Portfolio Allocation and Graph Theory

Aldj Souleïman
Buccafurri Arnaud
Gaillot Solène
M1 Économétrie-Statistiques - Paris 1 Panthéon-Sorbonne

January 2026

Abstract

In this project, we will attempt to build a dynamic portfolio that outperforms the DIJA index. To do this, we will determine which assets in the index are least correlated with the others and invest in those assets. Our portfolio is dynamic because we regularly analyze the dependencies between assets and adjust our portfolio accordingly.

Contents

1 Summary **3**

2 Introduction **4**

3 Description of the database **5**

4 Log returns normalization : GARCH(1,1) **9**

5 Statistical Dependence Between Assets **10**

5.1 Measurement limits of traditional correlation . . . 10

5.1.1 Pearson correlation coefficient . . . 10

5.1.2 Undirected graphs: introduction and notations . . . 10

5.1.3 Limitations of the undirected graph framework . . . 11

5.2 Directed graph and adjacency matrix . . . 11

5.3 Causality . . . 12

5.3.1 VAR(p) . . . 12

5.3.2 The Granger causality test . . . 14

5.4 Cross-correlation . . . 15

5.4.1 Cross-correlation formula . . . 15

5.4.2 The test of no correlation . . . 15

5.4.3 The non parametric approach . . . 15

6 Clustering in Binary Directed Networks **17**

6.1 Graph theory : degrees . . . 17

6.2 General coefficient of clustering . . . 17

6.3 Directed Triangles and Directional Clustering Coefficients . . . 19

7 Portfolio Construction **22**

7.1 Investing when portfolios are riskier than the index . . . 22

7.2 Not investing when portfolios are riskier than the index . . . 23

7.3 Investing in the index when portfolios are riskier than the index . . . 23

8 Results applied to the DIJA index **24**

8.1 Investing when portfolios are riskier than the index . . . 24

8.1.1 Granger-Causality . . . 24

8.1.2 Cross-correlation : Parametric test . . . 24

8.1.3 Cross-correlation : non parametric approach . . . 26

8.2 Not investing when portfolios are riskier than the index . . . 28

8.3 Investing in the index when portfolios are riskier than the index . . . 29

9 Conclusion **30**

10 Annexe **31**

10.1 Bibliography . . . 31

10.2 SAS code . . . 32

10.2.1 Common code for every method . . . 32

1 Summary

Based on a fixed investment universe composed of the Dow Jones Industrial Average (DJIA) actual components, we measure asset dependence using two distinct approaches, the cross-correlation and Granger causality. From these dependence matrices, we compute several clustering coefficients to identify the assets that are the least dependent on the rest of the index. Every two months, we create a new portfolio composed of these weakly dependent assets, optimized to maximize the Sortino ratio while trying to maintain a downside risk that is equal to or lower than that of the DJIA index. For the results, for the cross-correlation approach, the optimal strategy strongly outperforms the index, achieving a cumulative return close to 3000% over a 25-year period. In contrast, portfolios constructed using the Granger causality framework fail to outperform the index.

2 Introduction

Modern portfolio management has historically relied on the principle of diversification. Since Markowitz's foundational work, the central objective has been to maximize the expected return for a given level of risk. However, successive financial crises, the 2000 dot-com bubble, the 2008 subprime crisis, and the COVID-19 crisis, have highlighted the limits of traditional approaches. Indeed, financial markets are not simple collections of isolated assets, they are complex systems in which shocks propagate through networks of interdependencies.

In this project we use some research articles. First, we used the article of Eric Sigward: "Introduction à la théorie des graphes" to introduce graph theory basic notions like oriented graph and adjacency matrix. With this information, we found two methods to compute the dependencies between assets. To do that, we were help by the presentation of different correlation measures by Philippe de Peretti: "Measures of Association". And we then decided to set our choice on the cross-correlation method with the article of El Himdi, Khalid and Roy, Roch : "Tests for Noncorrelation of Two Multivariate ARMA Time Series".

And on the Granger-causality method with the article of Christophe Hurlin : "Un test simple de l'hypothèse de non-causalité dans un modèle de panel hétérogène".

Then, we needed to understand and conceptualize time series, we used the article of Régis Bourbonnais and Virginie Terreza: "Analyse de séries temporelles cours et exercices corrigés - Applications à l'économie et à la gestion". This article gives a panorama on time series (econometrics, ARIMA process, exponential smoothing...).

For the clustering method and to build our portfolio, we were inspired by Giorgio Fagiolo: "Clustering in Complex Directed Networks". We learnt how to do clustering coefficient and directed triangles.

Finnaly, we also wanted to normalize our error, we base ourselves on Jean-Michel ZAKOIAN: "Modèles GARCH et à volatilité stochastique".

Thus, we complete each article with one another and we apply them into a financial and dynamic context.

The most common dependence measure, Pearson's correlation coefficient, has major structural limitations, it is both static and symmetric, and it only captures correlation without providing any directional information. Yet, financial markets are empirically characterized by complex and asymmetric dynamics. In this context, we use directed graph theory, by modelling the market as a directed network. Then, we exploit this information for asset allocation and for the construction of a portfolio that is rebalanced every two months.

This leads us to the following question: "How can a dynamic allocation strategy based on directional clustering coefficients outperform the benchmark index while exhibiting a similar or lower level of risk?"

To answer this question, we will first show why traditional measures present limitations in the dynamic context of finance. We will then move to cross-correlation, with its parametric and non-parametric approaches. We will then discuss another way to construct the adjacency matrix using the Granger causality test. Following by the introduction of the essential elements of Fagiolo's clustering technique by focusing on the notions that are relevant for our thesis. Finally, we will explain concretely how we build our portfolio and conclude by commenting the obtained results, their limits, as well as possible improvement paths.

3 Description of the database

Throughout this paper, we apply the theoretical methods on financial assets from the DIJA (Dow Jones Industrial Average) index. This index consists of 30 American stocks equally weighted, considered as representative of the US economy.

However, the composition of the DIJA changes over time, which makes it difficult to compare results across different periods. To simplify our analysis, we created a fixed version of the index. First, we selected the stocks currently included in the DIJA and then, we retrieved their financial data for the period wanted, regardless of whether the stocks were always part of the index.

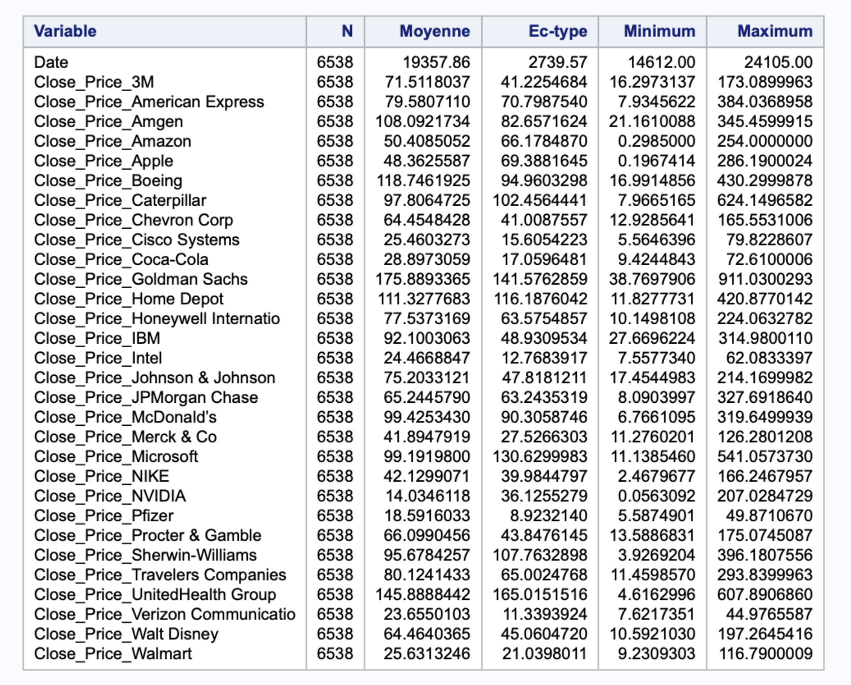
We extracted this data using the Python library "yfinance", which provides daily financial information from Yahoo Finance. The availability of historical data on the website influenced the period we could analyze. As a result, our dataset covers January 3, 2000, to December 31, 2025.

