cluster is considerably higher than outside of the cluster.
- The main idea is that clusters depend on point density.

OPTICS

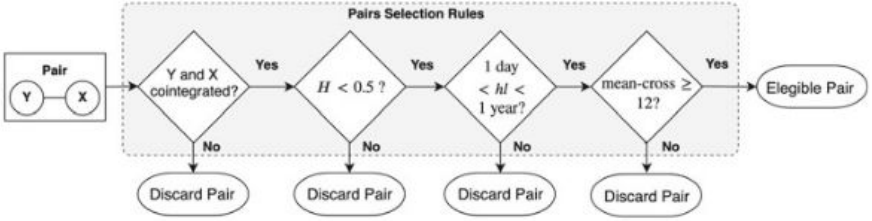
- No need to specify the number of clusters in advance.
- Robust to outliers.
- Suitable for clusters with varying density.



HUDSON
AND THAMES

Absolute Rules of Disqualification (ARODs)

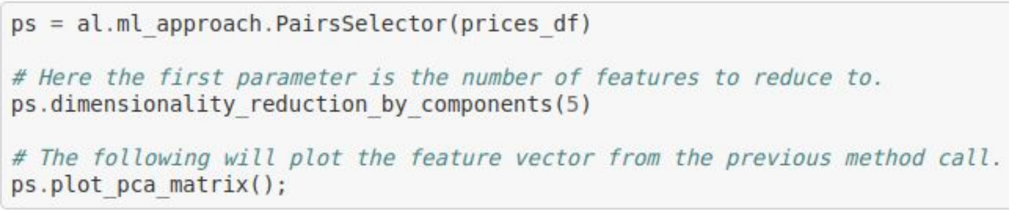
1. Check for Cointegration using Engle Granger test.
    a. Finds sound equilibrium relationships.
    b. The literature suggests cointegration performs better, when compared with minimum distance and correlation approaches.
2. Make sure that spread is mean reverting using the hurst exponent.
    a. Provides an extra layer of confidence to validate mean-reverting series.
3. Make sure the spread is tradable in the medium term (> 1 day and < 365 days).
4. Check spread reversion consistency.



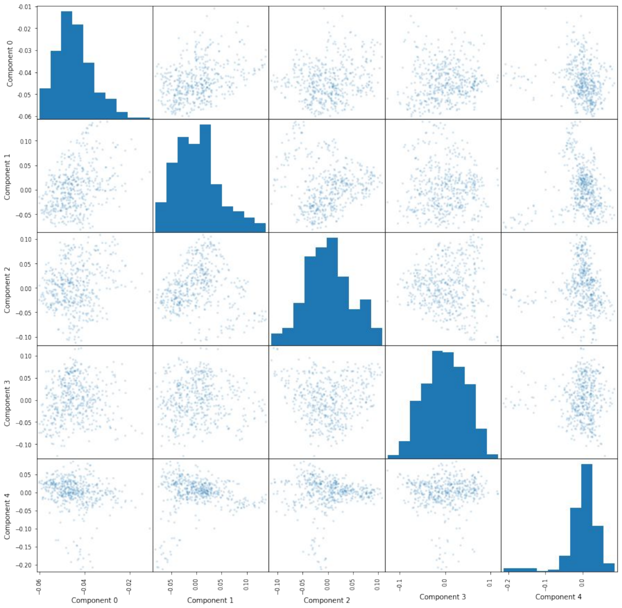
How to use the Module?

Step 1 - Dimensionality Reduction

- Initial processing/scaling of returns dataset.
- PCA reduction based on the number of components given.
- Feature Vector is stored in the class object but still publicly accessible if needed.
- Visualization helper method.



