HUDSON AND THAMES

Codependence with MlFinLab

Introduction: Illya Barziy

- Quantitative Researcher at Hudson & Thames.
- MA in Computer Science and Econometrics at University of Warsaw.

GitHub: @PanPip

Twitter: @IllyaBarziy

Overview

MlFInLab package

Functionality

Examples of use

Codependence Module

Risk Estimators Submodule

Applications in MlFinLab

References

MlFinLab

- Novel Quantitative Finance techniques from elite and peer-reviewed journals.
- Written in Python and available on PyPi

```
pip install mlfinlab
```

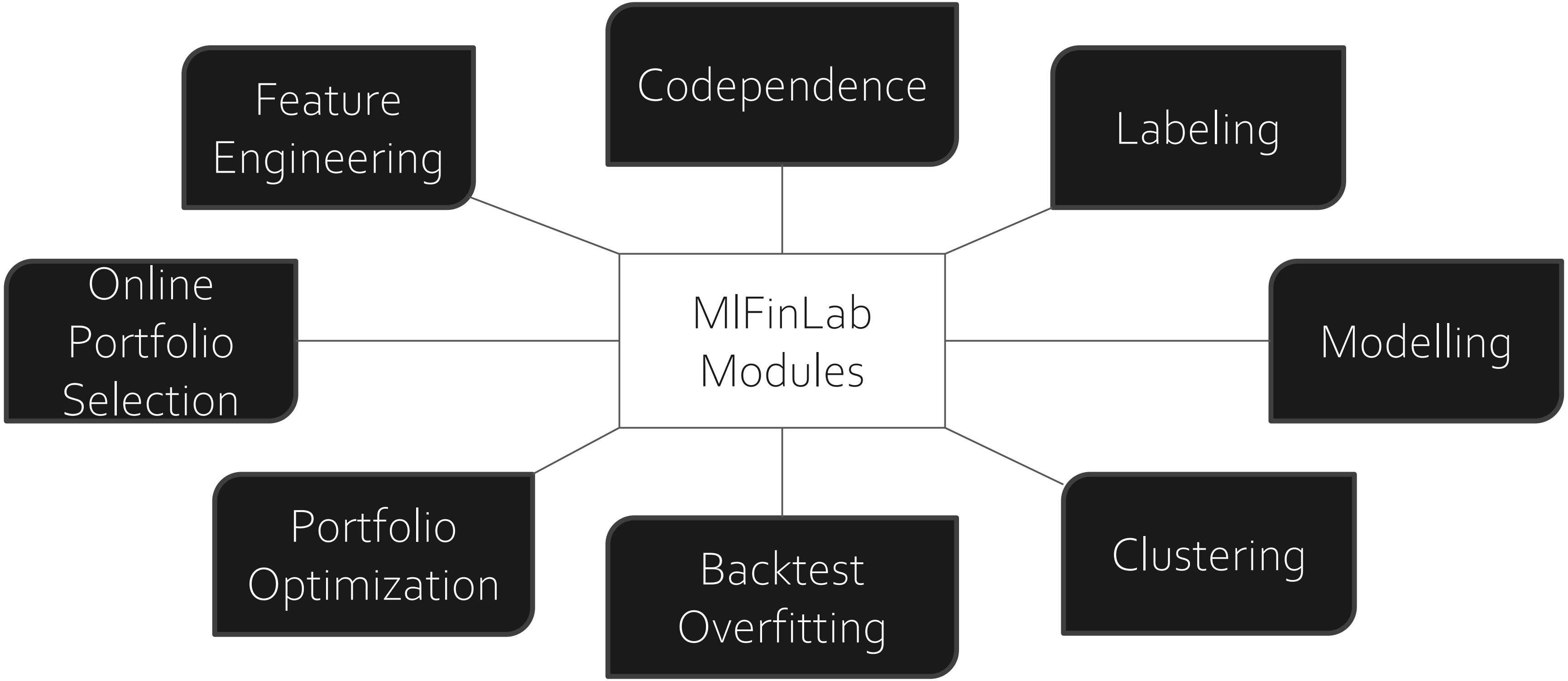
- Implementing algorithms since 2018
- Top 6-th algorithmic-trading package on GitHub

mlfinlab

MlFinlab helps portfolio managers and traders who want to leverage the power of machine learning by providing reproducible, interpretable, and easy to use tools.

Python 1.5k 403

github.com/hudson-and-thames/mlfinlab


MlFinLab Modules
Feature Engineering
Codependence
Labeling
Online Portfolio Selection
Modelling
Portfolio Optimization
Backtest Overfitting
Clustering

Examples of use

45+ Notebooks to try implemented functions

research

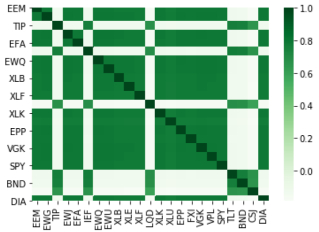
Notebooks based on financial machine learning.

