Covariance Matrix Cleaning

Hierarchical methods/Eigenvectors

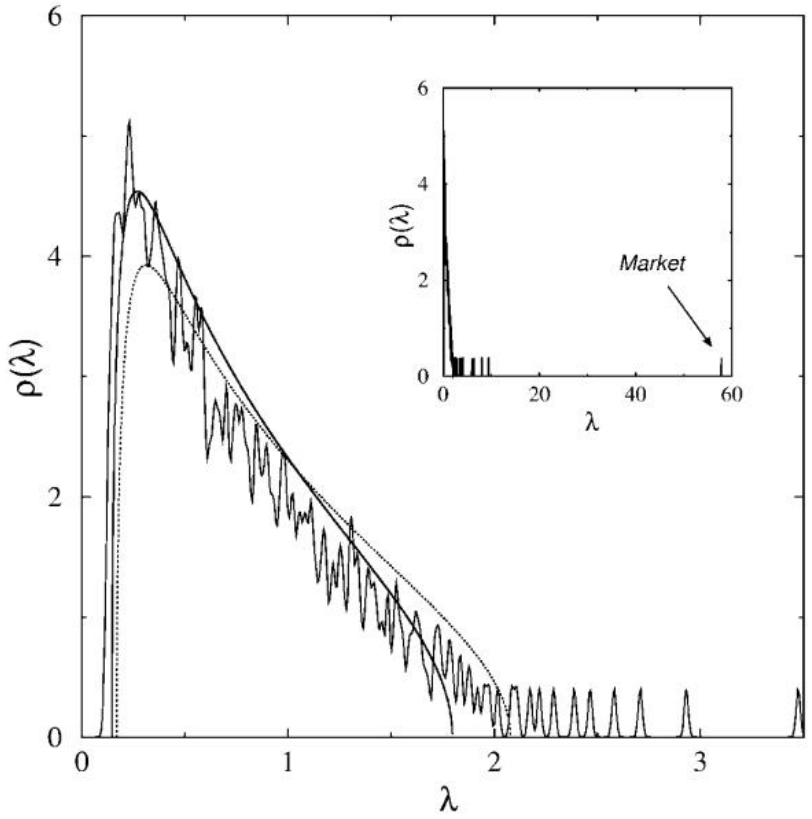
Christian Bongiorno | Allocation de Portefeuille | 02-03-2023

Topics of Today

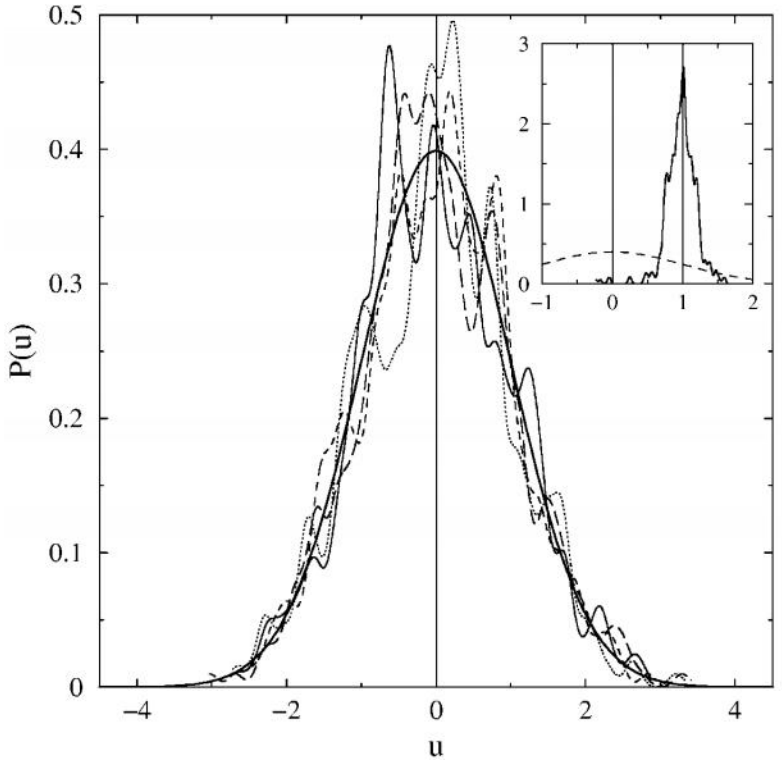
- We are going briefly to discuss the problem you have encountered in the assignments
- We will study a complementary approach to covariance filtering, i.e., Hierarchical methods
- I will show the relationship with eigenvector filtering
- I will discuss the last research results on this topic

Another Look at the Spectrum

Equity eigenvalue distribution



Distribution of eigenvector values



Laloux, L., Cizeau, P., Bouchaud, J. P., & Potters, M. (1999). Noise dressing of financial correlation matrices. Physical review letters, 83(7), 1467

Sectors and Eigenvectors (1)

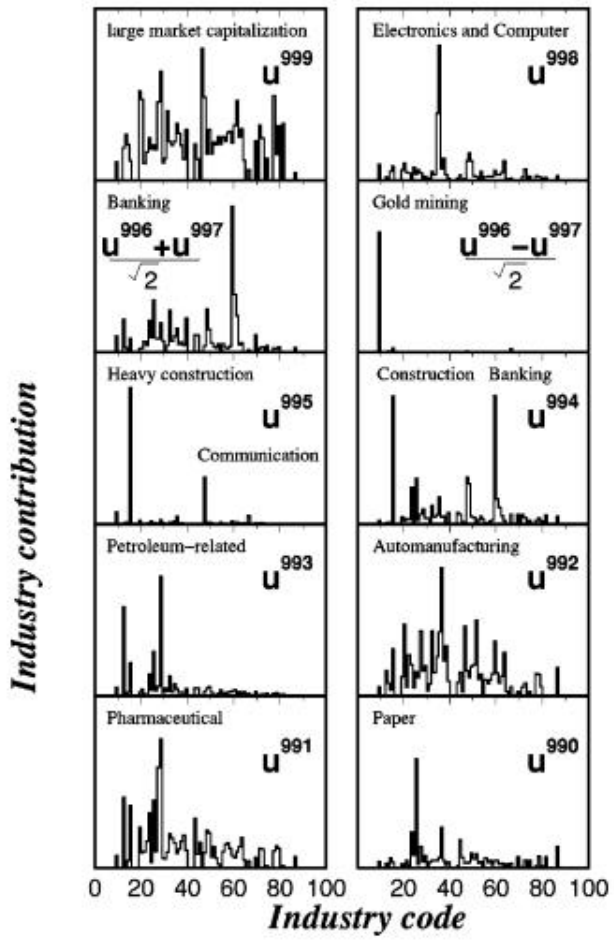


FIG. 1. Contribution X_I^k to industry sector I of eigenvector $\mathbf{u}^k$ for the deviating eigenvectors shows marked peaks at distinct values of SIC code, for all but $\mathbf{u}^{999}$ which contains stocks with large capitalizations as significant contributors.

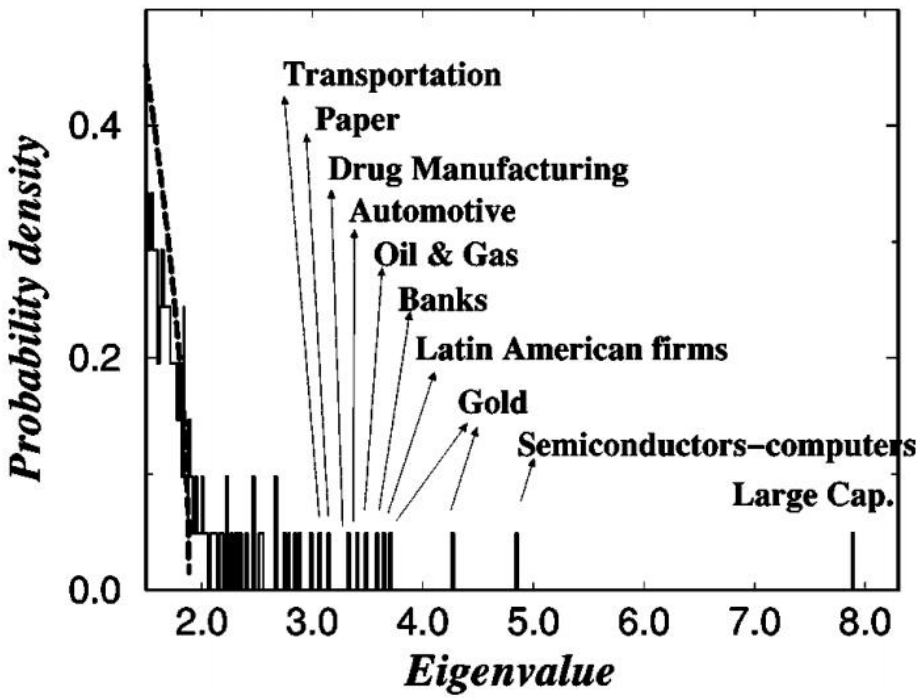
Gopikrishnan et al defined a metric to measure the contribution of the eigenvectors to specific sector codes

$$X_S^k = \sum_{i=1}^{n} P_{Si}\left[u_i^k\right]^2 \qquad P_{Si} = \begin{cases} \dfrac{1}{n_{Si}} & \text{if it belongs to the sector} \\ 0 & \text{if it not belongs to the sector} \end{cases}$$

where u_i^k is the i value of the k eigen-component, and
n_{s_i} is the number of elements of the sector S.

Gopikrishnan, P., Rosenow, B., Plerou, V., & Stanley, H. E. (2001). Quantifying and interpreting collective behavior in financial markets. Physical Review E, 64(3), 035106.

Sectors and Eigenvectors (2)



[nowadays we know that MP bulk even if seems random contains information]

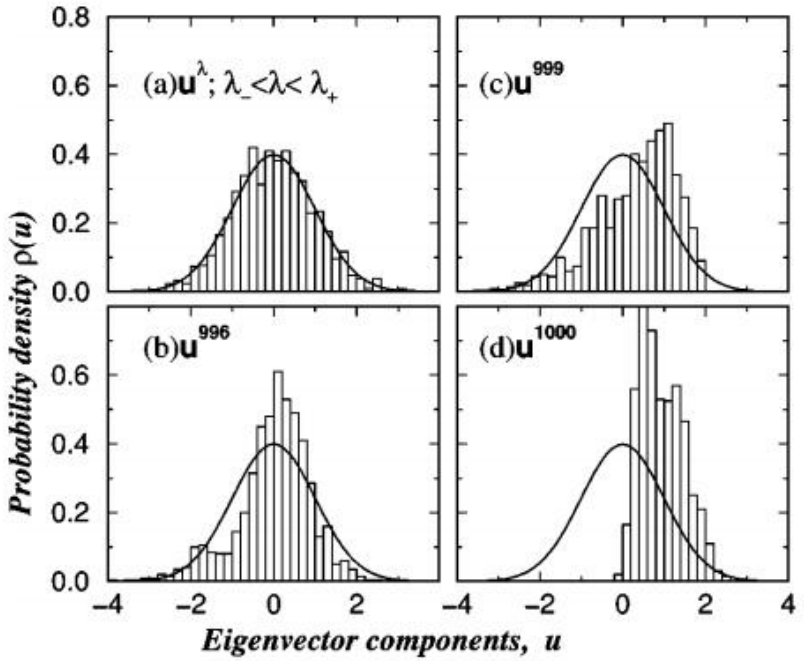


FIG. 8. (a) Distribution $\rho(u)$ of eigenvector components for one eigenvalue in the bulk $\lambda_- < \lambda < \lambda_+$ shows good agreement with the RMT prediction of Eq. (17) (solid curve). Similar results are obtained for other eigenvalues in the bulk. $\rho(u)$ for (b) u^{996} and (c) u^{999}, corresponding to eigenvalues larger than the RMT upper bound λ_+ (shaded region in Fig. 3). (d) $\rho(u)$ for u^{1000} deviates significantly from the Gaussian prediction of RMT. The above plots are for C constructed from 30-min returns for the 2-yr period 1994–1995. We also obtain similar results for C constructed from daily returns.

[1] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., & Stanley, H. E. (2002). Random matrix approach to cross correlations in financial data. Physical Review E, 65(6), 066126.