Two companies lacked complete data on Yahoo Finance and were therefore replaced by companies that are no longer part of the index. As a result, the final set of assets includes the following:

	Companies
1	3M
2	American Express
3	Amgen
4	Amazon
5	Apple
6	Boeing
7	Caterpillar
8	Chevron
9	Cisco Systems
10	Coca-Cola
11	Goldman Sachs
12	Home Depot
13	Honeywell International
14	IBM
15	Intel

	Companies
16	Johnson & Johnson
17	JPMorgan Chase
18	McDonald's
19	Merck & Co
20	Microsoft
21	NIKE
22	NVIDIA
23	Pfizer
24	Procter & Gamble
25	Sherwin-Williams
26	Travelers Companies
27	UnitedHealth Group
28	Verizon Communications
29	Walt Disney
30	Walmart

For each of these companies, we use the closing market price for every business day and obtain the following information:



Variable	N	Moyenne	Ec-type	Minimum	Maximum
Date	6538	19357.86	2739.57	14612.00	24105.00
Close_Price_3M	6538	71.5118037	41.2254684	16.2973137	173.0899963
Close_Price_American Express	6538	79.5807110	70.7987540	7.9345622	384.0368958
Close_Price_Amgen	6538	108.0921734	82.6571624	21.1610088	345.4599915
Close_Price_Amazon	6538	50.4085052	66.1784870	0.2985000	254.0000000
Close_Price_Apple	6538	48.3625587	69.3881645	0.1967414	286.1900024
Close_Price_Boeing	6538	118.7461925	94.9603298	16.9914856	430.2999878
Close_Price_Caterpillar	6538	97.8064725	102.4564441	7.9665165	624.1496582
Close_Price_Chevron Corp	6538	64.4548428	41.0087557	12.9285641	165.5531006
Close_Price_Cisco Systems	6538	25.4603273	15.6054223	5.5646396	79.8228607
Close_Price_Coca-Cola	6538	28.8973059	17.0596481	9.4244843	72.6100006
Close_Price_Goldman Sachs	6538	175.8893365	141.5762859	38.7697906	911.0300293
Close_Price_Home Depot	6538	111.3277683	116.1876042	11.8277731	420.8770142
Close_Price_Honeywell Internatio	6538	77.5373169	63.5754857	10.1498108	224.0632782
Close_Price_IBM	6538	92.1003063	48.9309534	27.6696224	314.9800110
Close_Price_Intel	6538	24.4668847	12.7683917	7.5577340	62.0833397
Close_Price_Johnson & Johnson	6538	75.2033121	47.8181211	17.4544983	214.1699982
Close_Price_JPMorgan Chase	6538	65.2445790	63.2435319	8.0903997	327.6918640
Close_Price_McDonald's	6538	99.4253430	90.3058746	6.7661095	319.6499939
Close_Price_Merck & Co	6538	41.8947919	27.5266303	11.2760201	126.2801208
Close_Price_Microsoft	6538	99.1919800	130.6299983	11.1385460	541.0573730
Close_Price_NIKE	6538	42.1299071	39.9844797	2.4679677	166.2467957
Close_Price_NVIDIA	6538	14.0346118	36.1255279	0.0563092	207.0284729
Close_Price_Pfizer	6538	18.5916033	8.9232140	5.5874901	49.8710670
Close_Price_Procter & Gamble	6538	66.0990456	43.8476145	13.5886831	175.0745087
Close_Price_Sherwin-Williams	6538	95.6784257	107.7632898	3.9269204	396.1807556
Close_Price_Travelers Companies	6538	80.1241433	65.0024768	11.4598570	293.8399963
Close_Price_UnitedHealth Group	6538	145.8888442	165.0151516	4.6162996	607.8906860
Close_Price_Verizon Communicatio	6538	23.6550103	11.3393924	7.6217351	44.9765587
Close_Price_Walt Disney	6538	64.4640365	45.0604720	10.5921030	197.2645416
Close_Price_Walmart	6538	25.6313246	21.0398011	9.2309303	116.7900009

For each asset, we calculate its daily log return. It measures how much the value of an asset has changed proportionally. The log return formula is written as follows:

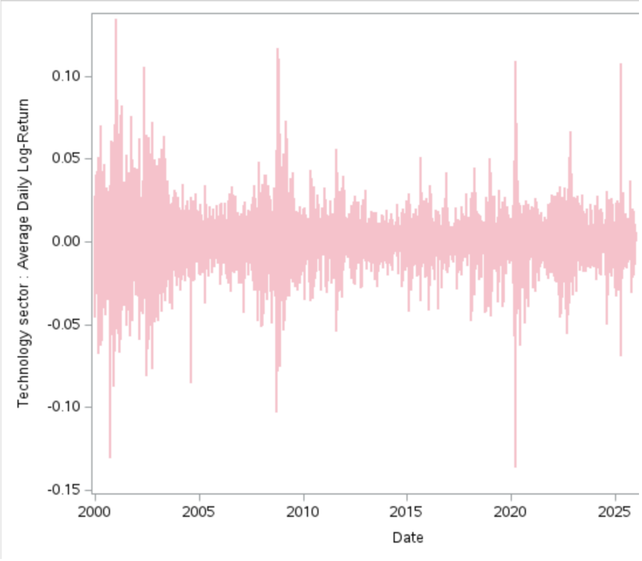
$$z_t = log(P_t) - log(P_{t-1})$$

Where P_t is the price of an asset at market close on day t.

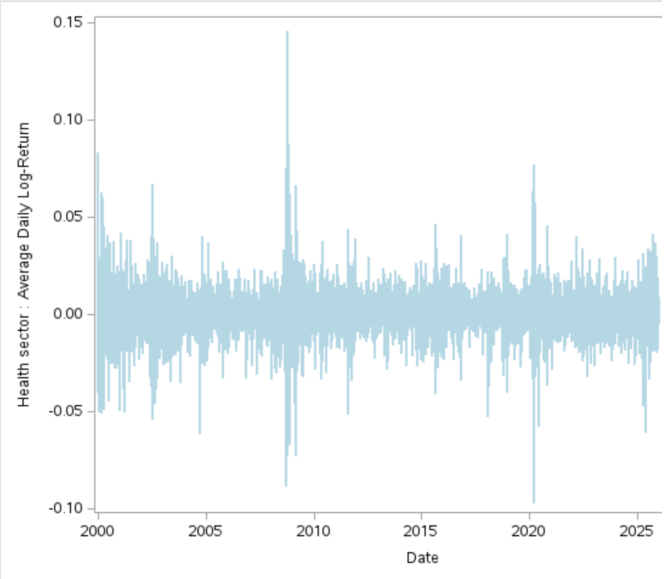
By taking the logarithmic returns, we obtain series that are approximately stationary, which justifies their use in correlation and clustering analysis.

We analyze the average log returns of companies belonging to the same industry:

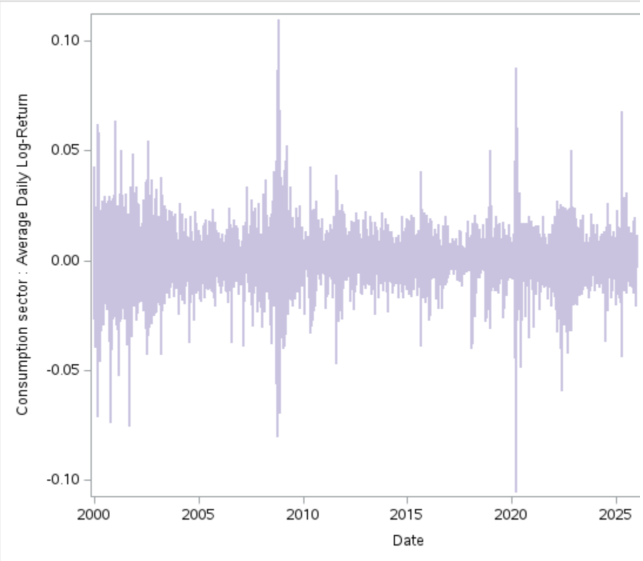
1. Technology (Apple, Microsoft, Intel, IBM, Cisco Systems, Verizon Communications, NVIDIA).



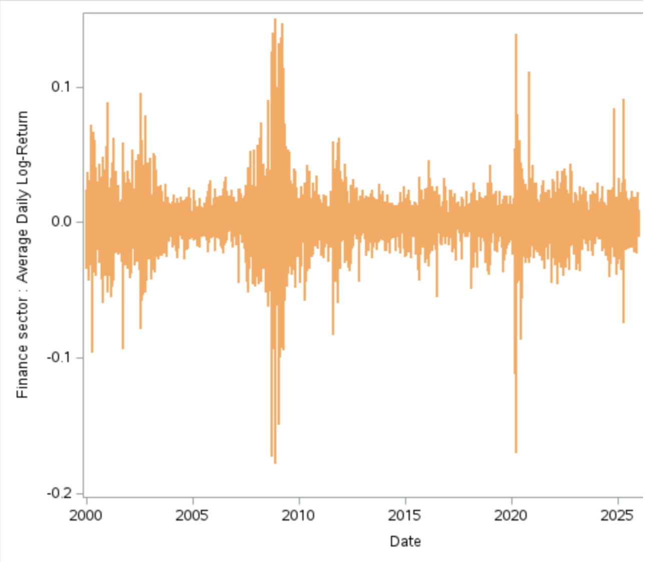
2. Health and Pharmaceuticals (Johnson & Johnson, Pfizer, Merck & Co, Amgen, UnitedHealth Group)