Jupyter Notebook 666 240

github.com/hudson-and-thames/research

In [7]:
```python
# Plotting the heatmap of the Theory-implied correlation matrix
sns.heatmap(etf_tic, cmap="Greens")
```

Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x133fdfe27c8>



We can see that the Theory-implied correlation matrix is less noisy and has a clearly defined structure in comparison to the Empirical correlation matrix.

If we want to measure the similarity of the Empirical correlation matrix and the Theory-implied correlation matrix, we can use the correlation matrix distance introduced by *Herdin* and *Bonek* in a paper **AMIMO Correlation Matrix based Metric for Characterizing Non-Stationarity** available here. The distance is calculated as:

$$d[\Sigma_1, \Sigma_2] = 1 - \frac{tr(\Sigma_1 \Sigma_2)}{\|\Sigma_1\|_f \|\Sigma_2\|_f}$$

Where Σ_1, Σ_2 are the two correlation matrices and the $\|.\|_f$ is the Frobenius norm.

"The distance $d[\Sigma_1, \Sigma_2]$ measures the orthogonality between the considered correlation matrices. It becomes zero if the correlation matrices are equal up to a scaling factor, and one if they differ to a maximum extent".

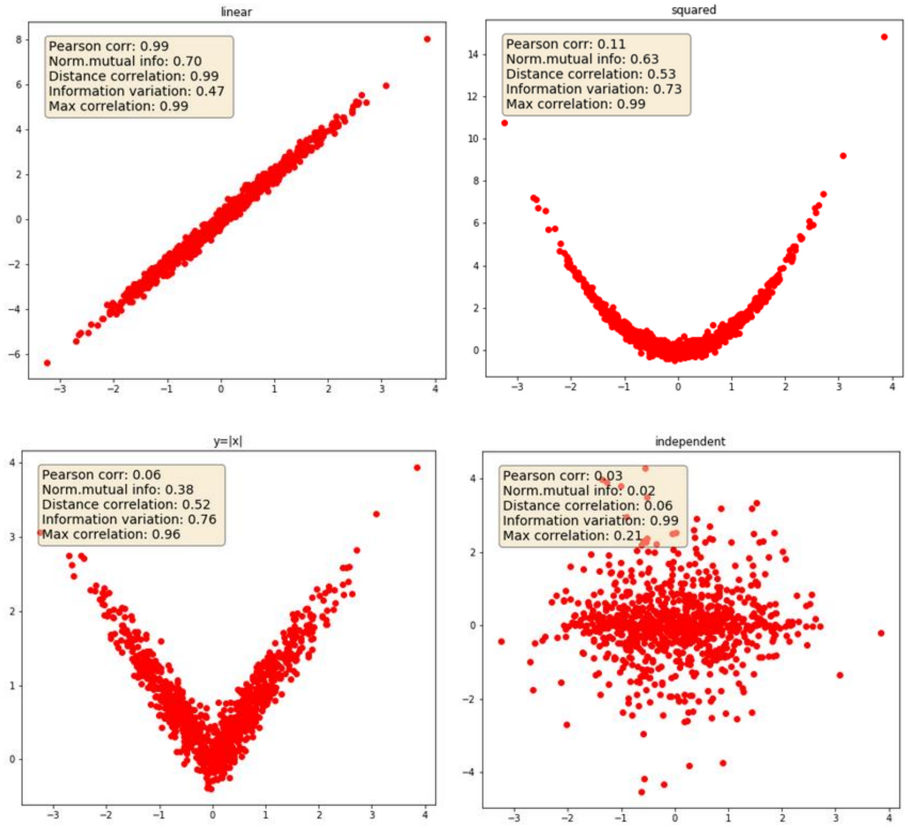
In [8]:
```python
# Calculating the correlation matrix distance
distance = tic.corr_dist(etf_corr, etf_tic)

# Printing the result
print('The distance between empirical and the theory-implied correlation matrices is' , distance)
```

```
The distance between empirical and the theory-implied correlation matrices is 0.035661302090136515
```

The correlation matrices are different but are not too far apart. This shows that the theory-implied correlation matrix blended theory-implied views with empirical ones.

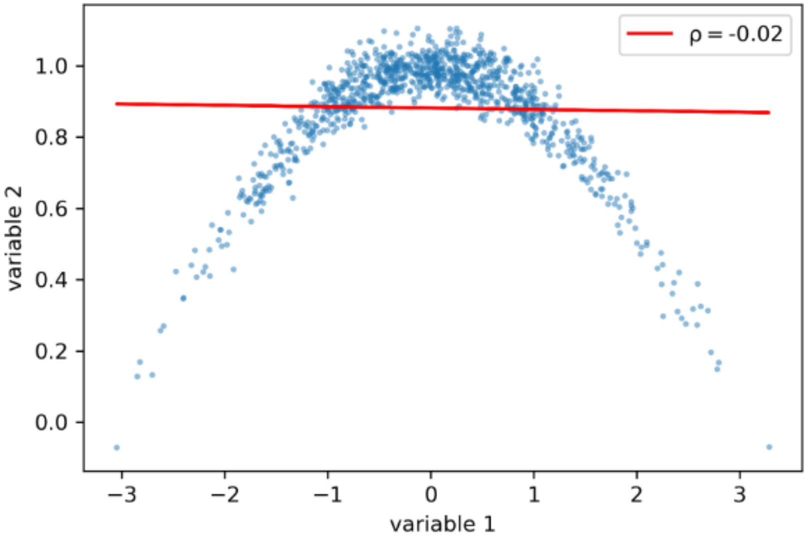
Codependence Module

- Pearson’s Correlation
- Distance correlation
- Angular distance
- Information-Theoretic Codependence
- GPR and GNPR distances



Pearson’s correlation

- Measures linear codependency neglecting non-linear relationships.
- Correlation is highly influenced by outliers.
- Correlation is typically meaningless unless the two variables follow a bivariate Normal distribution.