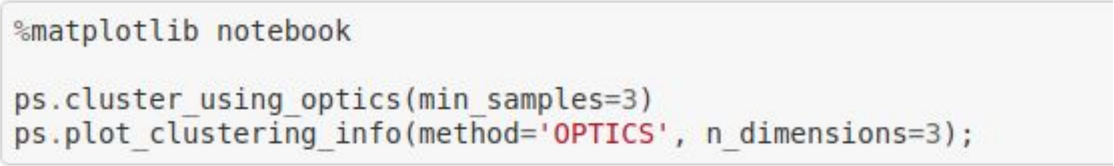
```
ps = al.ml_approach.PairsSelector(prices_df)

# Here the first parameter is the number of features to reduce to.
ps.dimensionality_reduction_by_components(5)

# The following will plot the feature vector from the previous method call.
ps.plot_pca_matrix();
```



HUDSON
AND THAMES

Step 2 - Clustering

- Two methods of clustering (OPTICS & DBSCAN).
- If in doubt use OPTICS.
- If more control on parameters is needed use DBSCAN.
- Visualization helper method allows for both 2 and 3 dimensional visual representations.

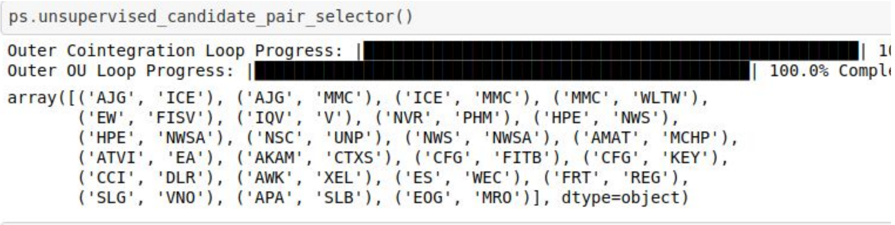


```
%matplotlib notebook

ps.cluster_using_optics(min_samples=3)
ps.plot_clustering_info(method='OPTICS', n_dimensions=3);
```

HUDSON AND THAMES

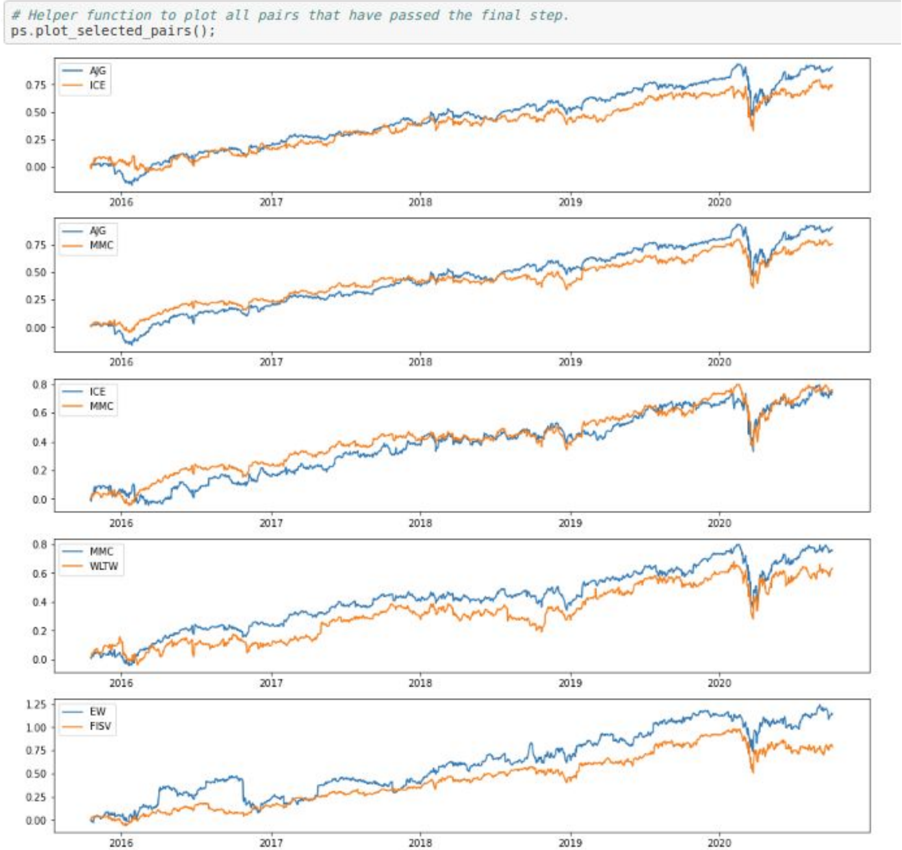
Step 3 - ARODs

- First, it is imposed that pairs are cointegrated, using a p-value of 1%.
- Then, the spread’s Hurst exponent, should be smaller than 0.5.
- Additionally, the half-life period should lay between one day and one year.
- Finally, it is imposed that the spread crosses a mean at least 12 times per year.