Hierarchical Structure of Financial Matrix

Eur. Phys. J. B **11**, 193–197 (1999)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences

Hierarchical structure in financial markets

R.N. Mantegna[a]

Istituto Nazionale per la Fisica della Materia, Unità di Palermo, 90128, Palermo, Italy
Dipartimento di Energetica ed Applicazioni di Fisica, Università di Palermo, Viale delle Scienze, 90128, Palermo, Italy

Received 24 March 1999 and Received in final form 28 June 1999

Abstract. I find a hierarchical arrangement of stocks traded in a financial market by investigating the daily time series of the logarithm of stock price. The topological space is a subdominant ultrametric space associated with a graph connecting the stocks of the portfolio analyzed. The graph is obtained starting from the matrix of correlation coefficient computed between all pairs of stocks of the portfolio by considering the synchronous time evolution of the difference of the logarithm of daily stock price. The hierarchical tree of the subdominant ultrametric space associated with the graph provides a meaningful economic taxonomy.

PACS. 02.50.Sk Multivariate analysis – 89.90.+n Other areas of general interest to physicists

Scientific Community Reception

Rosario Nunzio Mantegna

STAI SEGUENDO

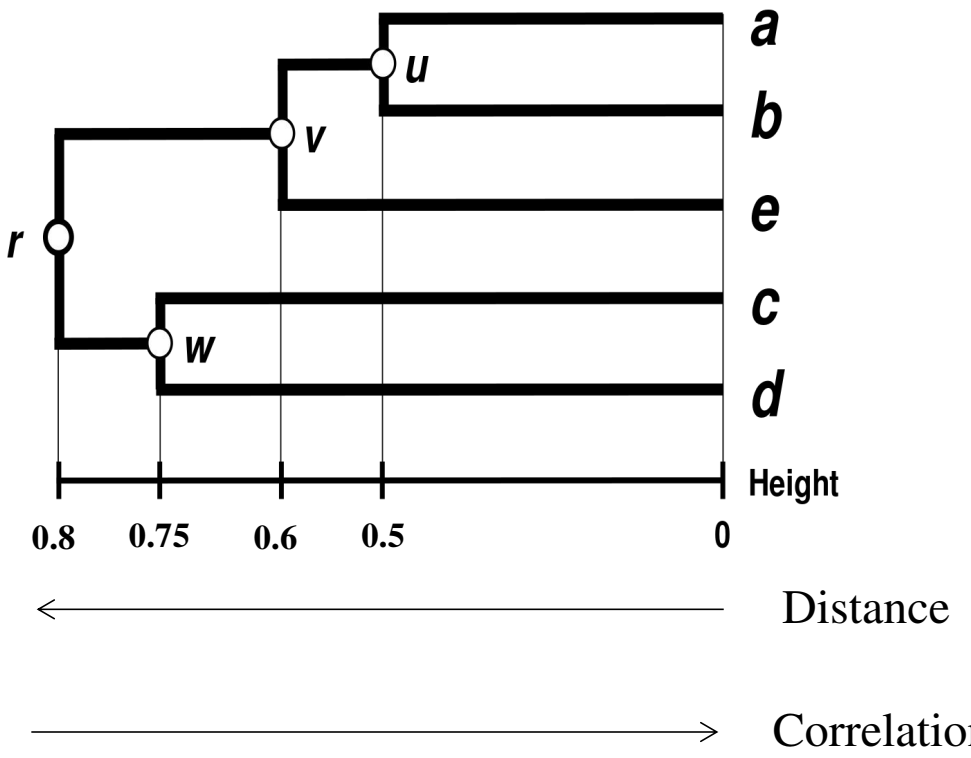
Professor of Applied Physics. Department of Physics and Chemistry, Università degli Studi di Palermo
Email verificata su unipa.it - Home page

Econophysics Statistical physics Complex systems Financial markets Information filtering

TITOLO	CITATA DA	ANNO
Introduction to econophysics: correlations and complexity in finance RN Mantegna, HE Stanley Cambridge university press	5173	1999
Introduction to econophysics: correlations and complexity in finance RN Mantegna, HE Stanley Cambridge university press	~~5085~~	1999
Scaling behaviour in the dynamics of an economic index RN Mantegna, HE Stanley Nature 376 (6535), 46-49	2237	1995
Hierarchical structure in financial markets RN Mantegna The European Physical Journal B-Condensed Matter and Complex Systems 11 (1 ...	1952	1999
Stochastic process with ultraslow convergence to a Gaussian: the truncated Lévy flight RN Mantegna, HE Stanley Physical Review Letters 73 (22), 2946	994	1994
Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis SV Buldyrev, AL Goldberger, S Havlin, RN Mantegna, ME Matsa, ... Physical Review E 51 (5), 5084	699	1995
A tool for filtering information in complex systems M Tumminello, T Aste, T Di Matteo, RN Mantegna Proceedings of the National Academy of Sciences 102 (30), 10421-10426	689	2005
Fast, accurate algorithm for numerical simulation of Levy stable stochastic processes RN Mantegna Physical Review E 49 (5), 4677	600	1994

Hierarchical Clustering

It is an agglomeration algorithm that groups recursively elements based on their distance



A simple way to define a distance from a correlation is:

$$d_{ij} = 1 - c_{ij}$$

Let $\sigma(x)$ the set of elements that cluster x contains

Linkage Rule

Single Linkage:

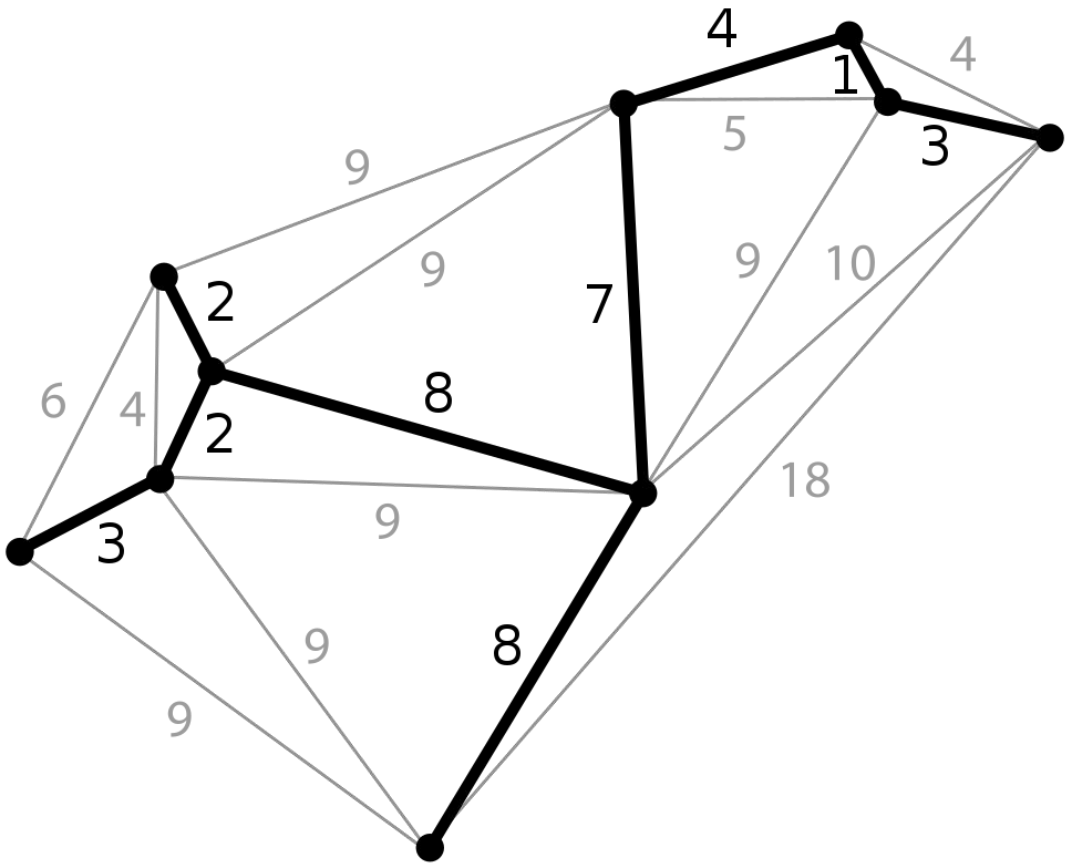
$$\rho_{hk} = min\{ d_{ij} : i \in \sigma(h),\ j \in \sigma(k) \}$$

Average Linkage:

$$\rho_{hk} = \frac{\sum_{i \in \sigma(h)} \sum_{j \in \sigma(k)} d_{ij}}{N_h N_k}$$

Minimum Spanning Tree

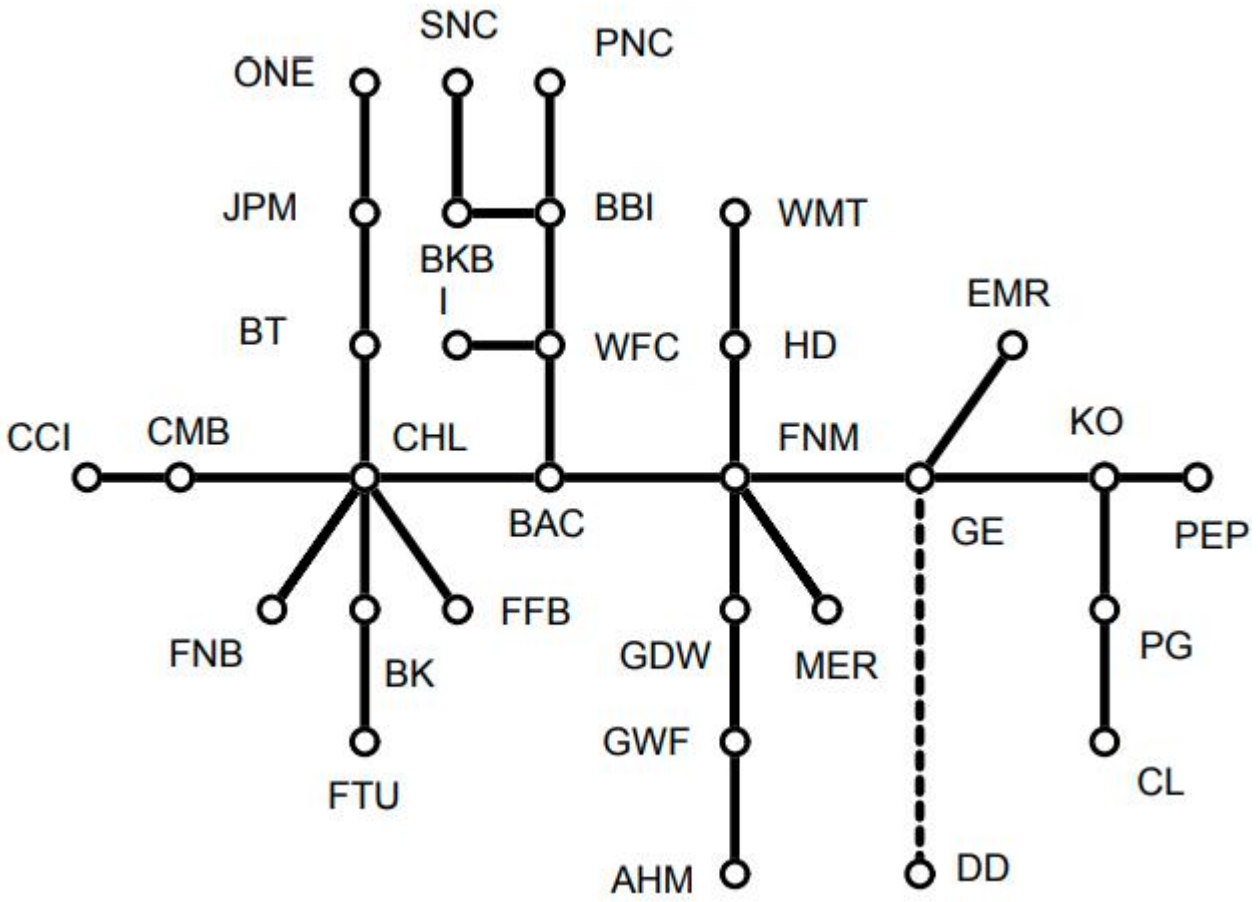
An MST is a subset of the edges of a connected, edge-weighted undirected graph that connects all the vertices together, without any cycles and with the minimum possible total edge weight



It can be shown that retaining only the links with distance smaller than h, the remaining connected components are coincident with the clusters of a single linkage HC after a cutoff at distance h.