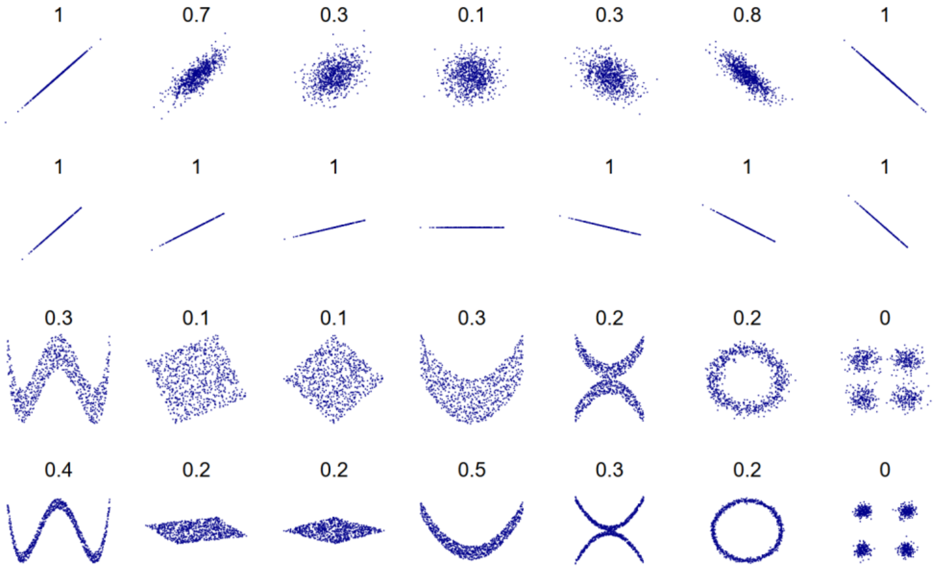
According to Lopez de Prado:

“Correlation is a flawed measure of financial codependence. Many financial relationships are non-linear, and correlation fails to recognize them”

HUDSON
AND THAMES

Distance correlation

A non-linear generalization of Pearson’s correlation introduced in 2005 by Gábor Székely.



Properties:

- $0 \leq \rho_{dist}(X, Y) \leq 1$
- $\rho_{dist}(X, Y) = 0 \iff X$ and Y are independent
- Computationally expensive at $O(n^2)$ vs $O(n)$ for Pearson’s correlation
- Has analogs for the ordinary moments: Distance variance, Distance standard deviation, and Distance covariance.

According to Lopez de Prado:

“Distance covariance can be interpreted as the average Hadamard product of the doubly-centered Euclidean distance matrices of X, Y ”

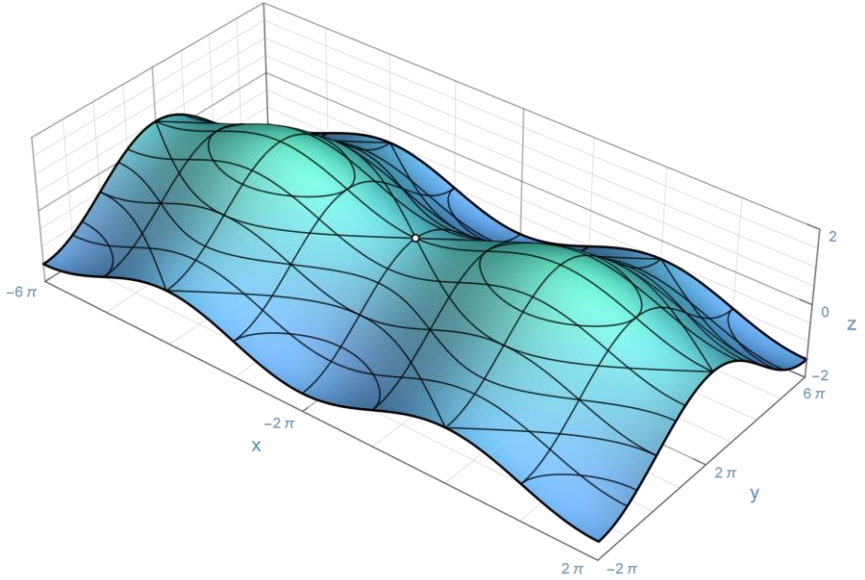
Angular distance

- Correlation is that it's not a metric.