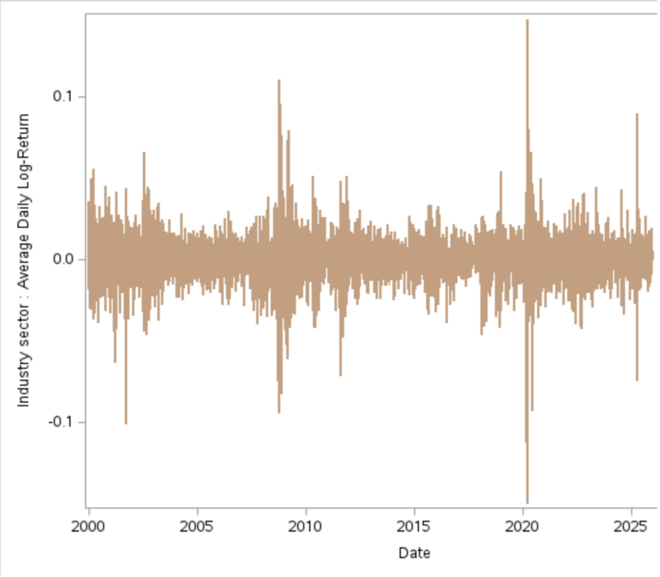
3. Consumption (Nike, McDonald's, Home Depot, Walt Disney, Coca-Cola, Procter & Gamble, Walmart, Amazon)



4.Finance (JPMorgan Chase, American Express, Goldman Sachs, Travelers Companies)



5.Industry (Boeing, Caterpillar, 3M, Honeywell International, Chevron Corp, Sherwin-Williams)



There are clear differences in behavior between sectors, especially in times of crisis such as the dot-com bubble in 2000, the subprime mortgage crisis in 2008, and the COVID crisis in 2020.

4 Log returns normalization : GARCH(1,1)

In the context of financial time series, the error variance is generally not homoscedastic. Generalized Autoregressive Conditional Heteroskedasticity (GARCH(K, q)) models account for time-varying conditional variance in the residuals. With a GARCH(1,1) model, we can normalize the time series using its conditional standard deviation.

Definition :(X_t) is a GARCH(K,q) processus if :

- $\epsilon_t = \sigma_t z_t$
- $\exists w, (\alpha_i)_{1 \leq i \leq q}, (\gamma_i)_{1 \leq i \leq K},\ \ \mathbb{V}[X_t | X_{t-1}] = \sigma_t^2 = w + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^2 + \sum_{i=1}^{K} \gamma_i \epsilon_{t-i}^2$

(z_t) is i.i.d with moment $\mathbb{E}[z_t] = 0$ and $\mathbb{V}[z_t] = 1$. We don't know the distribution of (z_t).
We restrein at the case where $q = K = 1$. (ϵ_t) who follow GARCH(1,1) has for variance :

$$\sigma_t^2 = \mathbb{V}[\epsilon_t | \epsilon_{t-1}] = w + \alpha\sigma_{t-1}^2 + \gamma\epsilon_{t-1}^2 \iff \sigma_t^2 = w + \sigma_{t-1}^2(\alpha z_{t-1} + \gamma)$$

The GARCH(1,1) model admit only one solution if :

$$\beta = \mathbb{E}[log(\alpha z_t^2 + \gamma)] < 0$$

In the context of heteroscedasticity, the OLS estimator are not BLUE, we can fix the problem with σ_t, we have $\epsilon_t = \sqrt{\sigma_t^2 z_t}$. We can divide the temporal serie (x_t) by its standard derivation. We fix $h(t) = \sqrt{\sigma_t^2}$

$$\tilde{x_t} = \frac{x_t}{h(t)}$$

If we suppose a (M) model for the serie $(x_t)_{1 \leq t \leq T}$, we have

$$(M) : x = Z\beta + \epsilon$$

Where $\mathbb{V}[\epsilon] = \Omega \neq \mathbb{I}_T$, $\Omega = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_T^2 \end{pmatrix}$ Now, we have (M') an equivalent model to (M)

$$(M') : \Omega^{-1/2}x = \Omega^{-1/2}Z\beta + \Omega^{-1/2}\epsilon$$

$$\iff (M') : \tilde{x} = \tilde{Z}\beta + \tilde{\epsilon}$$

$\forall t \in \{1, ..., T\}$, $\mathbb{V}[\tilde{\epsilon_t}] = \mathbb{V}[\frac{\epsilon_t}{\sqrt{\sigma_t^2}}] = \frac{1}{\sigma_t^2}\mathbb{V}[\epsilon_t] = 1$

5 Statistical Dependence Between Assets

5.1 Measurement limits of traditional correlation

5.1.1 Pearson correlation coefficient

Let $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ be two variables representing the returns of two financial assets. T denotes the number of temporal observations.

The covariance between X and Y is defined by

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Empirically the covariance is estimated by

$$\widehat{\operatorname{cov}}(X,Y) = \frac{1}{T-1}\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y}),$$

where $\bar{x}$ and $\bar{y}$ denote the empirical means.

The Pearson correlation coefficient is then defined as

$$\rho_{XY} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

In the empirical case we obtain

$$\hat{\rho}_{XY} = \frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2}\sqrt{\sum_{t=1}^{T}(y_t-\bar{y})^2}}.$$

By the Cauchy–Schwarz inequality we have $-1 \leq \hat{\rho}_{XY} \leq 1$.

In the financial context the Pearson correlation coefficient corresponds to a bilateral and symmetric relationship between two assets. It refers to the notion of co-movement in the sense that it measures how two assets tend to evolve jointly over a given period. In other words it does not provide information about a possible direction or a transmission mechanism. This point constitutes as we will see later an important limitation in the context of our thesis.

5.1.2 Undirected graphs: introduction and notations

In order to extend the analysis of dependence relations to a set of financial assets it is necessary to introduce some notions from graph theory.

An undirected graph G is defined by a pair $G = (V,E)$ where $V = \{x_1,\ldots,x_n\}$ denotes the set of vertices of the graph and $E \subset \mathcal{P}_2(V)$ the set of edges. Each vertex represents a financial asset while an edge corresponds to a dependence relationship between two assets. When an edge $e = \{x_i, x_j\} \in E$ connects two vertices x_i and x_j these vertices are said to be adjacent. The absence of orientation implies that the relationship between x_i and x_j does not have a privileged direction.

In our context an edge exists between two vertices when the returns of the corresponding assets exhibit a dependence measured by the Pearson correlation coefficient.

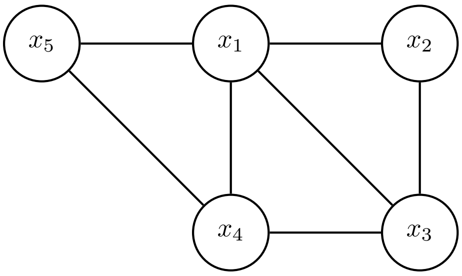


Figure 1: An undirected graph G with five vertices

5.1.3 Limitations of the undirected graph framework

First, this framework imposes a symmetric relationship between assets. However financial markets are characterized by asymmetric interactions. For instance a shock affecting Chevron Corp may impact Boeing whereas a shock affecting Chevron Corp does not necessarily generate an immediate or comparable effect on Boeing. Such asymmetric relationships cannot be captured by a symmetric dependence measure since $\rho_{ij} = \rho_{ji}$ by construction.

Second, the Pearson correlation coefficient is static in nature. It measures a correlation over a fixed period and does not take into account temporal delays in the transmission of shocks. In practice a shock affecting Chevron Corp may impact Boeing over several days rather than instantaneously.

Third, during calm periods correlations between assets tend to be moderate and sector-specific whereas during crisis periods correlations often increase sharply across the entire market. The Pearson correlation coefficient averages these different periods and does not distinguish between normal and stress periods.

Finally, the undirected graph framework describes pairwise co-movements between assets and it does not identify transmission channels central assets or contagion paths. The absence of directionality prevents the analysis of how shocks originate propagate and amplify across the financial system.

Many other dependence measures have similar limitations. For instance, measures such as Spearman's correlation or distance correlation capture alternative forms of dependence between two variables. However despite measuring non-linear associations these measures remain symmetric and are computed over a fixed period of time. As a result they fail to capture asymmetric interactions temporal delays or shock propagation mechanisms that characterize financial markets.

Consequently the introduction of directed graphs provides a more suitable theoretical framework for the objectives of this thesis.

5.2 Directed graph and adjacency matrix

To understand causality and cross-correlation, we need to introduce a directed graph G, which is described by two sets: $V = (x_1, ..., x_n)$, representing the vertices, and $E \subset V \times V$ representing the arcs. We denote the graph by $G = (V, E)$.