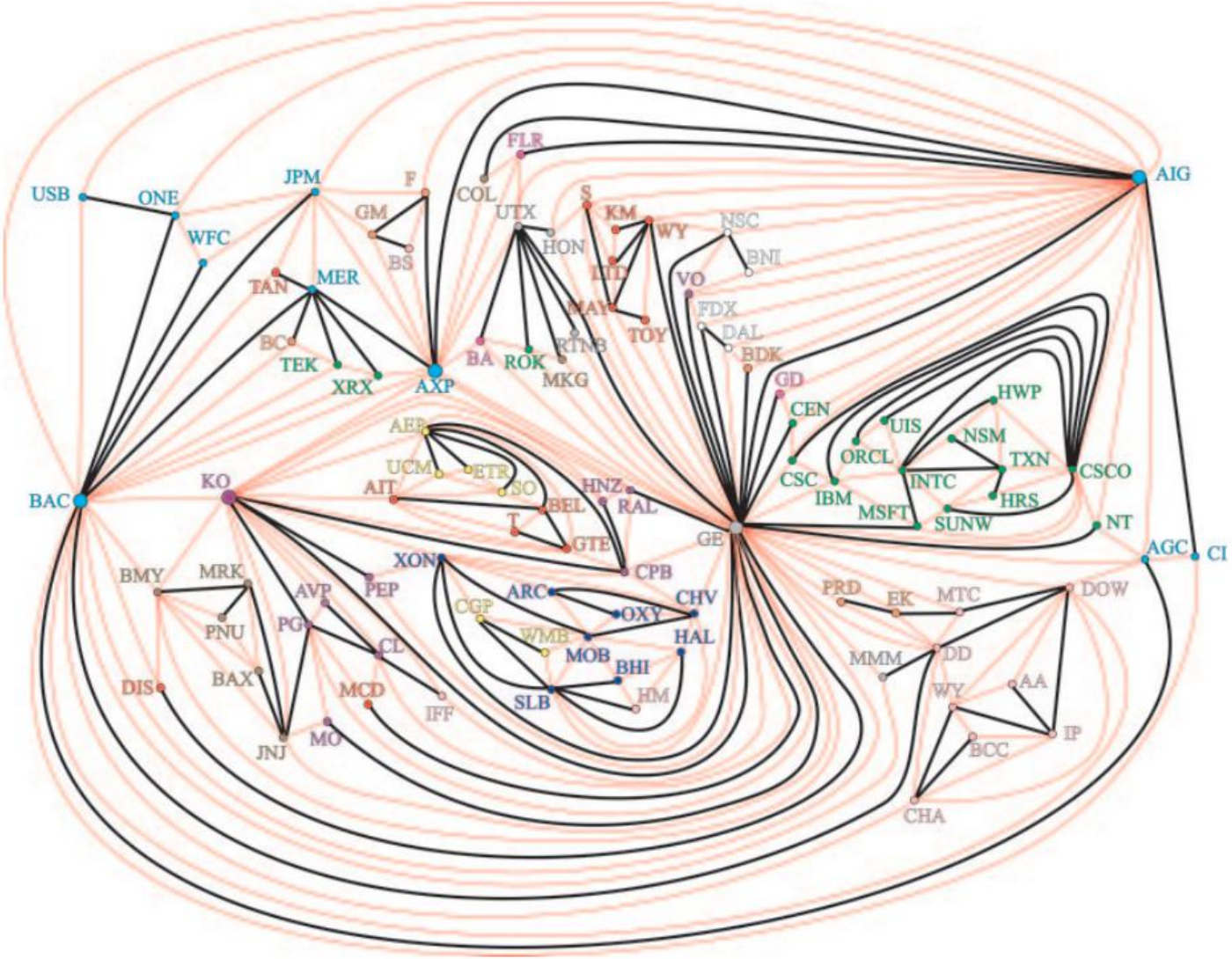
This means that an MST contains more information than a HC

Financial MST



Branches of the Tree are associated with economic sector taxonomy

Planar Graphs and Generalization



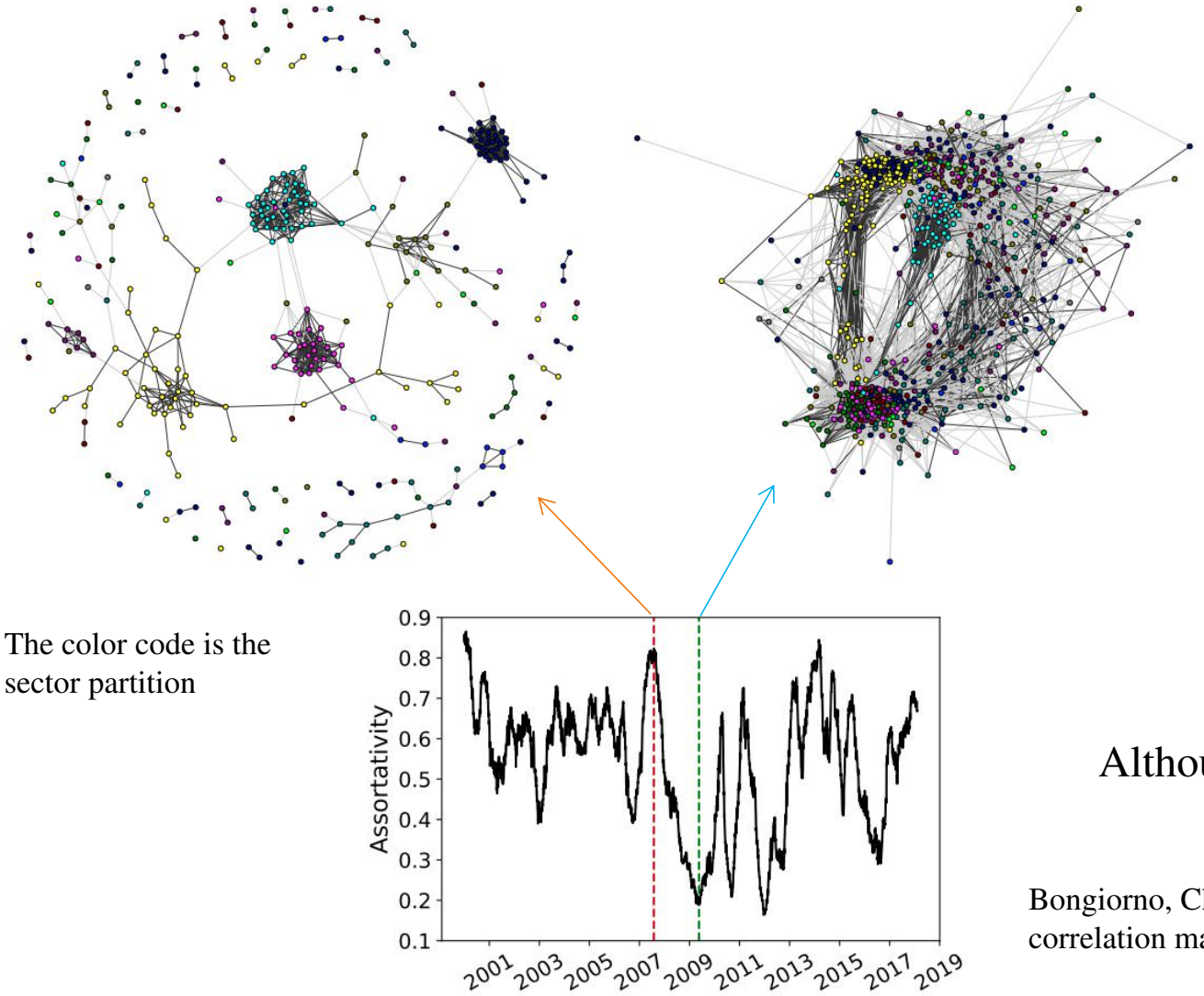
The MST can be generalized planar graph by putting different topological constrains.

Can be show that larger embedding spaces provides less severe filtering

However, this is not the topic of the course, if you are interested you can check the works of Aste.

Tumminello, M., Aste, T., Di Matteo, T., & Mantegna, R. N. (2005). A tool for filtering information in complex systems. Proceedings of the National Academy of Sciences, 102(30), 10421-10426

Temporal Persistence



The color code is the sector partition

We retained those links with a correlation coefficient significantly larger than a single-index factor model.

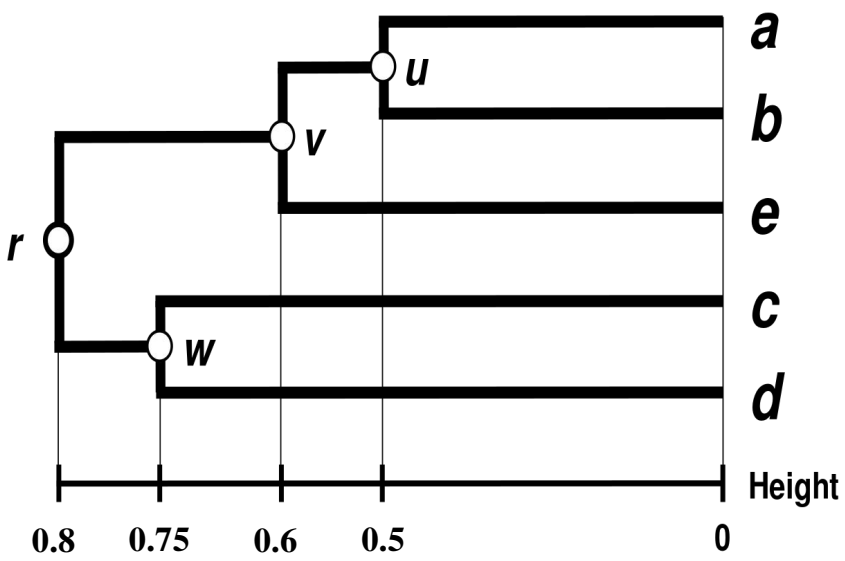
Assortativity measure the propensity to establish links within the same sectors

The assortativity is 1 if we observe links only within the same sector, it is zero if they are randomly distributed with respect to the sector partition.

Although it varies over time a certain persistence is observed.

Bongiorno, Christian, and Damien Challet. "Non-parametric sign prediction of high-dimensional correlation matrix coefficients." EPL (Europhysics Letters) 133.4 (2021): 48001.

HC on Correlations



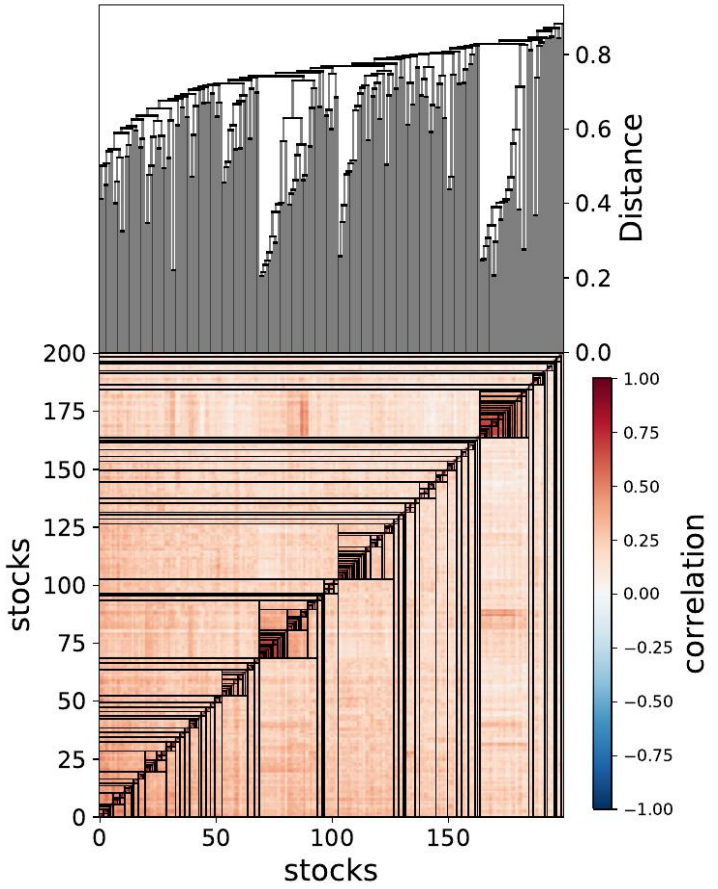
Distance

Correlation

Average Linkage:

$$\rho_{hk} = \frac{\sum_{i \in \sigma(h)} \sum_{j \in \sigma(k)} (1 - c_{ij})}{N_h \, N_k}$$