  - non-negativity: $\rho(X,Y) \geq 0$
  - identity of indiscernibles: $\rho(X,Y) = 0 \iff X = Y$
  - symmetry: $\rho(X,Y) = \rho(Y,X)$
  - subadditivity: $\rho(X,Z) \leq \rho(X,Y) + \rho(X,Y)$

- Problem? Incoherence example:

  "For instance, the difference between correlations (0.9,1.0) is the same as (0.1,0.2), even though the former involves a greater difference in terms of codependence". – Lopez de Prado



According to Lopez de Prado:

"Metric functions are important because they induce an intuitive topology on a given set, which allows us to apply notions of "proximity" or "similarity"

Angular distance

Defined metrics based on correlation:

- Angular distance:

$$\rho_d(X,Y) = \sqrt{\frac{1}{2} - \big(1 - \rho(X,Y)\big)}$$

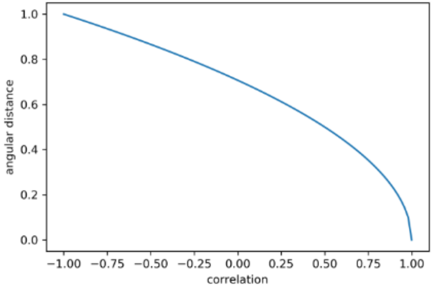
- Absolute angular distance:

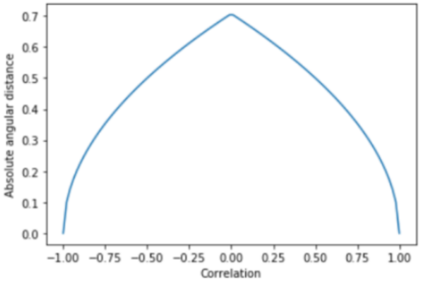
$$\rho_{|d|}(X,Y) = \sqrt{\frac{1}{2} - \big(1 - |\rho|(X,Y)\big)}$$

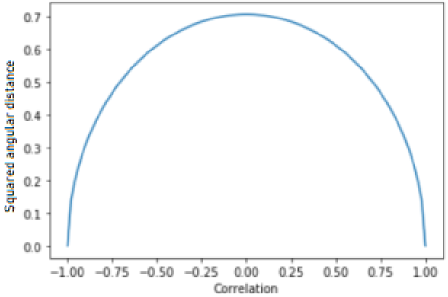
- Squared angular distance:

$$\rho_{d^2}(X,Y) = \sqrt{\frac{1}{2} - \big(1 - \rho^2(X,Y)\big)}$$

Proof available in [Lopez de Prado[2016]]







According to Lopez de Prado:

"In some financial applications, it makes more sense to apply a modified definition of angular distance, such that the sign of the correlation is ignored"

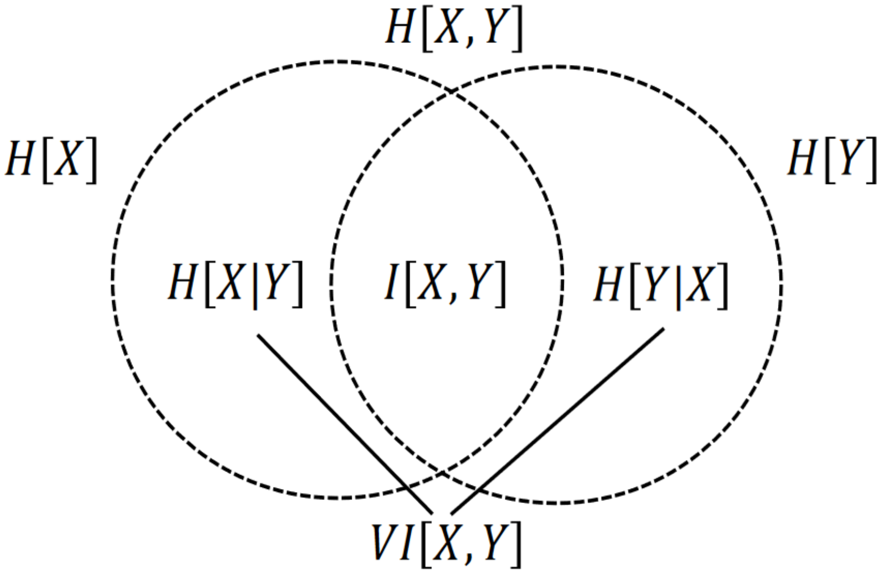
(*long-only portfolios vs* long-short portfolios)

Information-Theoretic Codependence

Measures from Information Theory:

- Entropy: $H(X) = -\sum_{x \in S_x} p(x) \log(p(x))$
- Joint entropy: $H(X,Y) = -\sum_{x,y \in S_x \times S_y} p(x,y) \log(p(x,y))$
- Conditional entropy: $H(X|Y) = H(X,Y) - H(Y)$
- Mutual information: $\mathrm{I}(X,Y) = H(X) - H(X|Y)$
- Variation of information: $\mathrm{VI}(X,Y) = H(X|Y) + H(Y|X)$

Reference and more information in [Lopez de Prado[2020]]