In this project, the vertices represent the assets, and the arcs represent the statistical dependence between

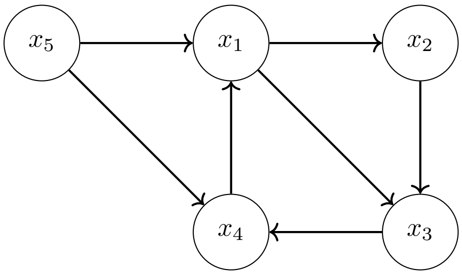


Figure 2: An oriented graph G with 5 verteces

two assets. If asset x_1 causes asset x_2, we define the relation $x_1 \to x_2$. For example, the asset Chevron Corp, a company in the energy and oil sector, may influence the asset Boeing, a firm in the transportation industry. Graph theory allows us to represent all the causal links between DIJA assets.

One way to model all these asset relationships is by using an adjacency matrix. The coefficient $(a_{i,j})$, indicates whether asset x_i causes asset x_j. In graph theory, the adjacency matrix $A = (a_{i,j})_{1 \leq i,j \leq n}$ contains only 0s and 1s. 1 if $x_i \to x_j$ and 0 otherwise.

$$A = (a_{i,j})_{1 \leq i,j \leq n} = \begin{cases} 1 & \text{if} \quad x_i \to x_j \\ 0 & \quad \text{else} \end{cases}$$

We can construct the adjacency matrix A of the graph G shown in Figure 2:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

5.3 Causality

5.3.1 VAR(p)

A VAR(p) process (Vector Autoregressive) models multiple variables. The VAR(p) model is an extension of the AR(p) model. The standard form of a VAR(p) model is:

$$y_t = c + X_1 y_{t-1} + X_2 y_{t-2} + ... + X_p y_{t-p} + \epsilon_t$$

$$\forall t \in \mathbb{R}, y_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{K,t} \end{pmatrix}, (A_i)_{1 \leq i \leq p} \text{ are coefficient matrices, and } \epsilon_t = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{K,t} \end{pmatrix} \sim \begin{pmatrix} BB(0, \sigma_1^2) \\ BB(0, \sigma_2^2) \\ \vdots \\ BB(0, \sigma_K^2) \end{pmatrix}$$

We want to determine whether one asset causes another, so we set $K = 2$. The model takes the following form:

$$\begin{pmatrix} x_{i,t} \\ x_{j,t} \end{pmatrix} = \begin{pmatrix} c_{i,t} \\ c_{j,t} \end{pmatrix} + \begin{pmatrix} a_{i,1}^{(1)} & a_{j,1}^{(1)} \\ a_{i,2}^{(1)} & a_{j,2}^{(1)} \end{pmatrix} \begin{pmatrix} x_{i,t-1} \\ x_{j,t-1} \end{pmatrix} + ... + \begin{pmatrix} a_{i,1}^{(p)} & a_{j,1}^{(p)} \\ a_{i,2}^{(p)} & a_{j,2}^{(p)} \end{pmatrix} \begin{pmatrix} x_{i,t-p} \\ x_{j,t-p} \end{pmatrix} + \begin{pmatrix} \epsilon_{i,t} \\ \epsilon_{i,t} \end{pmatrix}$$

In this equation, the two series $(x_{i,t})_{1 \leq t \leq T}$ and $(x_{j,t})_{1 \leq t \leq T}$ correspond to two DIJA assets. Using the VAR(p) model, we can examine whether the past values of one asset influence another.
We have $\mathbb{E}[y_t] = \mu = (I_K - A_1 - ... - A_p)^{-1} c$, $\forall t \in \mathbb{N}$ and we denote $\mathbb{E}[(y_t - \mu)(y_t - \mu)^t] = \Gamma_y(h)$ as the

autocovariance function.

$$\Gamma_y(h) = A_1\Gamma_y(h-1) + ... + A_p\Gamma_y(h-p)$$

The autocorrelation function can be calculated as follows:

$$R_y(h) = D^{-1/2}\Gamma_y(h)D^{-1/2}$$

Here, D is a diagonal matrix containing the variances of the two variables x_i and x_j on the diagonal, with 0 elsewhere.

There are several ways to estimate $\hat{\beta}$, particularly using maximum likelihood estimation, but here we only present the GLS estimator. Here $K = 2$, we can write:

$$\underbrace{\begin{pmatrix} | \\ x_1 \\ | \\ | \\ x_2 \\ | \end{pmatrix}}_{X\in\mathbb{R}^{2n}} = \underbrace{\begin{pmatrix} \overbrace{\begin{pmatrix} | & | & | & | & | & | & | \\ 1 & x_{1,t-1} & \cdots & x_{1,t-p} & x_{2,t-1} & \cdots & x_{2,t-p} \\ | & | & | & | & | & | & | \end{pmatrix}}^{Z_0} & 0_{\mathcal{M}_{n,2p+1}(\mathbb{R})} \\ 0_{\mathcal{M}_{n,2p+1}(\mathbb{R})} & \begin{pmatrix} | & | & | & | & | & | & | \\ 1 & x_{1,t-1} & \cdots & x_{1,t-p} & x_{2,t-1} & \cdots & x_{2,t-p} \\ | & | & | & | & | & | & | \end{pmatrix} \end{pmatrix}}_{Z\in\mathcal{M}_{2n,4p+2}(\mathbb{R})} \underbrace{\begin{pmatrix} \left.\begin{matrix} \beta_0 \\ \beta_{1,1} \\ \vdots \\ \beta_{2,p} \end{matrix}\right\} B_1 \\ \left.\begin{matrix} \beta'_0 \\ \beta'_{1,1} \\ \vdots \\ \beta'_{2,p} \end{matrix}\right\} B_2 \end{pmatrix}}_{B\in\mathbb{R}^{2p+2}} + \underbrace{\begin{pmatrix} | \\ u_1 \\ | \\ | \\ u_{2,t} \\ | \end{pmatrix}}_{U\in\mathbb{R}^{2n}} \tag{1}$$

We aim to estimate the VAR(p) model $X = ZB + U$ using GLS. The equation (1) is equivalent to

$$X = (I_2 \otimes Z_0)B + U,\ I_2 \otimes Z_0 = \begin{pmatrix} Z_0 & 0 \\ 0 & Z_0 \end{pmatrix}$$

Where $\otimes$ is the kroenecker product so $I_2 \otimes Z_0 = \begin{pmatrix} Z_0 & 0 \\ 0 & Z_0 \end{pmatrix}$

According to Zellner theoreme, in this case whe have $\hat{B}_{OLS} = \hat{B}_{GLS} = \hat{B} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta_{2,p}} \\ \hat{\beta}'_0 \\ \vdots \\ \hat{\beta'_{2,p}} \end{pmatrix}$

And by the OLS we have

$$\hat{B} = (Z'Z)^{-1}Z'X,\quad \hat{B} \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} B$$

We can calculate separetely B_1 and B_2

$$\begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = \left(\begin{pmatrix} Z_0 & 0 \\ 0 & Z_0 \end{pmatrix}' \begin{pmatrix} Z_0 & 0 \\ 0 & Z_0 \end{pmatrix}\right)^{-1} \begin{pmatrix} Z_0 & 0 \\ 0 & Z_0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \iff \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = \begin{pmatrix} (Z_0'Z_0)^{-1}Z_0X_1 \\ (Z_0'Z_0)^{-1}Z_0X_2 \end{pmatrix}$$

5.3.2 The Granger causality test

In our case, we estimate a VAR(1) model, as we want to know the dependence of one asset on another over a one-day period. So we can write the equation as follows:

$$\underbrace{\begin{pmatrix} x_{i,t} \\ x_{j,t} \end{pmatrix}}_{X} = \underbrace{\begin{pmatrix} 1 & x_{i,t-1} & x_{j,t-1} & 0 & 0 \\ 0 & 0 & 1 & x_{i,t-1} & x_{j,t-1} \end{pmatrix}}_{Z} \underbrace{\begin{pmatrix} c_i \\ \beta_{ii} \\ \beta_{ij} \\ c_j \\ \beta_{ji} \\ \beta_{jj} \end{pmatrix}}_{B} + \underbrace{\begin{pmatrix} u_{i,t} \\ u_{j,t} \end{pmatrix}}_{U}$$

In order to test whether the values of x_j at $t-1$ provide information for predicting x_i at t beyond x_i itself, we perform Granger's causality test.

1. The restricted model:

$$x_{i,t} = c_i + \beta_{ii} x_{i,t-1} + u^c_{i,t}$$

2. The unrestricted model:

$$x_{i,t} = c_i + \beta_{ii} x_{i,t-1} + \beta_{ij} x_{j,t-1} + u^{nc}_{i,t}$$

Then, we compare the variances of the residuals of the two equations, performing a Wald test at the 5% significance level, such that the null hypothesis is written as $H0$: $\beta_{ij} = 0 \Longleftrightarrow x_j$ does not Granger-cause x_i.

And we set the following test statistic:

$$1F = \frac{RSS^c - RSS^{nc}}{\frac{RSSnc}{T-3}} \xrightarrow{\mathcal{L}} \chi^2_{(1)}$$

With $RSS^c = \sum_{t=1}^{T} (u^c_{i,t})^2$ and $RSS^{nc} = \sum_{t=1}^{T} (u^{nc}_{i,t})^2$

Finally, we determine $P(1F > 2.71) = 0.10$, the rejection region for $H0$ for this test.