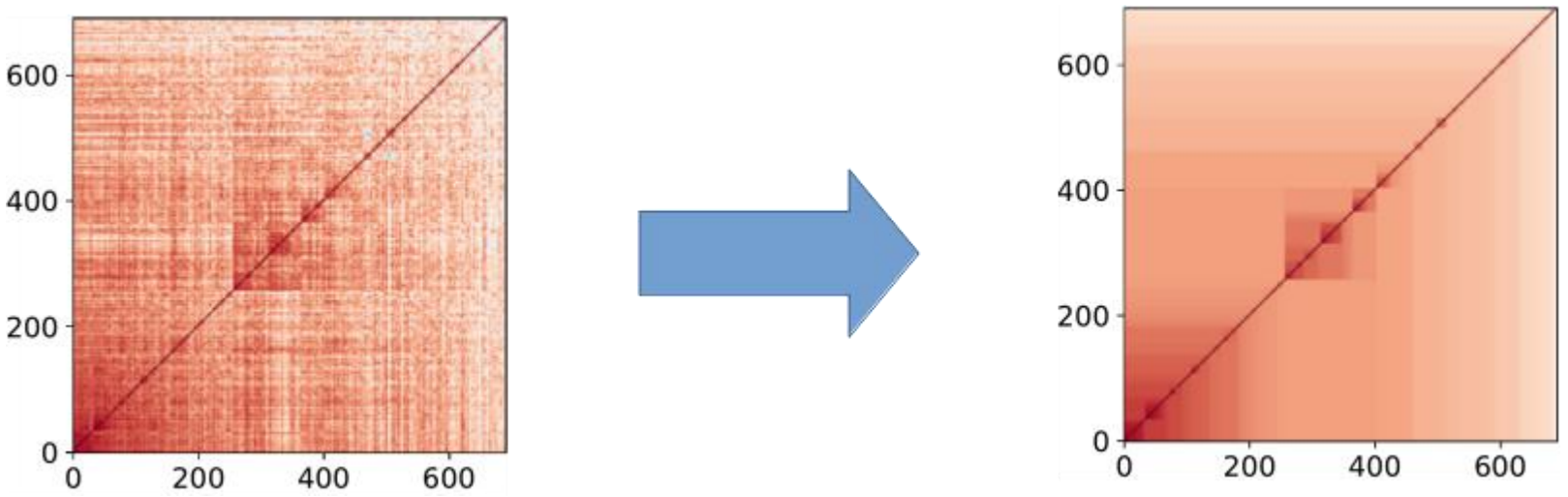
The levels of the HC are associated with the average value of the rectangular sub-matrices among blocks



HC Filtered Correlation Matrix

It is possible to substitute to each rectangular sub-matrix its average value.

Such a matrix is the simplest one, in terms of degree of freedom, that share the same dendrogram with the original matrix.



Factor Decomposition of HC filtered matrix

It is possible to define a factor decomposition that produce the HC filtered correlation matrix $C^{<}$

HC pattern Residual term

$$x_i(t) = \sum_{\alpha_h \in G(i)} \gamma_{\alpha_h} f^{(\alpha_h)}(t) + \eta_i \ \epsilon_i(t),$$

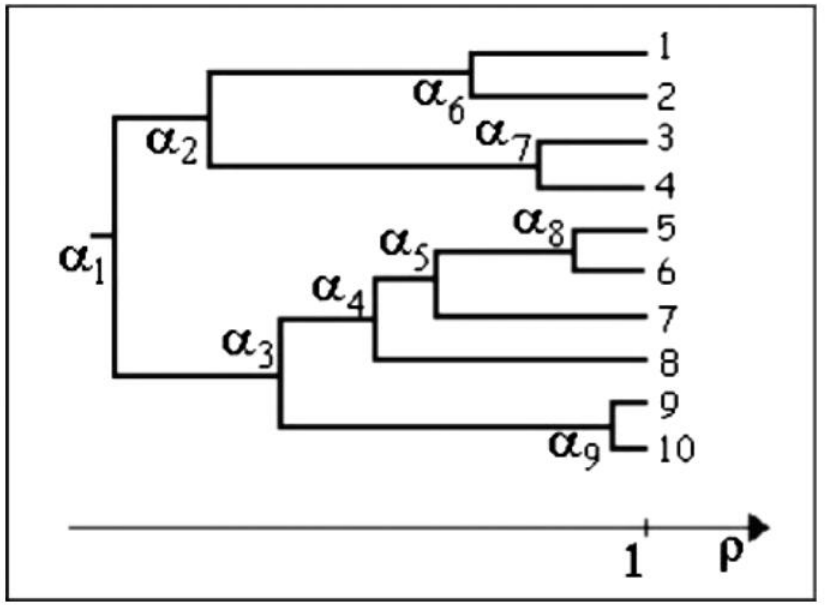
Residual are defined in such a way to guarantees 1 on the diagonal

$$\eta_i = [1 - \textstyle\sum_{\alpha_h \in G(i)} \gamma^2_{\alpha_h}]^{1/2}$$

Factor loading are defined from the difference of the levels between parent and son clusters

$$\gamma_{\alpha_1} = \sqrt{\rho_{\alpha_1}},$$

$$\gamma_{\alpha_h} = \sqrt{\rho_{\alpha_h} - \rho_{g(\alpha_h)}} \quad \forall\, h = 2, \ldots, n-1,$$



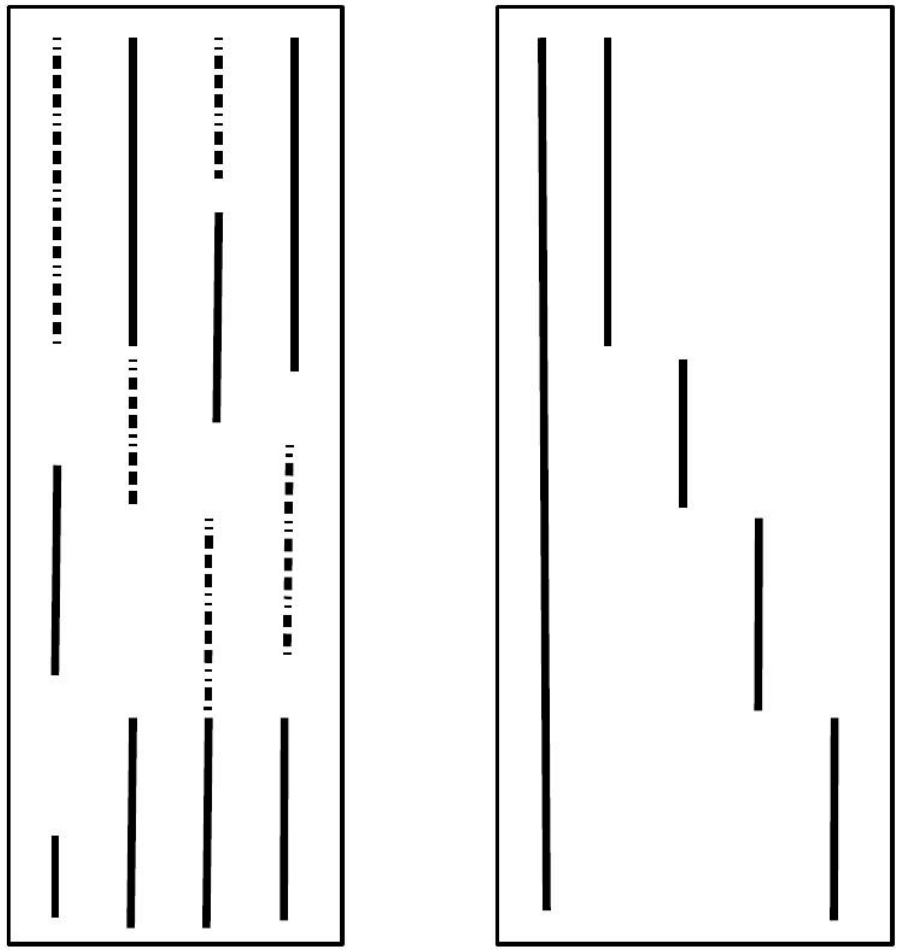
If $\rho_{\alpha_i} > 0$ for all i, then the filtered matrix is positive defined.

Tumminello, M., Lillo, F., & Mantegna, R. N. (2007). Hierarchically nested factor model from multivariate data. EPL (Europhysics Letters), 78(3), 30006

Complexity vs Simplicity

Complex

Simple



The factor loading matrix can have positive zero or negative values

I indicate here positive values as a straight line, negative with dotted lines and zero with empty space.

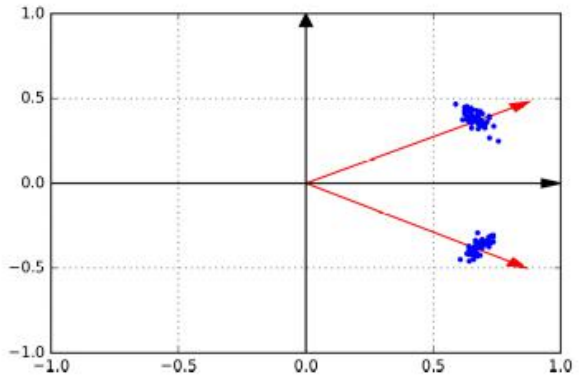
PCA or spectral decomposition (VariMax) address the parsimony principle by penalizing the simplicity

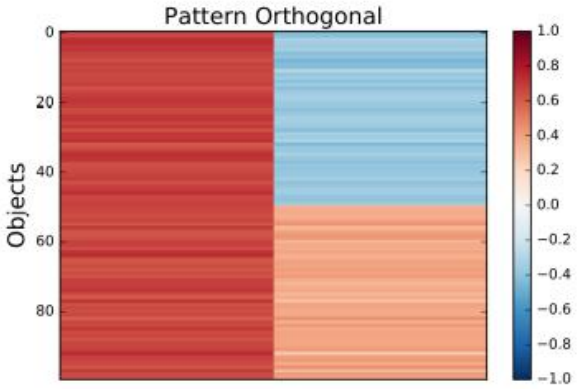
HC pattern is not a parsimonious representation, but it is simple

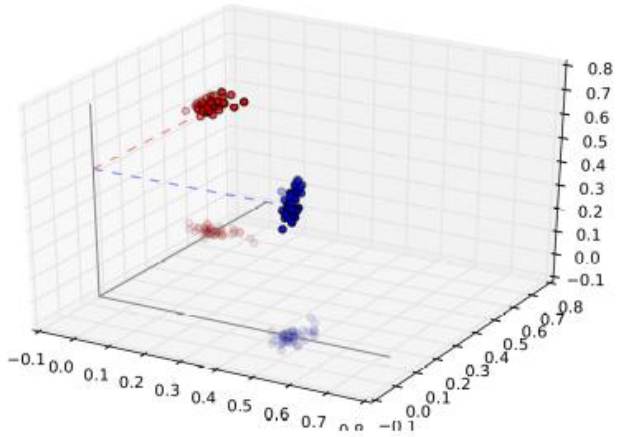
When a factor is composed by only positive elements you can associate it with that subset of elements.
If negative values are present you cannot do it.

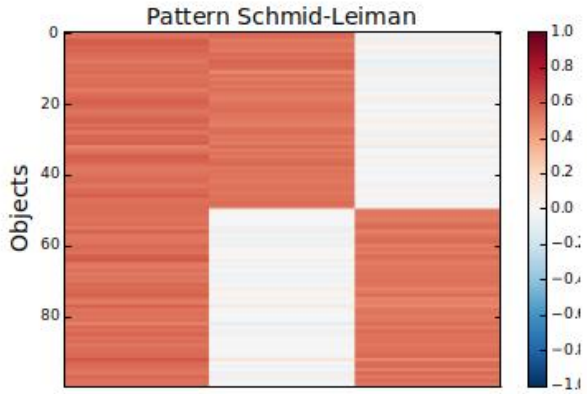
Thurstone, L. L. (1947). Multiple-factor analysis; a development and expansion of The Vectors of Mind

Validation of the Distribution









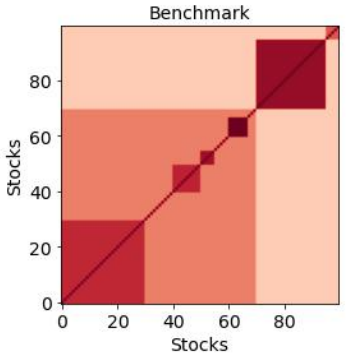
With PCA you are compressing all the information into the first two components, however, the second component cannot be associated to a specific cluster

You can imagine an oblique set of axis as the projection in a bi-dimensional space of a three-dimensional orthogonal set.