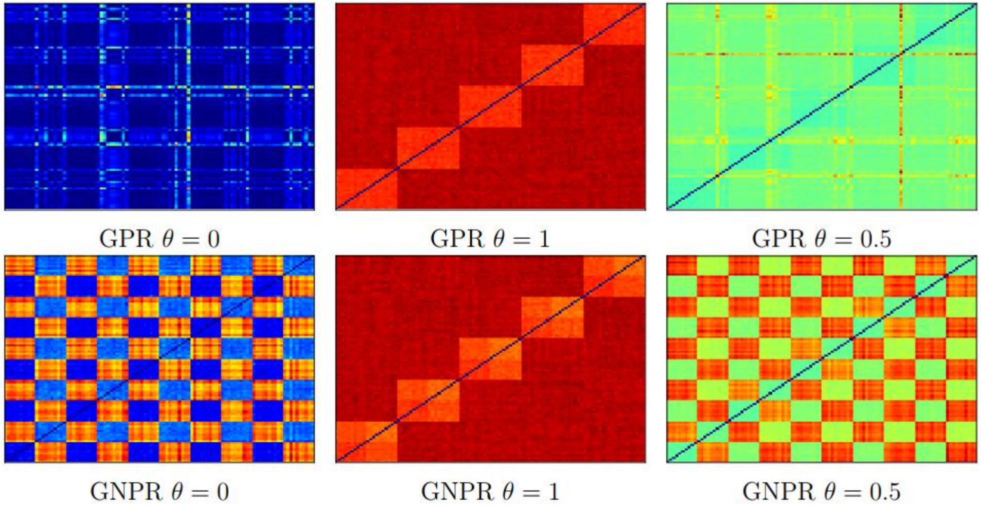
A picture from [Lopez de Prado[2020]]

GPR and GNPR distances

A novel distance described in the work by [Gautier Marti[2017]](). This distance allows to discriminate random variables both on distribution and dependence:

- **Distributional distances** focus on dissimilarity between probability distributions and quantify divergences in marginal behaviours.

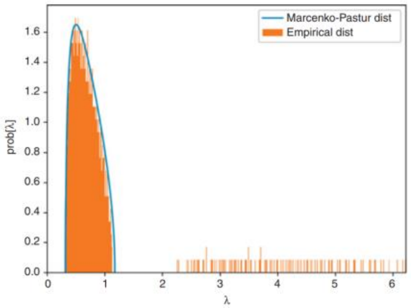
- **Dependence distances**, such as the distance correlation or copula-based kernel dependency measures focus on the joint behaviours of random variables, generally ignoring their distribution properties.

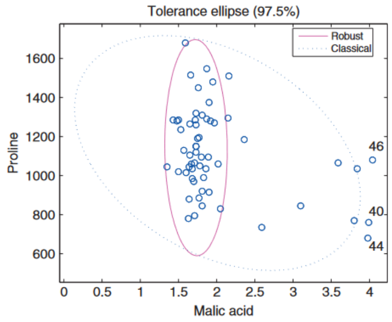


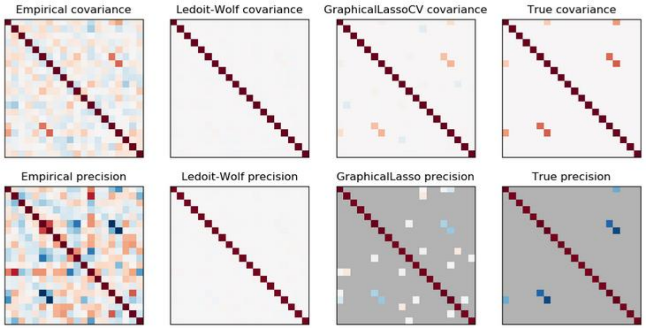
- *Both* $GPR(\theta = 1)$ *and* $GNPR(\theta = 1)$ *highlight the 5 correlation clusters*
- *Only* $GNPR(\theta = 0)$ *finds the 2 distributions subdividing them*
- $GNPR\ (\theta = 0.5)$ *can highlight the 10 original clusters*
- $GPR(\theta = 0.5)$ *simply adds noise on the correlation distance matrix it recovers*

Risk Estimators Submodule

- Minimum Covariance Determinant
- Covariance Estimator with Shrinkage
- Semi-Covariance Matrix
- De-noising and De-toning Covariance Matrix