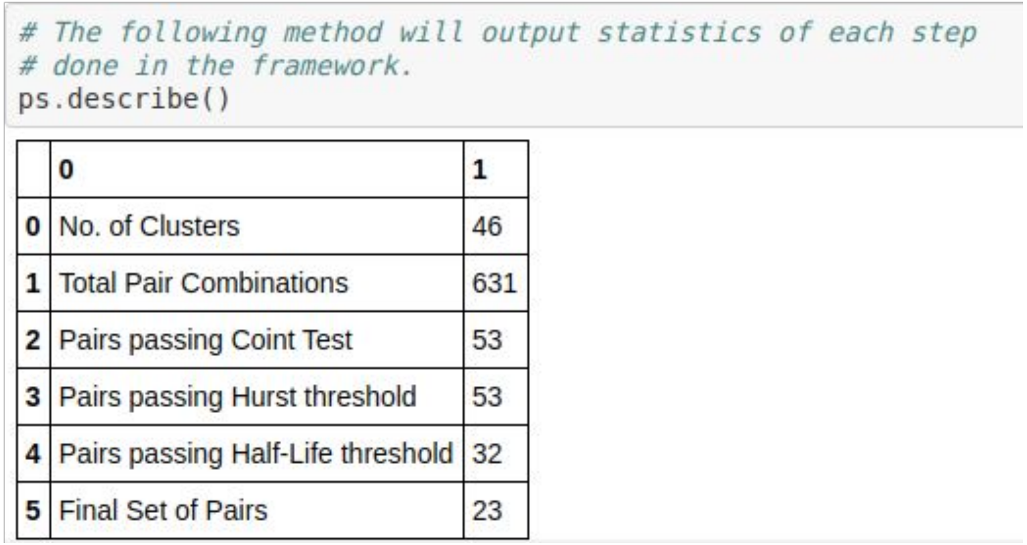
```
ps.unsupervised_candidate_pair_selector()
```

```
Outer Cointegration Loop Progress: |██████████████████████████████████████████████████| 1
Outer OU Loop Progress: |██████████████████████████████████████████████████| 100.0% Comple
array([('AJG', 'ICE'), ('AJG', 'MMC'), ('ICE', 'MMC'), ('MMC', 'WLTW'),
       ('EW', 'FISV'), ('IQV', 'V'), ('NVR', 'PHM'), ('HPE', 'NWS'),
       ('HPE', 'NWSA'), ('NSC', 'UNP'), ('NWS', 'NWSA'), ('AMAT', 'MCHP'),
       ('ATVI', 'EA'), ('AKAM', 'CTXS'), ('CFG', 'FITB'), ('CFG', 'KEY'),
       ('CCI', 'DLR'), ('AWK', 'XEL'), ('ES', 'WEC'), ('FRT', 'REG'),
       ('SLG', 'VNO'), ('APA', 'SLB'), ('EOG', 'MRO')], dtype=object)
```



```
# Helper function to plot all pairs that have passed the final step.
ps.plot_selected_pairs();
```

Results



```
# The following method will output statistics of each step
# done in the framework.
ps.describe()
```

	0	1
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

References

- Sarmento, S.M. and Horta, N., 2020. Enhancing a Pairs Trading strategy with the application of Machine Learning. Expert Systems with Applications, p.113490.
- Van Der Maaten, L., Postma, E., and Van den Herik, J., 2009. Dimensionality reduction: a comparative. J Mach Learn Res, 10(66-71), p.13.
- Avellaneda, M. and Lee, J.H., 2010. Statistical arbitrage in the US equities market. Quantitative Finance, 10(7), pp.761-782.
- 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.

Thank you for your time!

HUDSON
AND THAMES

Questions?