The HC approach is less parsimonious but simpler

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. Psychometrika, 22(1), 53-61

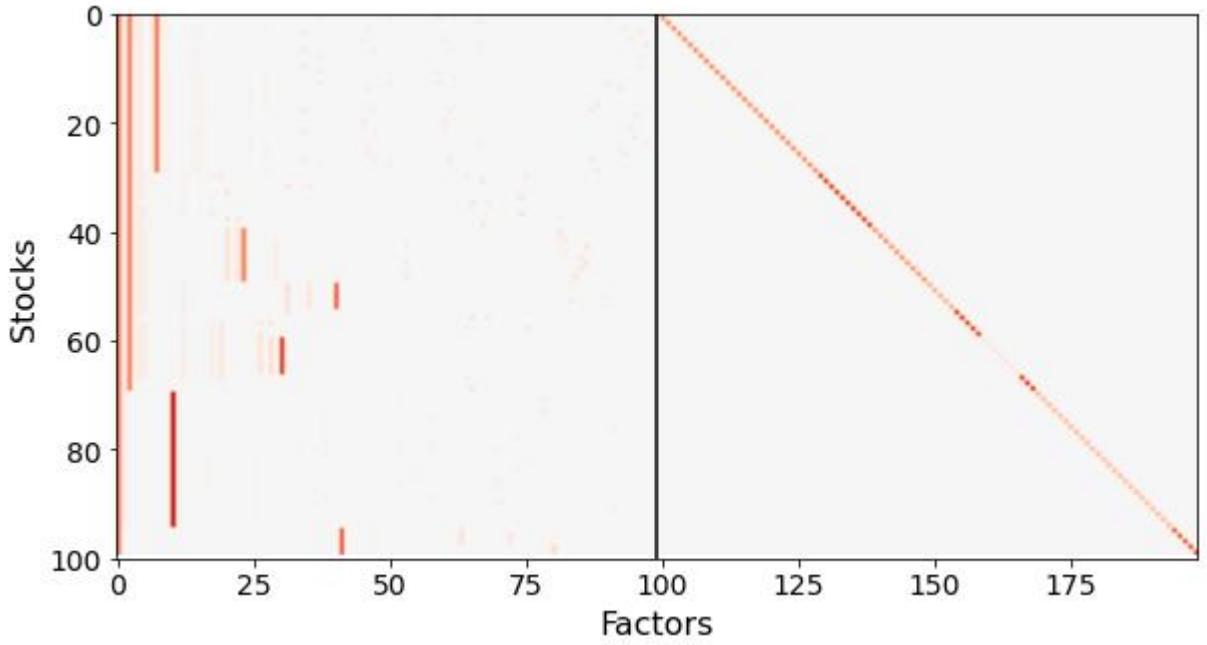
Interpretations



HC factor decomposition requires $2N - 1$ factors. PCA only N

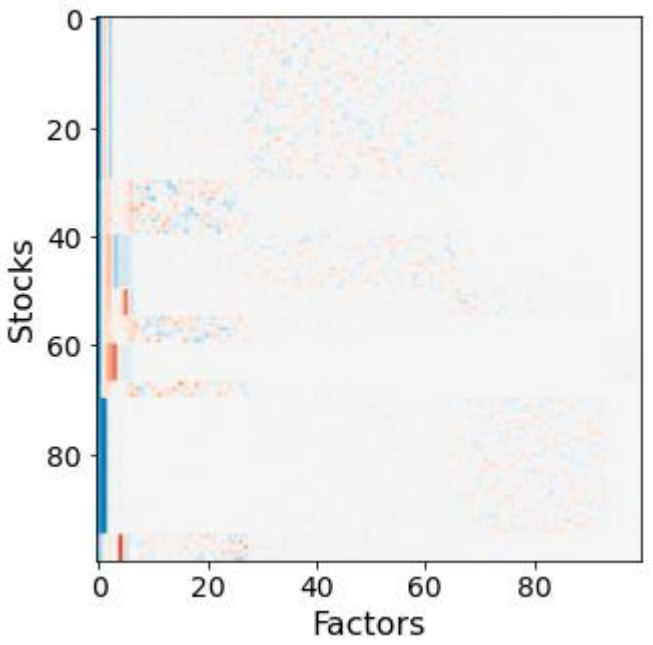
Lower components on PCA are difficult to distinguish from noise.
In HC lower components are associated to the higher correlations, so they are robust.
Higher components are associated to average with many elements so they are robust too

HC pattern [N-1]

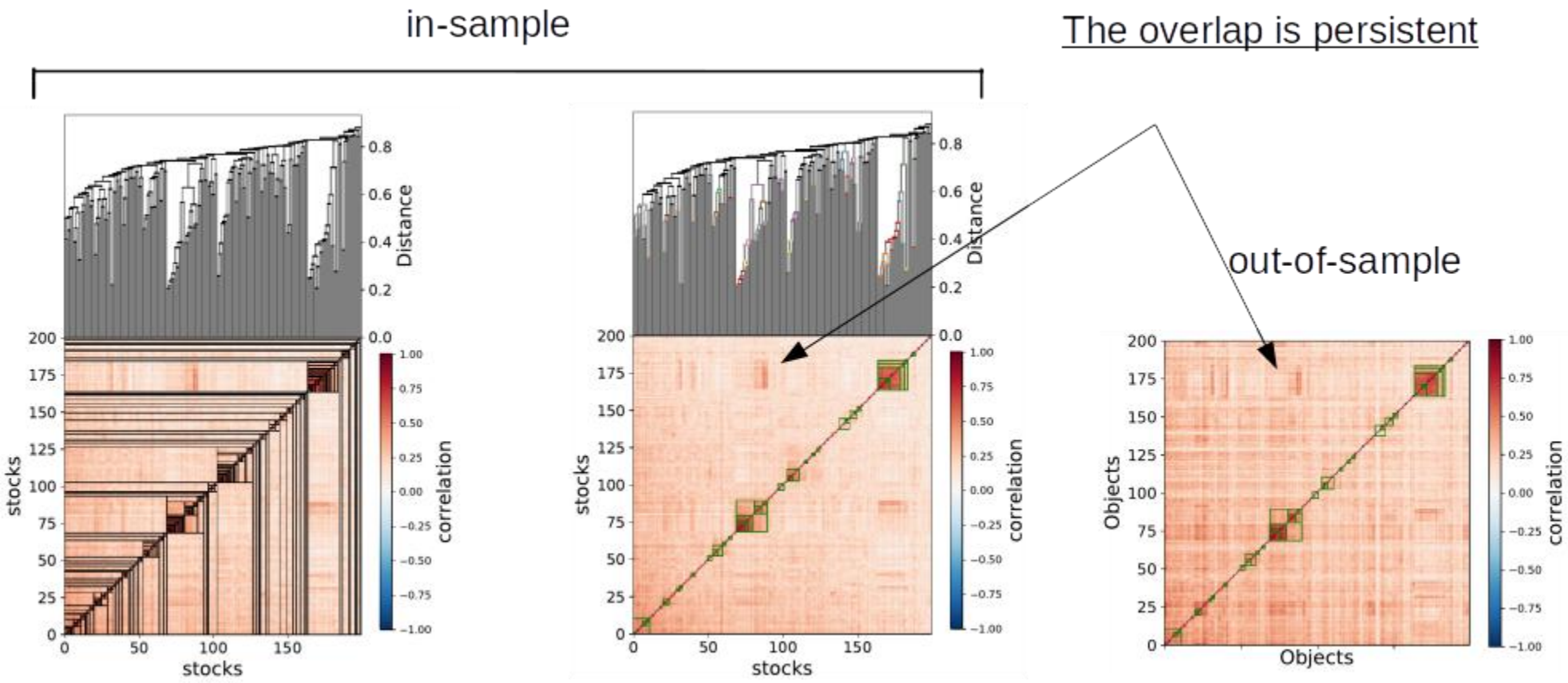


Residuals [N]

PCA pattern [N]

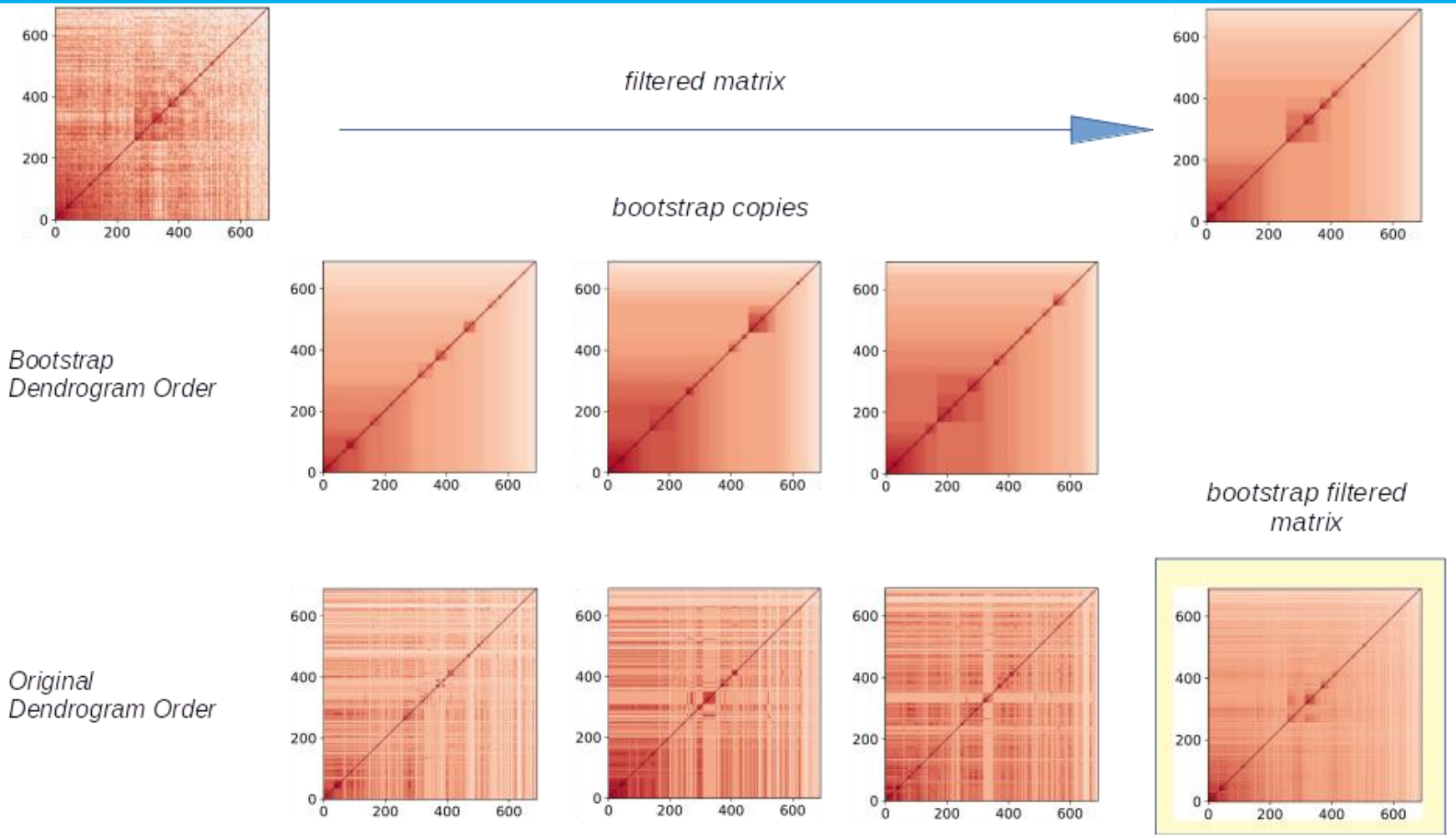


Is Hierarchical Clustering Enough?