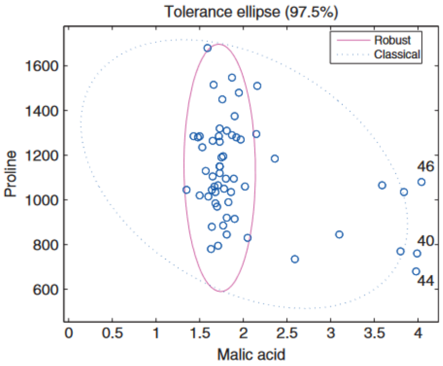




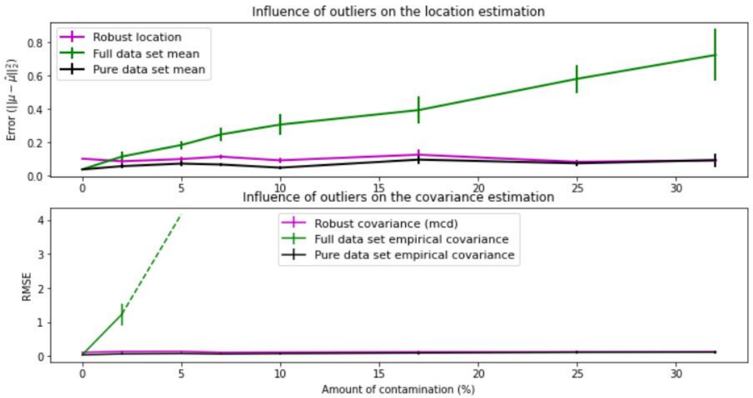
Minimum Covariance Determinant

- The MCD algorithm adjusts covariance to omit estimation errors caused by outliers in the data. This estimator was introduced by P.J. Rousseeuw.

- Main idea behind the MCD is to determine a subset of non-outliers and compute their empirical covariance matrix.



Hubert, Mia & Debruyne, Michiel. (2010)



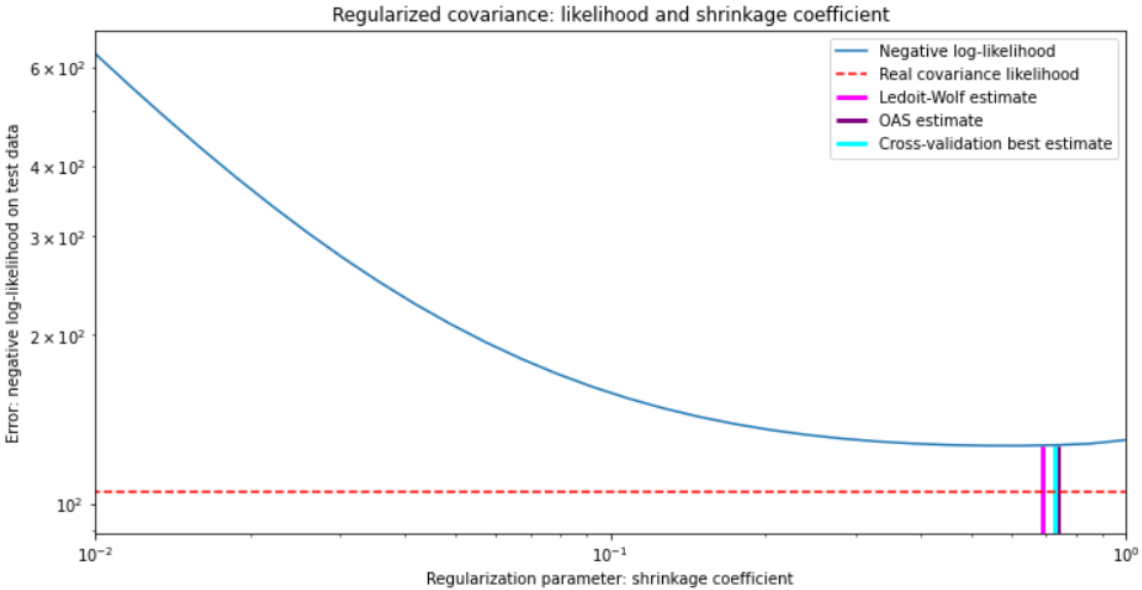
Reproduced with scikit-learn MCD code snippet.

Covariance Estimator with Shrinkage

- In some situations due to numerical reasons the estimated covariance matrix cannot be inverted. Shrinkage is used to avoid this problem.

- Main idea behind shrinkage is to reduce the ratio between the smallest and the largest eigenvalues of the empirical covariance matrix.

$$\sum\nolimits_{cov\_shrunk} = (1-\alpha)\sum\nolimits_{cov} + \alpha \frac{Tr\sum_{cov}}{p} Id$$



Reproduced with scikit-learn LedoitWolf vs OAS code snippet.