However, we want to perform these tests on a dynamic network, so we use the rolling window method to do so.

For each pair of assets (i, j), and for each date t, we perform tests on a sample of size $[t-w, t]$. Where w is the window size, which represents the number of past days taken into account to estimate the relationships between assets. We can vary w in order to compare the results obtained.

Next, still using the same window, we construct the adjacency matrix $A(t)$, which takes:

$$A(t) = (a_{i,j})_{1 \le i,j \le n} = \begin{cases} 0 & \text{if} \quad r_{i,j}(k) = 0 \\ 1 & \text{else} \end{cases}$$

Then, using this matrix, it is possible to plot the associated directed graph.

We then repeat this procedure in a new window that we shift by one day, i.e., $[t+1-w, t+1]$.

5.4 Cross-correlation

5.4.1 Cross-correlation formula

We need a cross-correlation measure that is consistent with a directed graph. Therefore, the cross-correlation must be oriented and not symmetric.

In financial applications, daily log-returns typically exhibit a negligible conditional mean, $\mathbb{E}[r_t] \approx 0$, and very weak linear autocorrelation in the conditional mean, while displaying strong conditional heteroskedasticity. As a consequence, modeling the conditional mean dynamics using ARMA-type residuals is not necessary at the daily frequency.

For this reason, we do not rely on ARMA residuals in our model. To introduce directionality, we compute an oriented cross-correlation at lag $k = 1$. For two assets i and j, we measure the dependence between $x_{i,t}$ and $x_{j,t-1}$.

On a window of size T, the directed cross-correlation used in our code is given by:

$$\hat{r}_{i,j}(1) = \frac{\sum_{t=2}^{T} z_{i,t}\, z_{j,t-1}}{\sqrt{\sum_{t=2}^{T} z_{i,t}^2}\,\sqrt{\sum_{t=1}^{T-1} z_{j,t}^2}}.$$

This measure is not symmetric in general since $\hat{r}_{i,j}(1) \neq \hat{r}_{j,i}(1)$. Thus, when we compute $\hat{r}_{i,j}(1)$, we are measuring the influence of asset j at time $t-1$ on asset i at time t.

$$\mathbf{R}(1) = \begin{pmatrix} \hat{r}_{1,1}(1) & \hat{r}_{1,2}(1) & \cdots & \hat{r}_{1,n}(1) \\ \hat{r}_{2,1}(1) & \hat{r}_{2,2}(1) & \cdots & \hat{r}_{2,n}(1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{n,1}(1) & \hat{r}_{n,2}(1) & \cdots & \hat{r}_{n,n}(1) \end{pmatrix}.$$

5.4.2 The test of no correlation

After calculating the cross-correlation at lag k = 1, we perform a statistical test to verify whether the correlation is significant.

We define the test statistic for a finite sample $QH(k)$, which we set in our case as:

$$QH(1) = \tfrac{n^2}{n-1} * \hat{r}_{i,j}(k)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$$

With n the sample size.

We set up the test such that the null hypothesis ($H0$) is the hypothesis of no correlation ($r_{i,j}(k) = 0$), which we perform at a 5% significance level.

For this significance level and degree of freedom, we are in the rejection region for $H0$ when $QH(1) > 3.84$.

We therefore perform several tests, still within the rolling window framework presented above, and construct the corresponding adjacency matrices.

5.4.3 The non parametric approach

To construct the adjacency matrices $A(t)$, we can also manually determine a threshold s, such that:

$$A(t) = (a_{i,j})_{1 \leq i,j \leq n} = \begin{cases} 0 & \text{if } \hat{r}_{i,j}(k) < s \\ 1 & \text{if } \hat{r}_{i,j}(k) > s \end{cases}$$

In this case, we no longer use the non-correlation test seen previously, but remain within the rolling window approach.

6 Clustering in Binary Directed Networks

6.1 Graph theory : degrees

We consider a binary oriented network $G = (V, E)$. $V = \{x_1, ..., x_{30}\}$ are the 30 Dow Jones assets and $E \subset V \times V$ are the causality representation. This network is represented by the adjacent matrix $A(t) = (a_{i,j})_{1 \leq i,j \leq 30}$ with $a_{i,j} = 1$ if the cross-correlation coeficient $r_{i,j}(k) \neq 0$ and $a_{i,j} = 0$ else.

In graph theori, the degree of an edge is the number of vertex that connect this edge with another one. Using the adjacent matrix for clustering is essential. d_i^{in} is the in-degre of note i is the number of edge pointing to i and d_i^{out} is the out-degre of note i is the number of edge originating to i. We can calculate this two values with the adjacent matrix.

$$d_i^{in} = \sum_{j \neq i} a_{j,i} = (A^T)_i \mathbf{1}_{30}$$

$$d_i^{out} = \sum_{j \neq i} a_{i,j} = (A)_i \mathbf{1}_{30}$$

In these equations, A is the adjacent matrix, A^T it transpose and $(A)_i$ the i-line of the A matrix. We also have $\mathbf{1}_{30}$ who is the vector of 1 in $\mathbb{R}^{30}$ For understand the signfication of degree in a oriented network, if we have $d_i^{out} = k$, that means the i asset cause k others assets. Conversely, if we have $d_i^{in} = k$, that means k assets cause i asset.

We can also introduce $d_i^{\leftrightarrow}$ the bilateral vertex. If a i asset cause the j asset and j cause i then, the vertex toward i and j is bilateral.

$$d_i^{\leftrightarrow} = (A^2)_{i,i}$$

where $(A^2)_{i,i}$ is the (i,i) coeficient of A^2 matrix.

An another degree measure is important in our case, d_i^{tot} is the sum of in-degree and out-degree. This measure is useful because it take care of how many asset cause itself and how many asset it cause.

$$d_i^{tot} = d_i^{in} + d_i^{out} = (A^T + A)_i \mathbf{1}_{30}$$

Now, we have four different degree measure, $d_i^{in}, d_i^{out}, d_i^{tot}$ and $d_i^{\leftrightarrow}$. These degree measure will be used in the next part for calculate clustering coefficient with directed triangles.

6.2 General coefficient of clustering

We must introduce the most general form of the clustering coefficient applied to binary directed networks (BDNs).

We begin by writing the expression of the coefficient, then we describe each term to fully understand where the formula comes from and what it measures:

$$C_i^D(A) = \frac{t_i^D}{T_i^D} = \frac{\frac{1}{2} \sum_{j \neq i} \sum_{h \neq (i,j)} (a_{ij} + a_{ji})(a_{ih} + a_{hi})(a_{jh} + a_{hj})}{d_i^{\text{tot}}\big(d_i^{\text{tot}} - 1\big) - 2d_i^{\leftrightarrow}}.$$

t_i^D denotes the number of directed triangles effectively formed around node i, while T_i^D represents the total number of directed triangles potentially formable by node i.

By construction, the clustering coefficient satisfies

$$C_i^D \in [0,1],$$

since the numerator is always less than or equal to the denominator. It can therefore be interpreted as follows: $C_i^D = 0$ means that no triangle is formed around asset i; $C_i^D = 1$ corresponds to a perfectly closed neighborhood, in which all possible triangle configurations are realized.

This expression can also be written in matrix form:

$$C_i^D(A) = \frac{(A + A^T)^3_{ii}}{2\big[d_i^{\text{tot}}\big(d_i^{\text{tot}} - 1\big) - 2d_i^{\leftrightarrow}\big]}.$$

This basic form takes into account all triangles that can be formed around node i, regardless of edge direction. Node i constitutes the central point of the analysis. We then consider all its possible neighbors, but only in pairs of two neighbors (j, h).

The numerator relies on products of the form

$$(a_{ij} + a_{ji})(a_{ih} + a_{hi})(a_{jh} + a_{hj}),$$

which play the role of triangle indicators.

For example, for any pair of nodes involving i and j, we have:

$$a_{ij} + a_{ji} = \begin{cases} 0, & \text{if there is no edge between } i \text{ and } j, \\ 1, & \text{if there is an edge in only one direction,} \\ 2, & \text{if there are two edges, hence a bilateral relationship.} \end{cases}$$

The same reasoning applies to pairs (i, h) and (j, h).

The sums

$$\sum_{j \neq i} \sum_{h \neq (i,j)}$$

allow us to traverse the set of distinct neighbor pairs of node i. We thus consider all possible configurations involving the triplet (i, j, h), with $j \neq i$ and $h \neq (i, j)$. This allows us to account for the set of directed triangles involving node i.

The factor $\frac{1}{2}$ corrects the double counting related to index order, since the same pair of neighbors (j, h) and (h, j) would otherwise lead to counting the same triangle twice.