Bongiorno, Christian, Salvatore Micciché, and Rosario N. Mantegna. "Statistically validated hierarchical clustering: Nested partitions in hierarchical trees." Physica A: Statistical Mechanics and its Applications (2022): 126933.

Bootstrap Average Hierarchical Clustering (BAHC)



It is an average element-wise of the HC filtering of bootstrap realizations

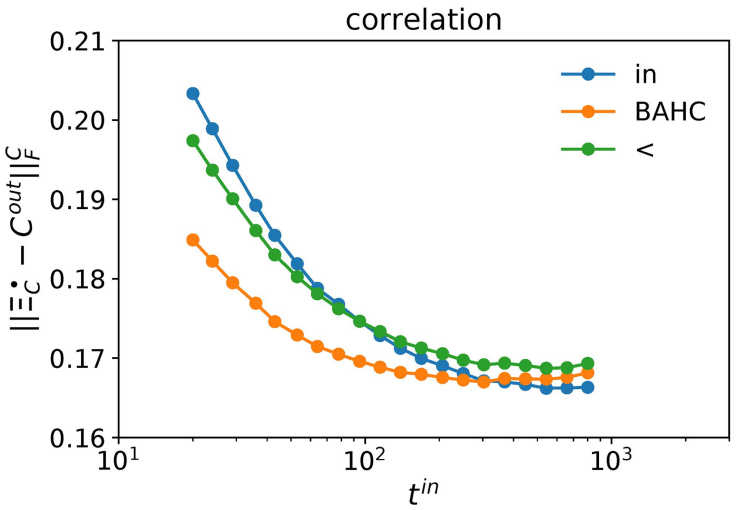
$$C^{\mathrm{BAHC}} = \frac{\sum_{i=1}^{k} C^{(i)<}}{k}$$

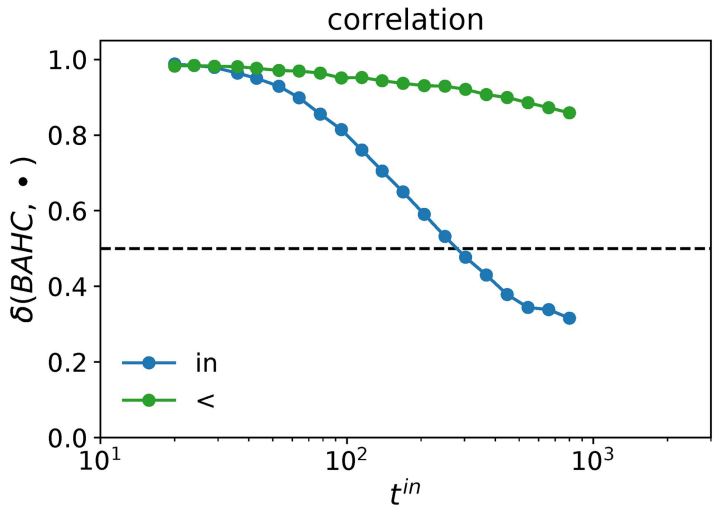
For the covariance:

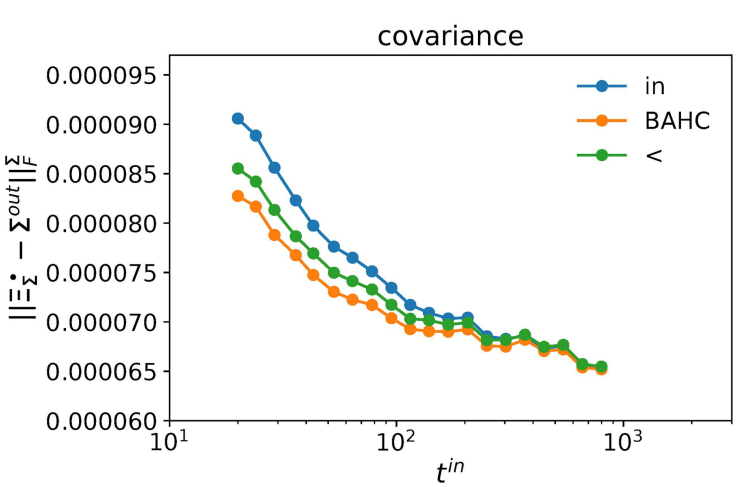
$$\Sigma_{ij}^{\mathrm{BAHC}} = C_{ij}^{\mathrm{BAHC}} \sqrt{\Sigma_{ii}\Sigma_{jj}}$$

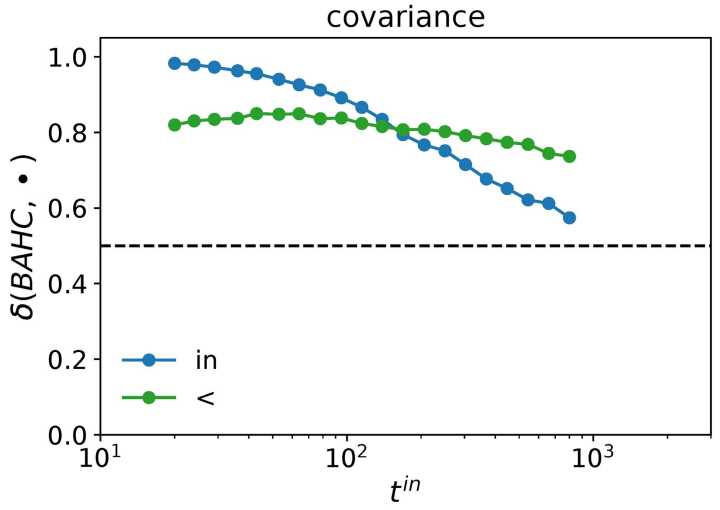
Bongiorno, C., & Challet, D. (2021). Covariance matrix filtering with bootstrapped hierarchies. PloS one, 16(1), e0245092

Eigenvector Overlap







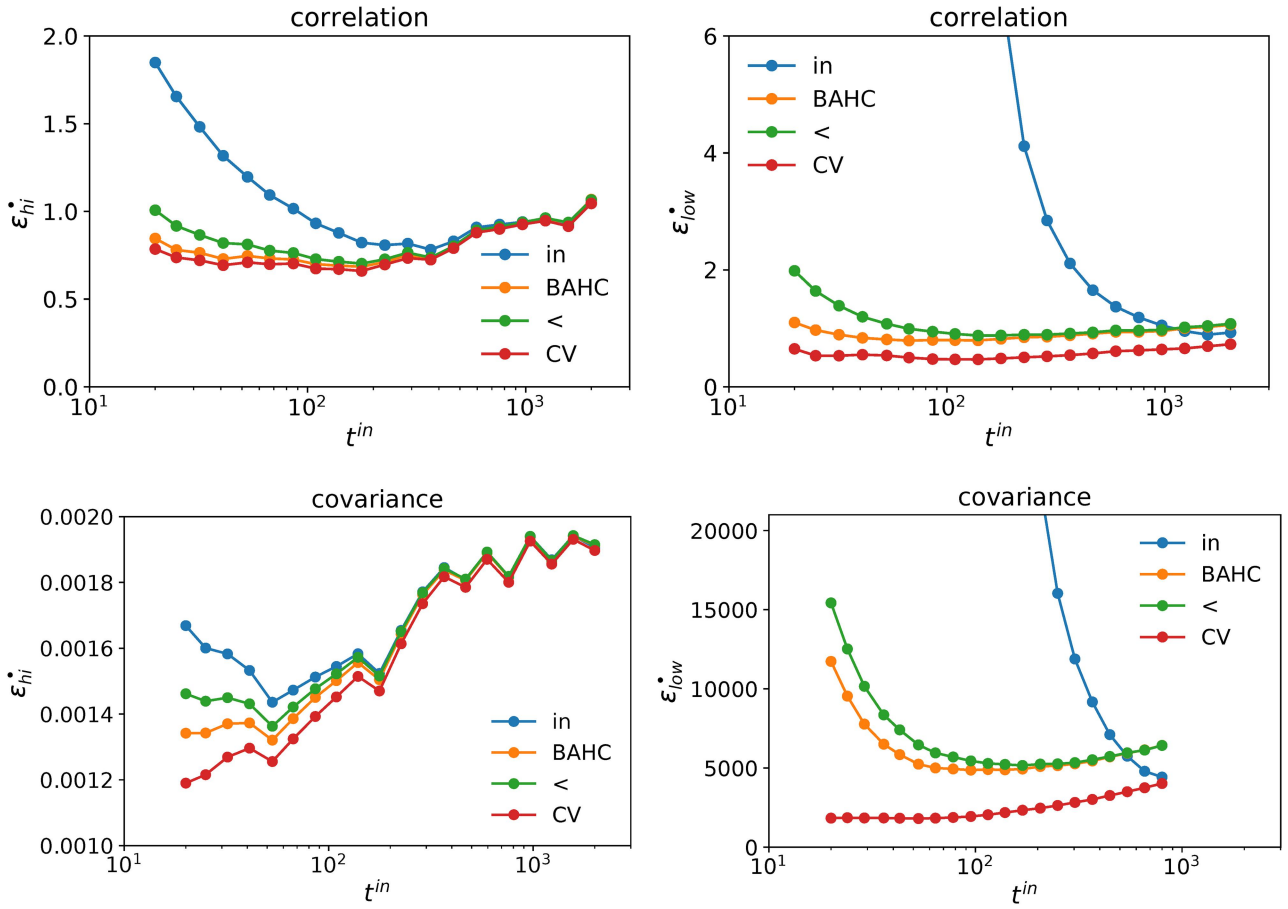


Note that the oracle $\Xi = C^{\text{out}}$ only if $V^{\text{in}} = V^{\text{out}}$

$\delta(\text{BAHC}, \blacksquare)$ is the fraction of times BAHC outperforms the method $\blacksquare$ over independent realizations

10’000 independent realizations

Eigenvalues Overlap



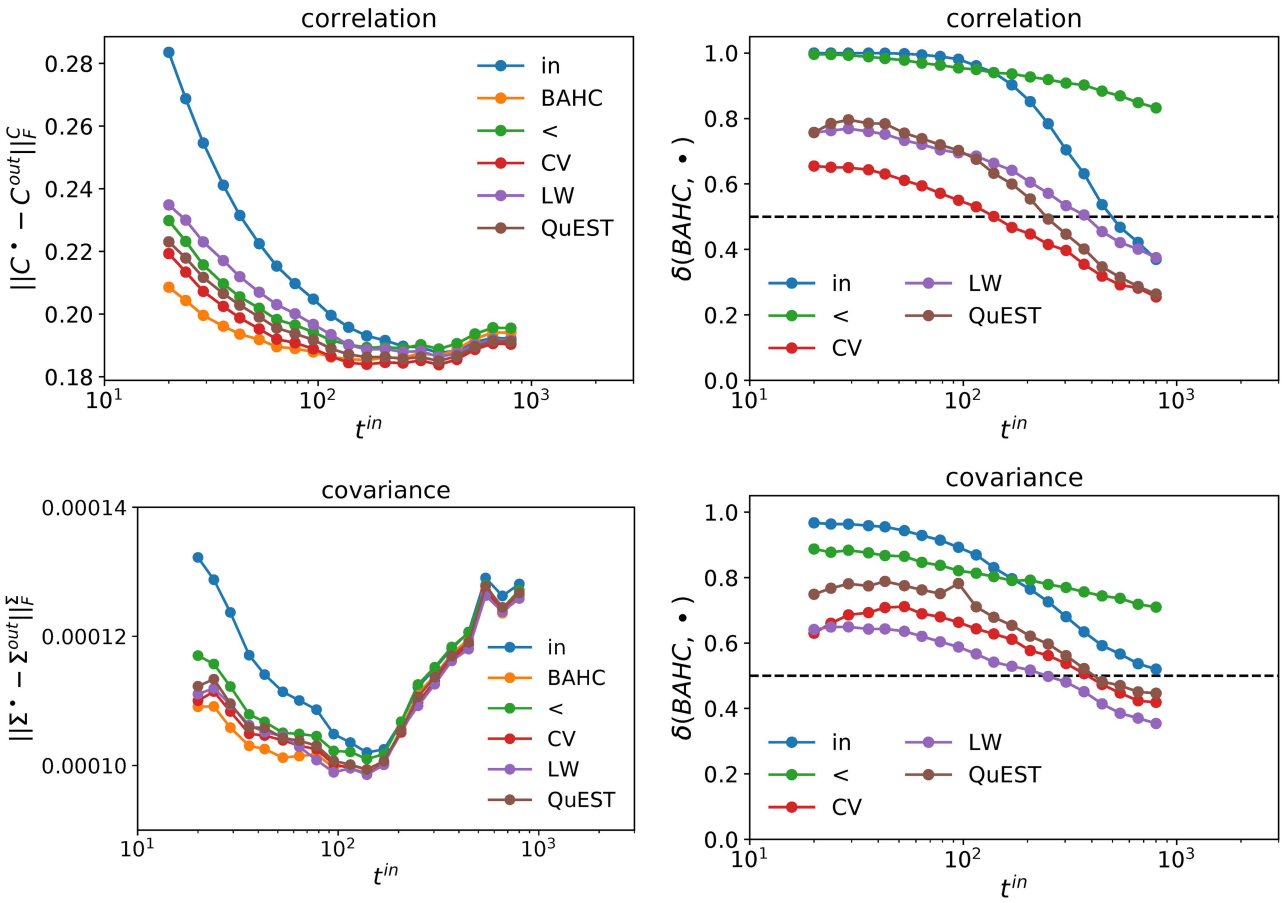
To have an over-all estimator of the eigenvalue overlap we can measure the deviation from the oracle

$$\epsilon_{hi} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - z_i)^2}$$

$$\epsilon_{low} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\lambda_i} - \frac{1}{z_i}\right)^2},$$