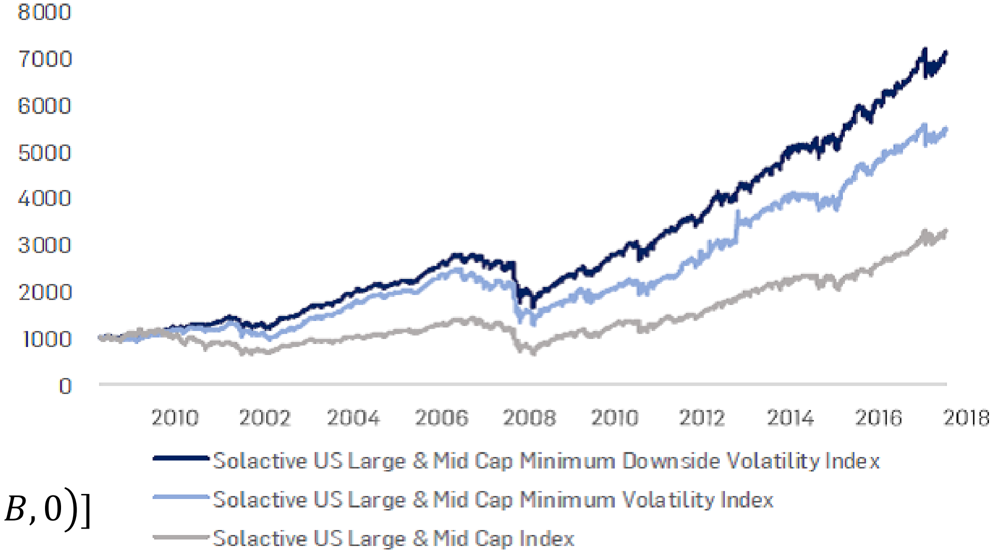
Semi-Covariance Matrix

- Semi-Covariance matrix is a measure of the downside volatility of a portfolio. This metric also allows measuring the volatility of returns below a specific threshold.

- This tool is used to minimize negative returns. It was mentioned in the works of Markowitz, however is still not widely used.

$$SemiCov_{i,j} = \frac{1}{T}\sum_{t=1}^{T}[Min(R_{i,t} - B, 0) * Min(R_{j,t} - B, 0)]$$

Figure 1. Simulated Historical Performance: Solactive US Large & Mid Cap Minimum Downside Volatility Index and Minimum Volatility Index (1999-2018)



Solactive Research Paper. (2017)

De-noising and De-toning Covariance Matrix

- **De-noising** is an alternative to shrinkage that discriminates between eigenvalues that are associated with noise and eigenvalues associated with signal components

- **De-toning** is based on removing the eigenvector associated with the market component. According to Lopez de Prado:

  “By removing the market component, we allow a greater portion of the correlation to be explained by components that affect specific subsets of the securities. It is similar to removing a loud tone that prevents us from hearing other sounds”

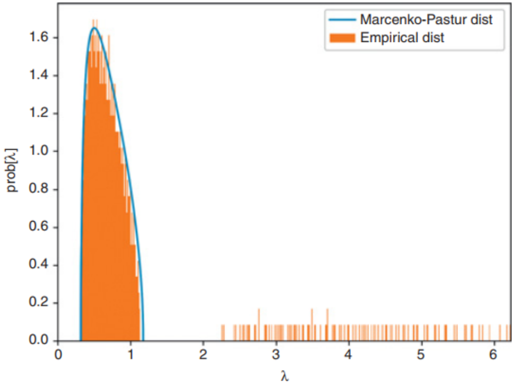


Figure 2.2 Fitting the Marcenko–Pastur PDF on a noisy covariance matrix.

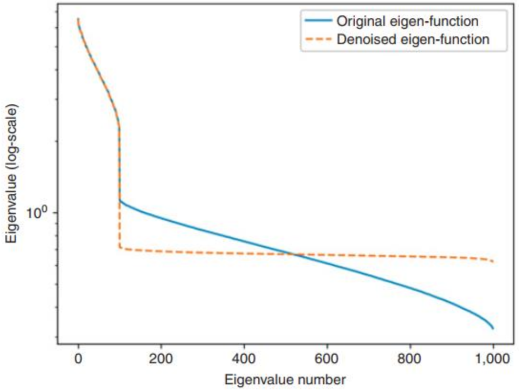


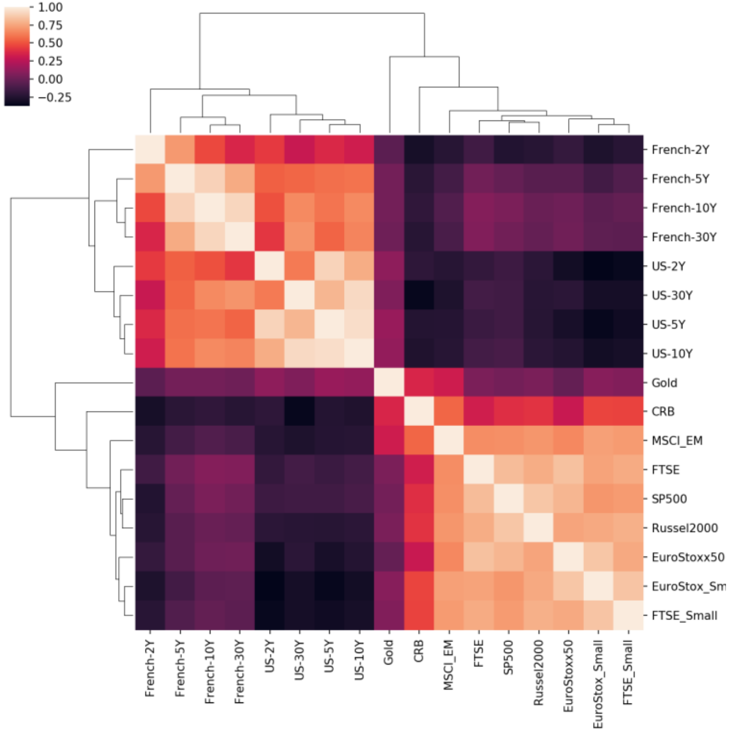
Figure 2.3 A comparison of eigenvalues before and after applying the residual eigenvalue method.