Consequently, the product

$$(a_{ij} + a_{ji})(a_{ih} + a_{hi})(a_{jh} + a_{hj}),$$

and thus the numerator

$$\frac{1}{2} \sum_{j \neq i} \sum_{h \neq (i,j)} (a_{ij} + a_{ji})(a_{ih} + a_{hi})(a_{jh} + a_{hj}),$$

are non-zero if, and only if, the three connections exist. This indicates the presence of a triangle connecting i, j, and h, regardless of the orientation of the edges.

The denominator

$$d_i^{\text{tot}}\left(d_i^{\text{tot}}-1\right)$$

corresponds to the maximum number of triangles that node i could form.

If node i has d_i^{tot} incident edges, the number of ways to choose two distinct neighbors without taking order into account is given by

$$\binom{d_i^{\text{tot}}}{2}=\frac{d_i^{\text{tot}}!}{2!\,(d_i^{\text{tot}}-2)!}=\frac{d_i^{\text{tot}}(d_i^{\text{tot}}-1)(d_i^{\text{tot}}-2)!}{2\,(d_i^{\text{tot}}-2)!}=\frac{d_i^{\text{tot}}(d_i^{\text{tot}}-1)}{2}.$$

However, in a directed network, the order of neighbors is important, because a pair (j,h) and the pair (h,j) can lead to two different directed triangle configurations. We therefore reason in terms of arrangements of two neighbors among d_i^{tot}, which gives

$$A^2_{d_i^{\text{tot}}}=2\times\binom{d_i^{\text{tot}}}{2}=d_i^{\text{tot}}\left(d_i^{\text{tot}}-1\right).$$

The corrective term $-2d_i^{\leftrightarrow}$ that we subtract is justified by the existence of configurations that do not correspond to true triangles. This occurs when node i is connected to the same neighbor j by two opposite edges $(i\to j)$ and $(j\to i)$. These two edges can be selected as a pair, whereas they involve only one neighbor and therefore cannot form a triangle.

There are $d_i^{\leftrightarrow}$ situations of this type for node i, and each generates two degenerate configurations. This is why we subtract the corrective term $2d_i^{\leftrightarrow}$, in order to exclude these "false triangles" and obtain the effective number of real potential directed triangles.

Finally, the term

$$(A+A^T)^3_{ii}$$

corresponds to the general approach adopted for the study of triangles around asset i. At this stage, the causal direction of edges, and thus the precise form of directed triangles, does not yet specifically interest us.

This transformation does not mean that directional information is ignored or lost, but that it is voluntarily set aside in a first step. The objective is simply to get a clear idea of the number of neighbors of asset i that are connected to each other, regardless of link orientation.

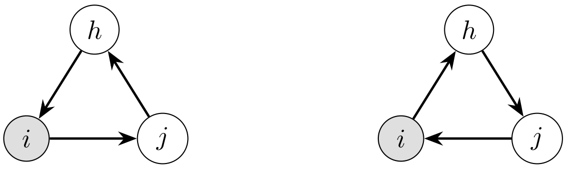
6.3 Directed Triangles and Directional Clustering Coefficients

We are now specifically interested in directed triangle configurations, which constitute the central element of the clustering measure in directed networks. In a directed network, and from the point of view of a given node i, there are in total eight possible configurations of directed triangles in which i appears as one of the vertices.

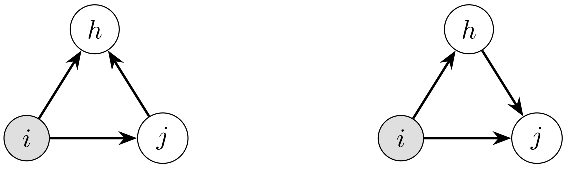
However, these eight configurations are not all distinct from a structural point of view. Indeed, they can be grouped into four fundamental motifs, each motif corresponding to two symmetric edge orientations. These motifs translate different flow patterns around node i and are the origin of the definition of the directional clustering coefficient.

More precisely, we distinguish four fundamental motifs of directed triangles from the point of view of node i. Each of these motifs corresponds to two symmetric orientations, which leads to a total of eight possible configurations.

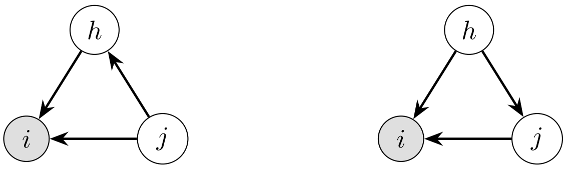
Cycle: Relationships of type $i \to j \to h \to i$ (or reverse orientation). This motif reflects a cyclic circulation of information between node i and two of its neighbors.



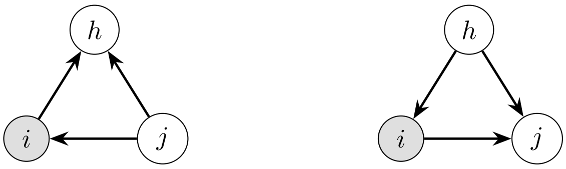
Out: Node i emits two outgoing edges towards its neighbors, typically $i \to j$ and $i \to h$. It then acts as a transmitter in the triangle structure.



In: Node i receives two incoming edges from its neighbors, for example $j \to i$ and $h \to i$. It then plays a receiver role in the triangle structure.



Middleman: Node i plays an intermediation role between two other nodes, according to a structure of type $j \to i \to h$ (or symmetrically $h \to i \to j$).



The eight possible directed triangles involving node i, organized by structural motif. Each motif is represented by two symmetric orientations.

The total number of directed triangles t_i^D decomposes into four distinct terms derived from the matrix expansion. We identify each of these terms with a precise structural motif:

$$t_i^D = (A + A^T)^3_{ii} = \underbrace{(A^3)_{ii}}_{t_i^{\text{cyc}}} + \underbrace{(AA^TA)_{ii}}_{t_i^{\text{mid}}} + \underbrace{(A^TA^2)_{ii}}_{t_i^{\text{in}}} + \underbrace{(A^2A^T)_{ii}}_{t_i^{\text{out}}}.$$

Thus, we have the following correspondences for the numerator of each coefficient:

$$t_i^{\text{cyc}} = t_{i1}^D, \quad t_i^{\text{mid}} = t_{i2}^D, \quad t_i^{\text{in}} = t_{i3}^D, \quad t_i^{\text{out}} = t_{i4}^D.$$

The maximum number of triangles that is theoretically possible to form, T_i^D, decomposes according to the in-degree, out-degree, and reciprocal degree of node i. We associate each motif with its theoretical potential:

$$T_i^D = \underbrace{\left[d_i^{\text{in}} d_i^{\text{out}} - d_i^{\leftrightarrow}\right]}_{T_i^{\text{cyc}}} + \underbrace{\left[d_i^{\text{in}} d_i^{\text{out}} - d_i^{\leftrightarrow}\right]}_{T_i^{\text{mid}}} + \underbrace{\left[d_i^{\text{in}}(d_i^{\text{in}} - 1)\right]}_{T_i^{\text{in}}} + \underbrace{\left[d_i^{\text{out}}(d_i^{\text{out}} - 1)\right]}_{T_i^{\text{out}}}.$$

Finally, for each motif, the clustering coefficient is defined as the ratio between the number of observed triangles and the maximum number of possible triangles for this specific motif.

We thus obtain the following four formulas:

$$CC_{\text{cyc}} = \frac{t_i^{\text{cyc}}}{T_i^{\text{cyc}}} = \frac{(A^3)_{ii}}{d_i^{\text{in}} \cdot d_i^{\text{out}} - d_i^{\leftrightarrow}}$$

$$CC_{\text{mid}} = \frac{t_i^{\text{mid}}}{T_i^{\text{mid}}} = \frac{(AA^TA)_{ii}}{d_i^{\text{in}} \cdot d_i^{\text{out}} - d_i^{\leftrightarrow}}$$

$$CC_{\text{in}} = \frac{t_i^{\text{in}}}{T_i^{\text{in}}} = \frac{(A^TA^2)_{ii}}{d_i^{\text{in}}(d_i^{\text{in}} - 1)}$$

$$CC_{\text{out}} = \frac{t_i^{\text{out}}}{T_i^{\text{out}}} = \frac{(A^2A^T)_{ii}}{d_i^{\text{out}}(d_i^{\text{out}} - 1)}$$

7 Portfolio Construction

7.1 Investing when portfolios are riskier than the index

For each of the four types of clustering (In, Out, Cyc, Mid), we test a grid of 100 different thresholds (ranging from 0.01 to 1). For each threshold, it selects assets whose clustering value is below the threshold (thus creating a "test" portfolio for each threshold and each CC). A precaution is built in to ensure that each "test" portfolio contains at least three assets in order to guarantee minimum diversification.

For each "clustering/threshold" combination, we then calculate its associated Sortino Ratio. The Sortino ratio measures the performance of a portfolio relative to a risk-free asset (here $R_f = 0$), while penalizing only downside risk.