Note that this is an over-all estimator not a punctual one, so it does not help you to understand the ptf composition

Discrepancy of the Cor/Cov Estimators

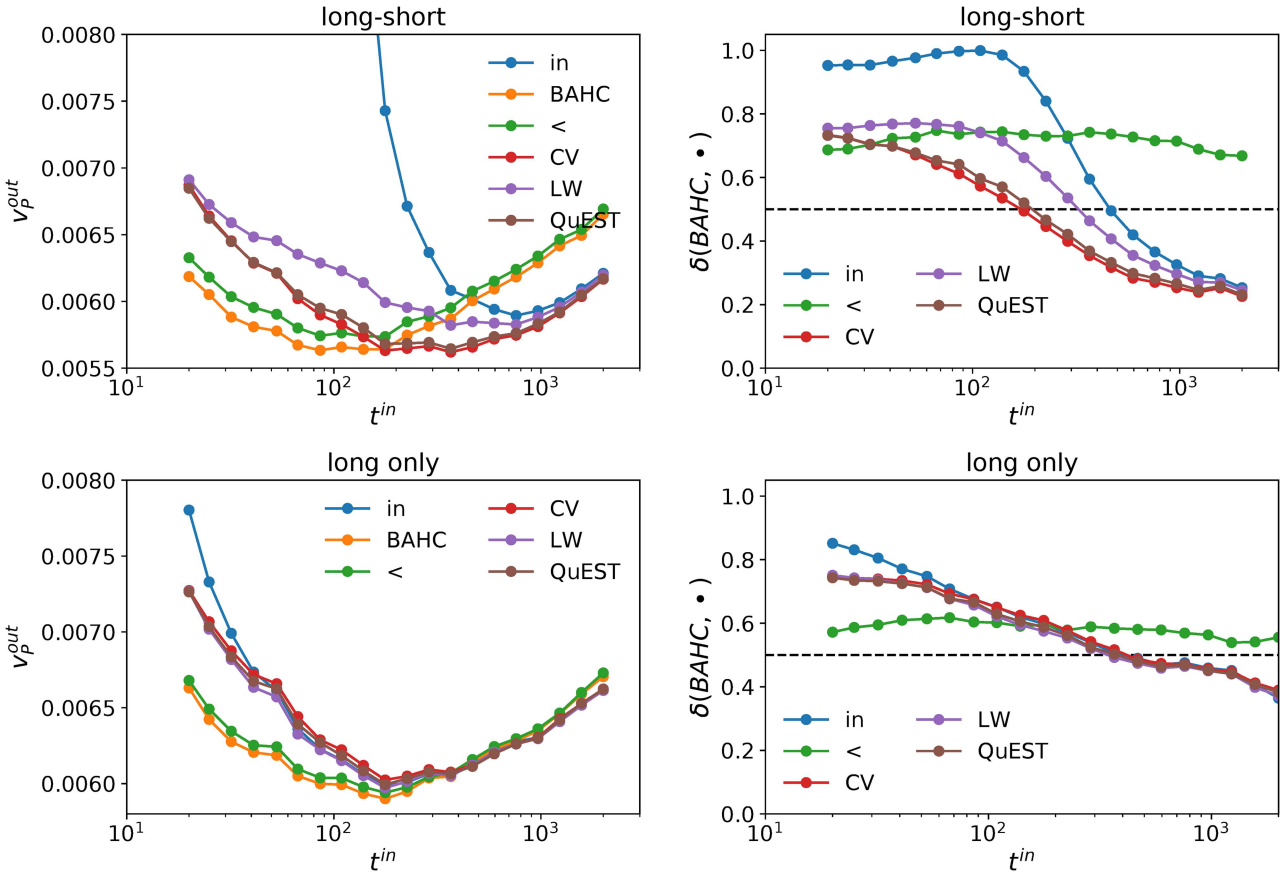


BAHC outperforms in the low rank regime

However, when the calibration window is large is not optimal

BAHC outperform < over the whole range

Risk of the Portfolio



BAHC outperforms all the estimators in the low rank regime.

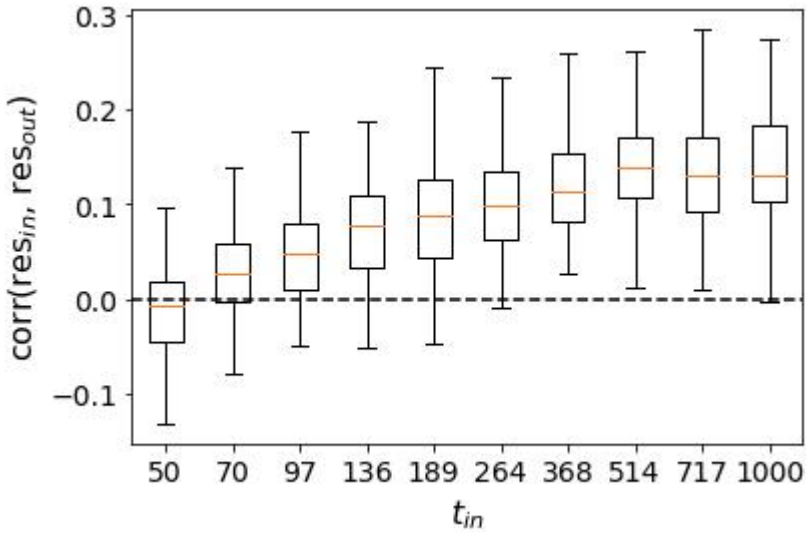
for long-short it obtain its optimal performance equivalent to CV with less data points.

For long-only it reaches the absolute minimum

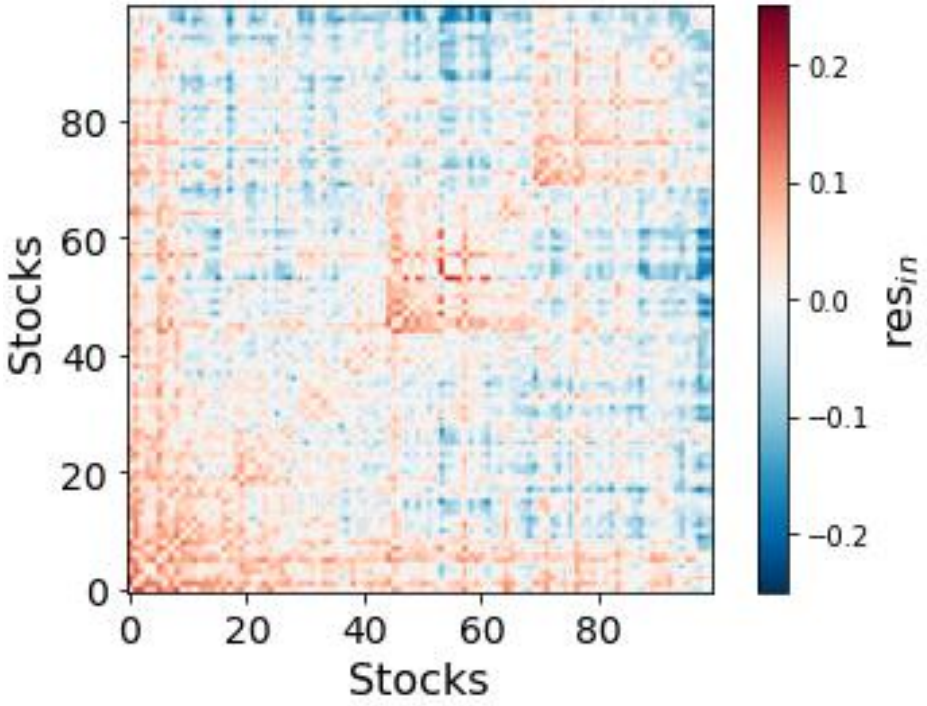
Sorry the risk is not annualized

Deviations from BAHC

The correlation between $C_{ij}^{\mathrm{BAHC}} - C_{ij}^{in}$ *and* $C_{ij}^{\mathrm{BAHC}} - C_{ij}^{\mathrm{out}}$ is significant and increases with the calibration window size



The left out seems to have a pattern



k-fold BAHC

We can cluster and filter the residue recursively with HC filtering.

$$k = 0 \text{ as } \mathbf{C}^{<}_{(0)} = \mathbf{0}.$$

$$\mathbf{E}_{(k)} = \mathbf{C} - \mathbf{C}^{<}_{(k)}.$$

$$\mathbf{C}^{<}_{(k+1)} = \mathbf{C}^{<}_{(k)} + \mathbf{E}^{<}_{(k)}.$$

It is possible to obtain a $C^{<}_{(k)}$ for each bootstrap scenario and the average them [1].

Note that $C^{<}_{(k)}$ for k finite is not non-negative defined, so one should fined the most similar correlation matrix [2].

When k go to infinity

$$\lim_{k\to\infty} \mathbf{C}^{<}_{(k)} = \mathbf{C}$$

for $k \to \infty$ the average is equivalent to the bootstrap average regularization

Note that these are preliminary results and the procedure might change

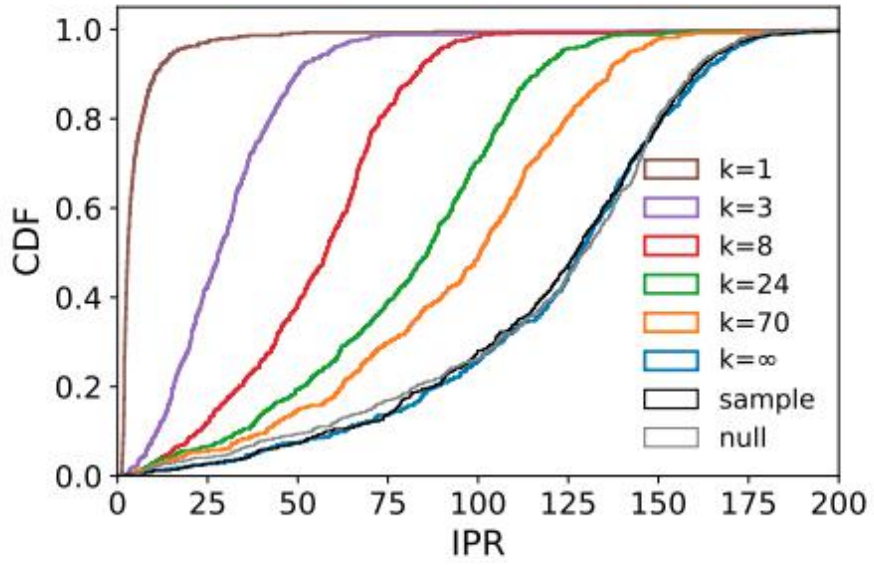
[1] Bongiorno, Christian, and Damien Challet. "Reactive global minimum variance portfolios with k-BAHC covariance cleaning." The European Journal of Finance (2021): 1-17