Lopez de Prado. (2020)

Applications in MlFinLab

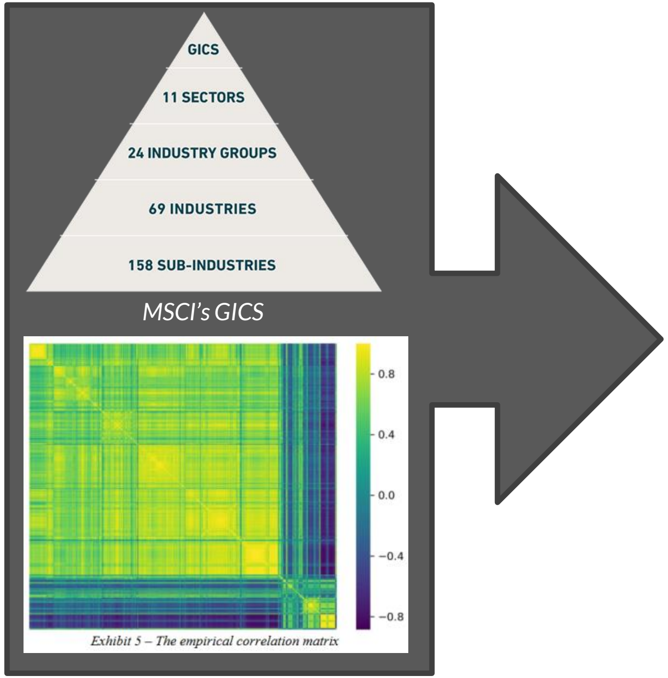
- Optimal Number of Clusters (ONC)
- Portfolio Optimization (MVO,CLA, HRP, HERC)
- Theory-Implied Correlation (TIC)

Correlation Matrix of Asset Returns



Theory-Implied Correlation (TIC)

- The TIC algorithm is aiming to estimate a forward-looking correlation matrix based on economic theory. The method is using a theoretical classification of assets (hierarchical structure) and fits the empirical correlation matrix to the theoretical structure.



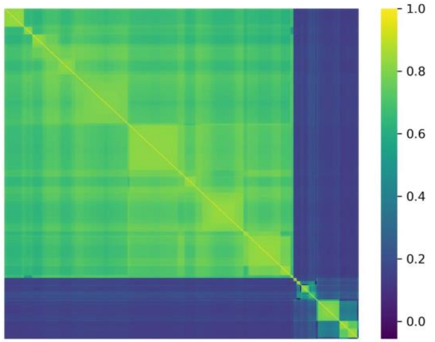


Exhibit 4 – The de-noised TIC matrix

Lopez de Prado. (2019)

"A problem of empirical correlation matrices is that they are purely observation driven, and do not impose a structural view of the investment universe, supported by economic theory" – Lopez de Prado

References I

- **Codependence presentation:** López de Prado, Marcos, Codependence (Presentation Slides) (January 2, 2020). doi: http://dx.doi.org/10.2139/ssrn.3512994
- **Angular distances:** López de Prado, Marcos, Building Diversified Portfolios that Outperform Out-of-Sample (May 23, 2016). Journal of Portfolio Management, 2016; doi: https://doi.org/10.3905/jpm.2016.42.4.059
- **GPR and GNPR distances:** Marti, Gautier. (2017). Some contributions to the clustering of financial time series and applications to credit default swaps. ResearchGate: https://www.researchgate.net/publication/322714557
- **Minimum Covariance Determinant:** Hubert, Mia & Debruyne, Michiel. (2010). Minimum covariance determinant. Wiley Interdisciplinary Reviews: Computational Statistics. 2. 36 - 43. https://doi.org/10.1002/wics.61
- **Covariance Estimator with Shrinkage:** O. Ledoit, M. Wolf. (2001). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis Volume 88, Issue 2, February 2004, Pages 365-411. doi: https://doi.org/10.1016/S0047-259X(03)00096-4

References II

- **Covariance Estimator with Shrinkage:** Chen, Yilun & Wiesel, Ami & Eldar, Yonina & Hero, Alfred. (2010). Shrinkage Algorithms for MMSE Covariance Estimation. Signal Processing, IEEE Transactions on. 58. 5016 - 5029. doi: https://doi.org/10.1109/TSP.2010.2053029
- **Semi-Covariance Matrix:** Minimum Downside Volatility Indices, Solactive Research Paper. (2017) Link: https://www.solactive.com/wp-content/uploads/2018/04/Solactive_Minimum-Downside-Volatility-Indices.pdf
- **Information-Theoretic Codependence/De-noising and De-toning:** López de Prado, Marcos, Machine Learning for Asset Managers. (2020). doi: https://doi.org/10.1017/9781108883658
- **Theory-Implied Correlation:** López de Prado, Marcos, Estimation of Theory-Implied Correlation Matrices (November 9, 2019). doi: http://dx.doi.org/10.2139/ssrn.3484152


HUDSON
AND THAMES
Thank You