The Sortino ratio is defined here as:

$$S_o = \frac{\mathbb{E}[R_p]}{\sigma_p^-}$$

where $\mathbb{E}[R_p]$ denotes the expected portfolio return and σ_p^- is the downside deviation, computed using only negative portfolio returns:

$$\sigma_p^- = \sqrt{\mathbb{E}\left[\min(R_p, 0)^2\right]}$$

Unlike the Sharpe ratio, the Sortino ratio only penalizes negative volatility (downside risk), which is relevant when trying to minimize losses.

Next, we look for the threshold that maximizes the Sortino ratio while requiring that the portfolio's risk of loss be less than or equal to that of the DIJA index. If no threshold allows us to create a portfolio with a downside risk inferior or equal to the index, we select the combination that offered the best absolute Sortino ratio (riskier than the index).

Algorithm 1 Portfolio Construction via Clustering Threshold Selection

```
Inputs:
  Clustering coefficients $\{CC^{IN}, CC^{OUT}, CC^{MID}, CC^{CYC}\}$
  Threshold grid $\Theta = \{0.01, 0.02, \dots, 1.00\}$
for each rebalancing date do
    for each clustering motif $r \in \{IN, OUT, MID, CYC\}$ do
        for each threshold $\theta \in \Theta$ do
            Select assets such that $CC_r < \theta$
            if number of selected assets < 3 then
                Continue
            end if
            Construct an equally weighted portfolio
            Compute portfolio returns over the holding period
            Compute portfolio Sortino ratio
            if portfolio downside risk exceeds benchmark downside risk then
                Discard portfolio
            end if
        end for
        Retain the threshold $\theta_r^*$ that maximizes the Sortino ratio for motif $r$
    end for
    Select the motif $r^*$ associated with the highest Sortino ratio
    Construct the final equally weighted portfolio using $(r^*, \theta_{r^*}^*)$
    Store portfolio returns
end for
```

7.2 Not investing when portfolios are riskier than the index

Same method as explained above, except that if the risk of portfolio loss is not less than or equal to that of the DIJA index, then we do not invest in any assets during the period concerned.

7.3 Investing in the index when portfolios are riskier than the index

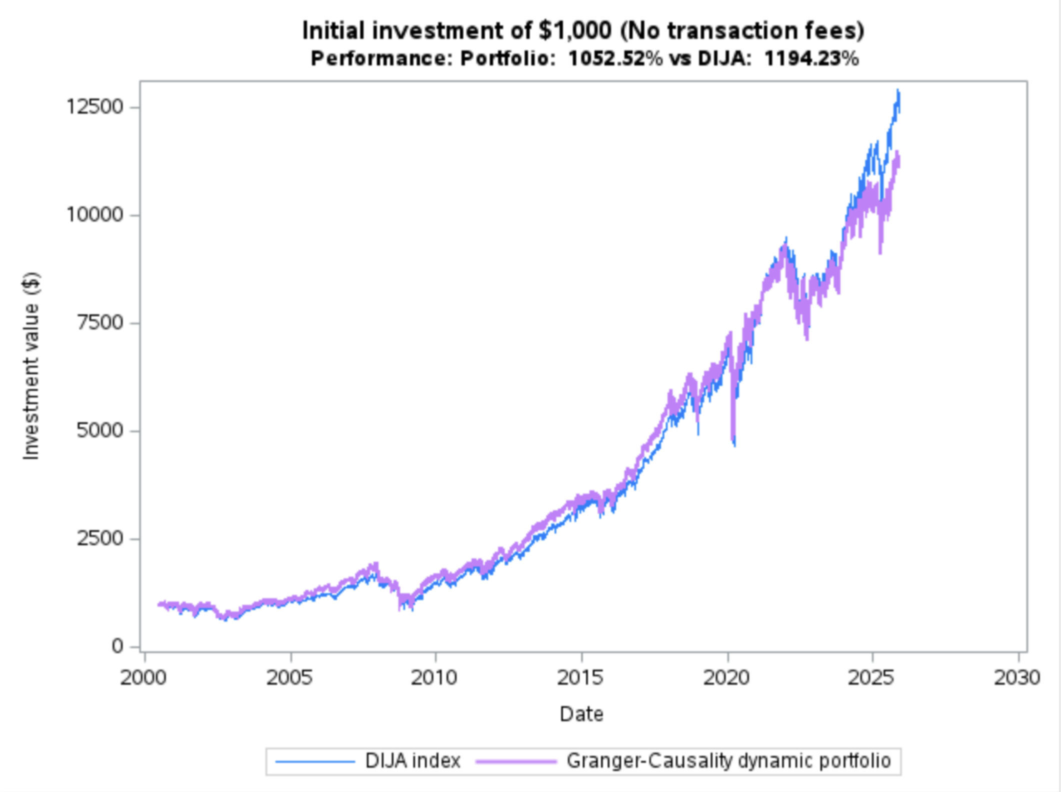
On the contrary, if the risk of loss of the portfolio is not less than or equal to that of the DIJA index, we then invest in all the assets of the index during the period concerned.

8 Results applied to the DIJA index

8.1 Investing when portfolios are riskier than the index

8.1.1 Granger-Causality

In a first approach, we use Granger causality to construct our portfolio. We set the significance level of the Granger causality test at $\alpha = 0.10$ due to the limited number of observations per period. The network window is $T = 132$, and one observation is lost because of the VAR(1) process. The data are normalized using a GARCH(1,1) model, and the results of the Granger causality test are as follows.

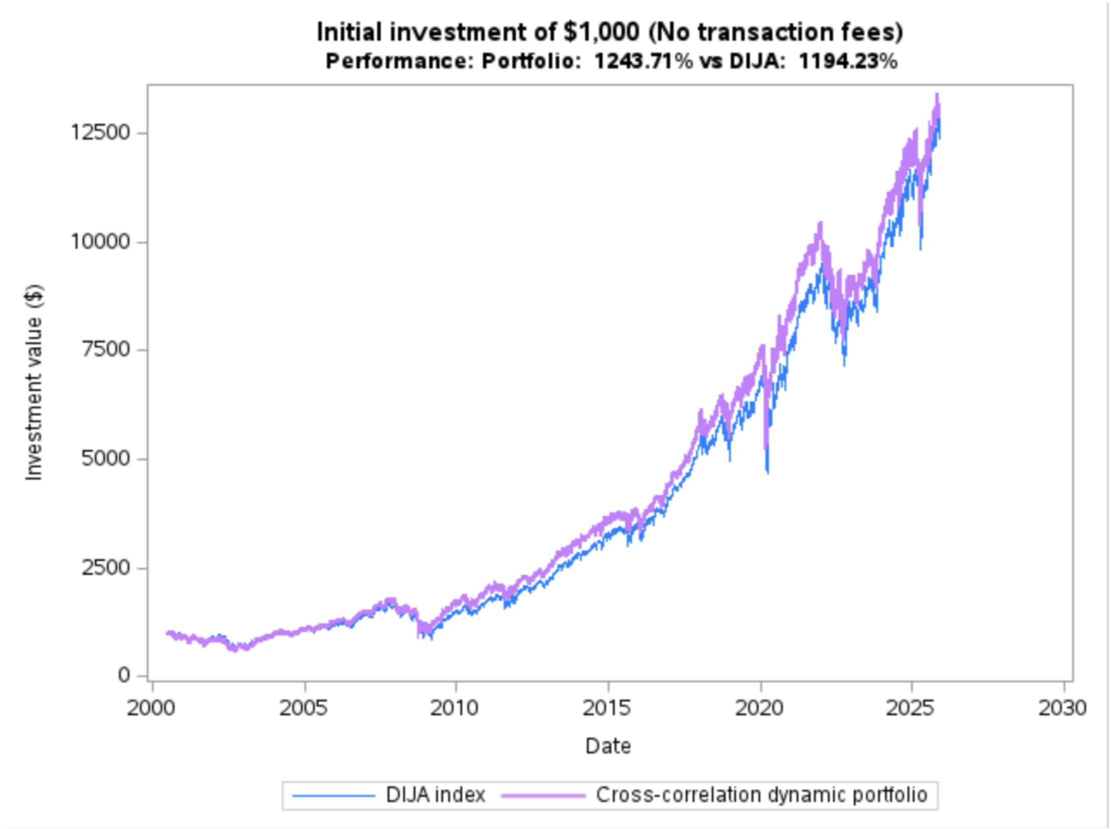


Our strategy does not outperform the DJIA index. We attribute this result to the limited number of observations. In each period, we perform $N(N-1) = 870$ tests. With a significance level of $\alpha = 0.10$, this implies approximately 87 false positives. The large number of tests therefore leads to multiple false positives, which in turn generate biased results. Using a more stringent significance level, such as $\alpha = 5\%$ or $\alpha = 2.5\%$, does not improve the results. Indeed, we are not in an asymptotic setting, and the errors are not Gaussian. With only 132 observations, the test lacks sufficient power and is therefore inefficient. Another reason is the structure of the test itself. The Granger causality test only assesses whether the past values of one asset help predict another. It captures temporal causality rather than structural causality and does not account for underlying microeconomic mechanisms.