[2] Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. IMA journal of Numerical Analysis, 22(3), 329-343

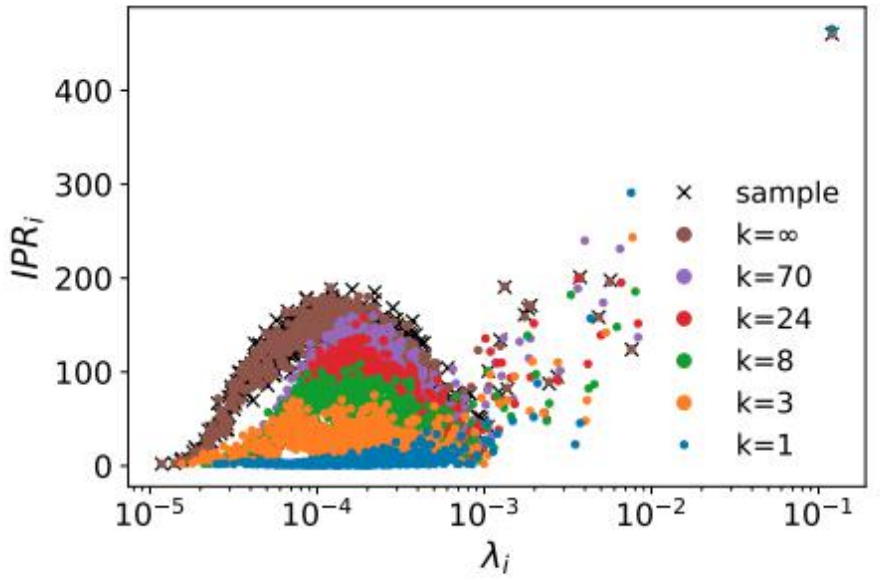
Eigenvalues and Eigenvectors

$$\mathrm{IPR}_j = \frac{1}{\sum_{i=1}^{N} v_{ij}^4}$$

The Inverse Proportion Ratio measures the localization of the eigenvectors, i.e., if the reach high values on a few number of stocks.

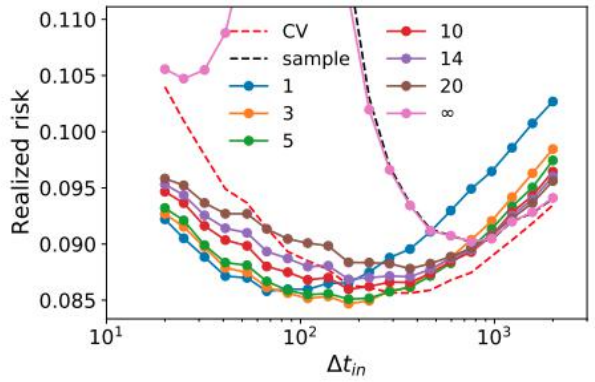


(a) IPR cumulative distribution.

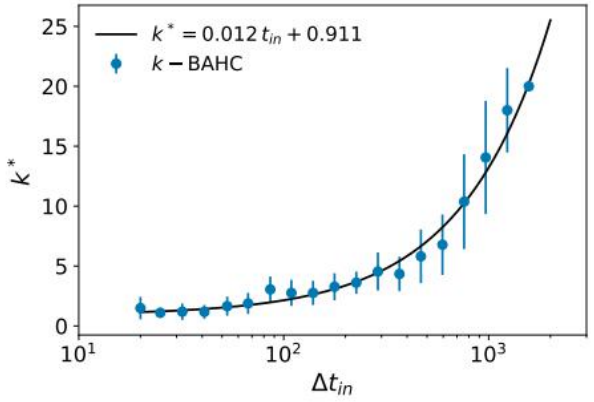


(b) IPRs and eigenvalues scatter plot.

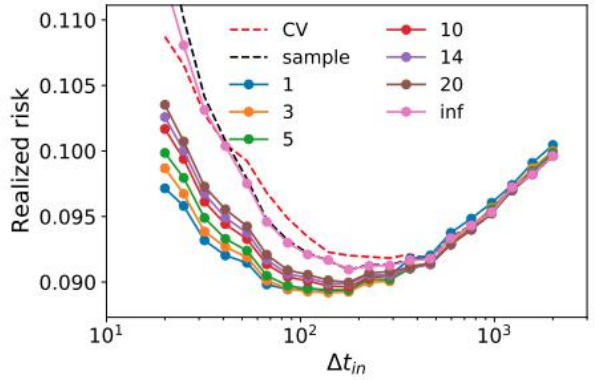
Realized Risk for k-BAHC



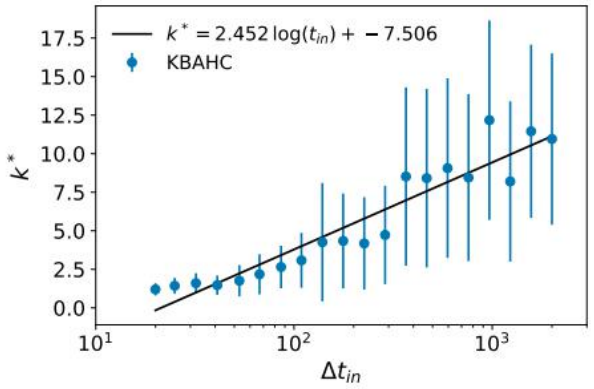
(a) Long-short realized risk.



(b) Long-short optimal k..



(c) Long-only realized risk.



(d) Long-only optimal k..

Finally it outperforms CV

The optimal k increases with the calibration window size.

The longer is the calibration window that more information you can extract from data

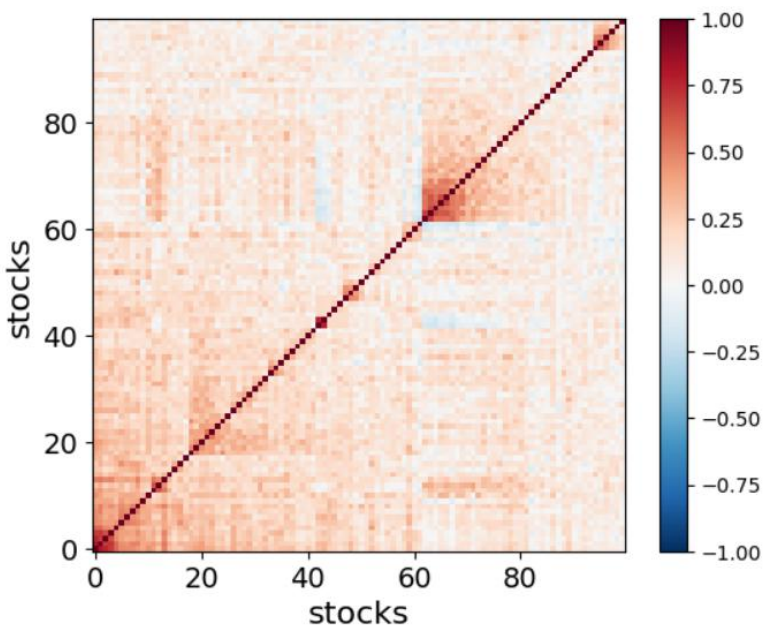
Eigenvalues-Shrinkage+Hierarchical Methods

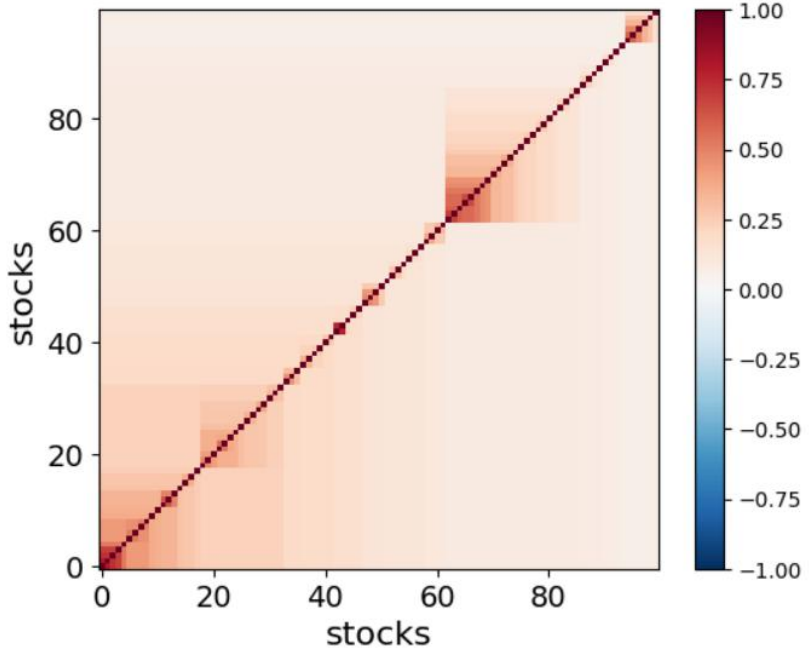
$X_r = \Sigma_<^{-1/2} X$ *Remove the structure*

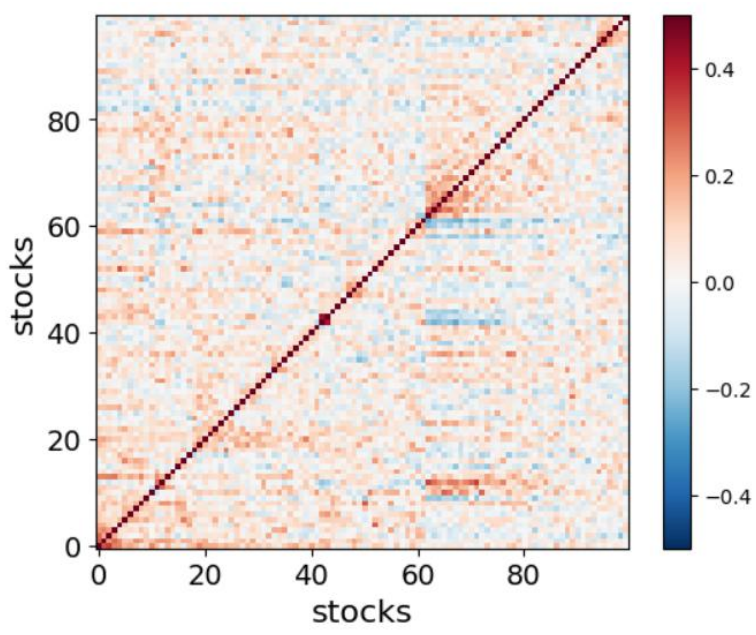
$\Sigma_r \rightarrow f(X_r)$ *Apply a filter to the residuals*

$\Sigma_f = \Sigma_<^{1/2} \Sigma_r \Sigma_<^{1/2}$

Bongiorno, Christian, Vincent Tan, and Stefan Zohren. "Large Covariance Estimation by Bootstrapped Hierarchies and Shrinkage." Available at SSRN (2022).







Questions?

Lab on Hierarchical Filtering

Using the the data you should implement the hierachical clustering filtring.

from scipy.cluster.hierarchy import average, leaves_list

average(dist) is the average-linkage clustering. It requires as input the upper-triangular part of the distance matrix (flatten).

leaves_list(out) takes as input the output of average(dist) and returns the ordered list of elements, which is useful to order the rows and columns of the correlation matrix before the plot

Lab on Hierarchical Filtering

```
>>> Z = average(y)
>>> Z
array([[ 0.        ,  1.        ,  1.        ,  2.        ],
       [ 3.        ,  4.        ,  1.        ,  2.        ],
       [ 6.        ,  7.        ,  1.        ,  2.        ],
       [ 9.        , 10.        ,  1.        ,  2.        ],
       [ 2.        , 12.        ,  1.20710678,  3.        ],
       [ 5.        , 13.        ,  1.20710678,  3.        ],
       [ 8.        , 14.        ,  1.20710678,  3.        ],
       [11.        , 15.        ,  1.20710678,  3.        ],
       [16.        , 17.        ,  3.39675184,  6.        ],
       [18.        , 19.        ,  3.39675184,  6.        ],
       [20.        , 21.        ,  4.09206523, 12.        ]])
```

each row is a new clade

index clade-1 index clade-2 height