8.1.2 Cross-correlation : Parametric test

We then try the cross-correlation method followed by the parametric test at a 5% threshold. To do this, we still normalize our data with GARCH(1,1), and we keep the same time horizons, (6 months of analysis and 2 months of portfolio management).

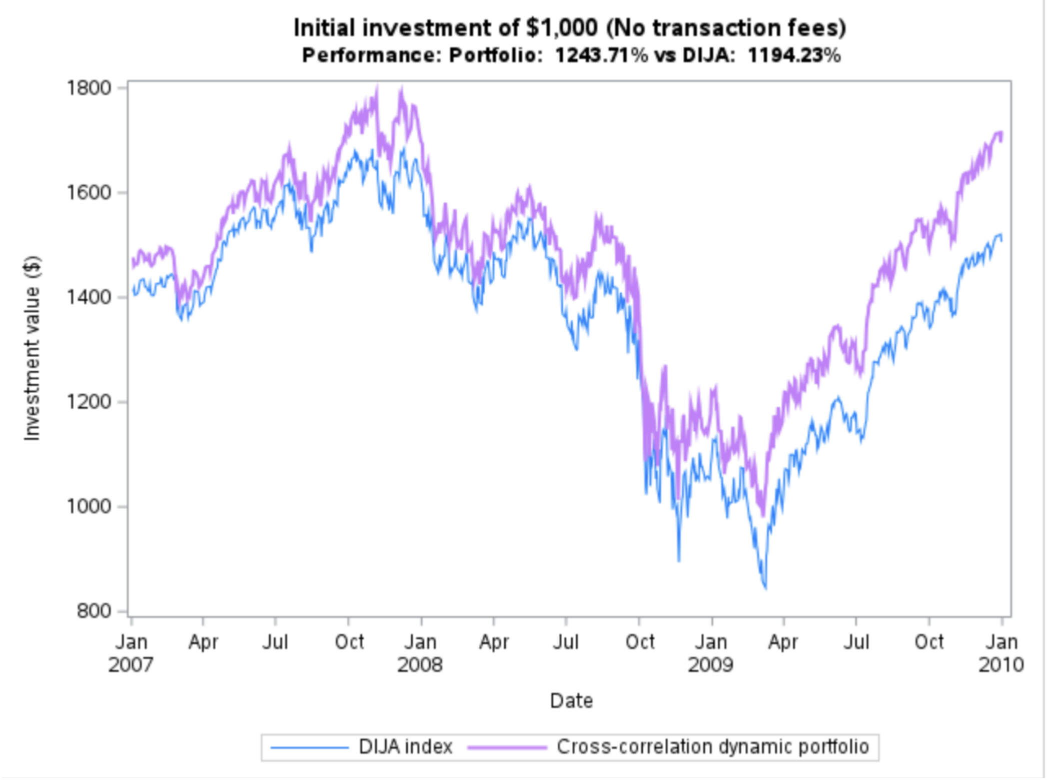
We obtain the following results:



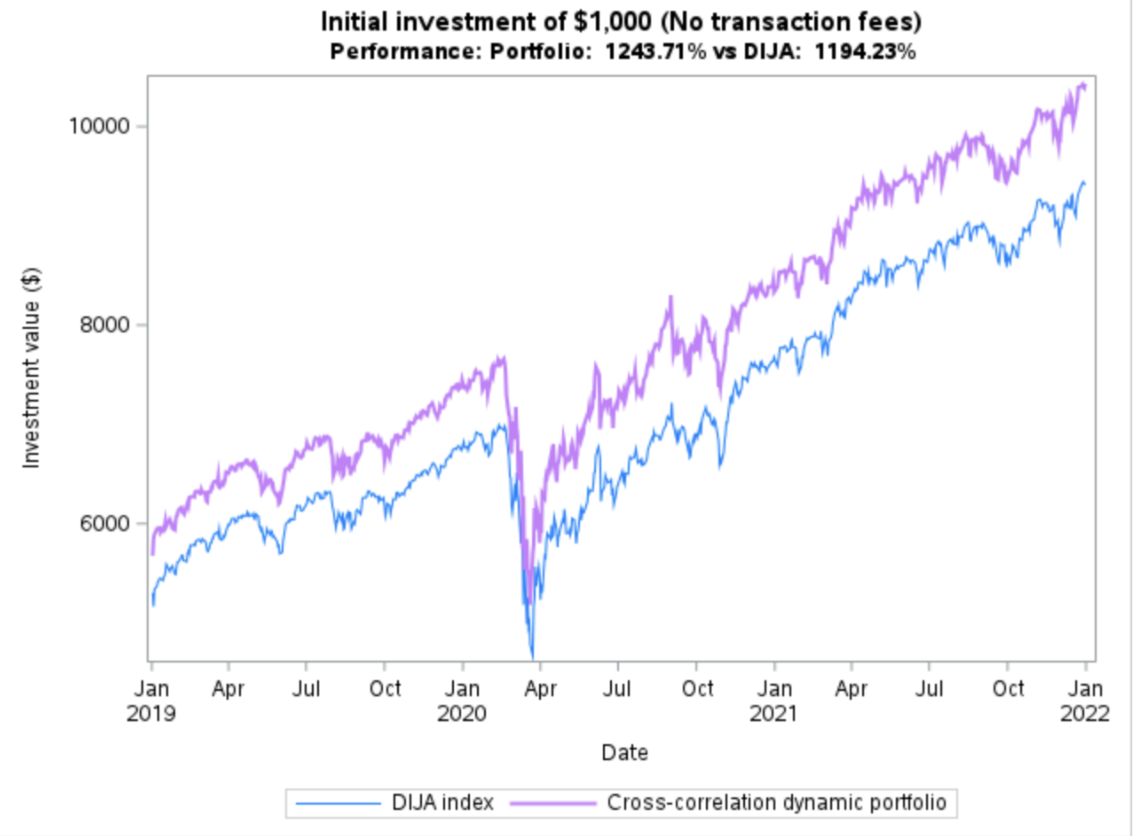
With a performance of 1243%, cross-correlation allows us to outperform the index.

We also want to check how our model performs during periods of financial crisis.

In 2008, we see a drop in values in both cases, but the index records losses, falling below $1,000, while our model's dynamic portfolio does not fall below the initial investment.

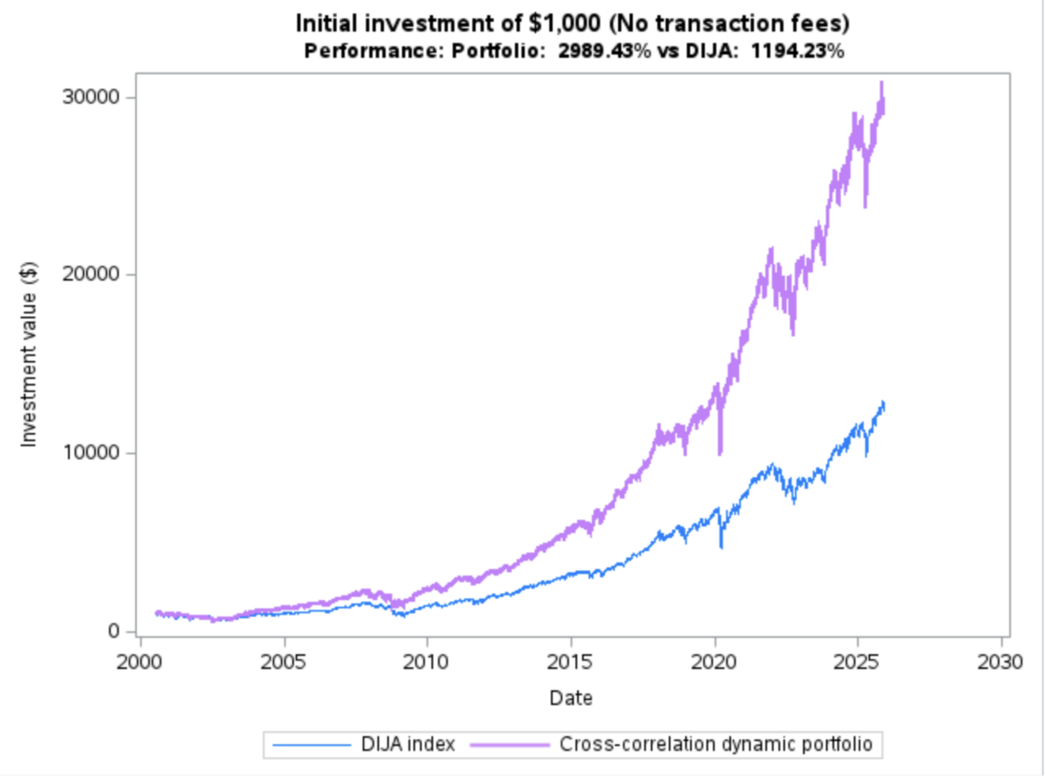


For the Covid crisis in 2020, we note that the value of our portfolio fell as sharply as the index. However, our portfolio recovered more quickly after the crisis.



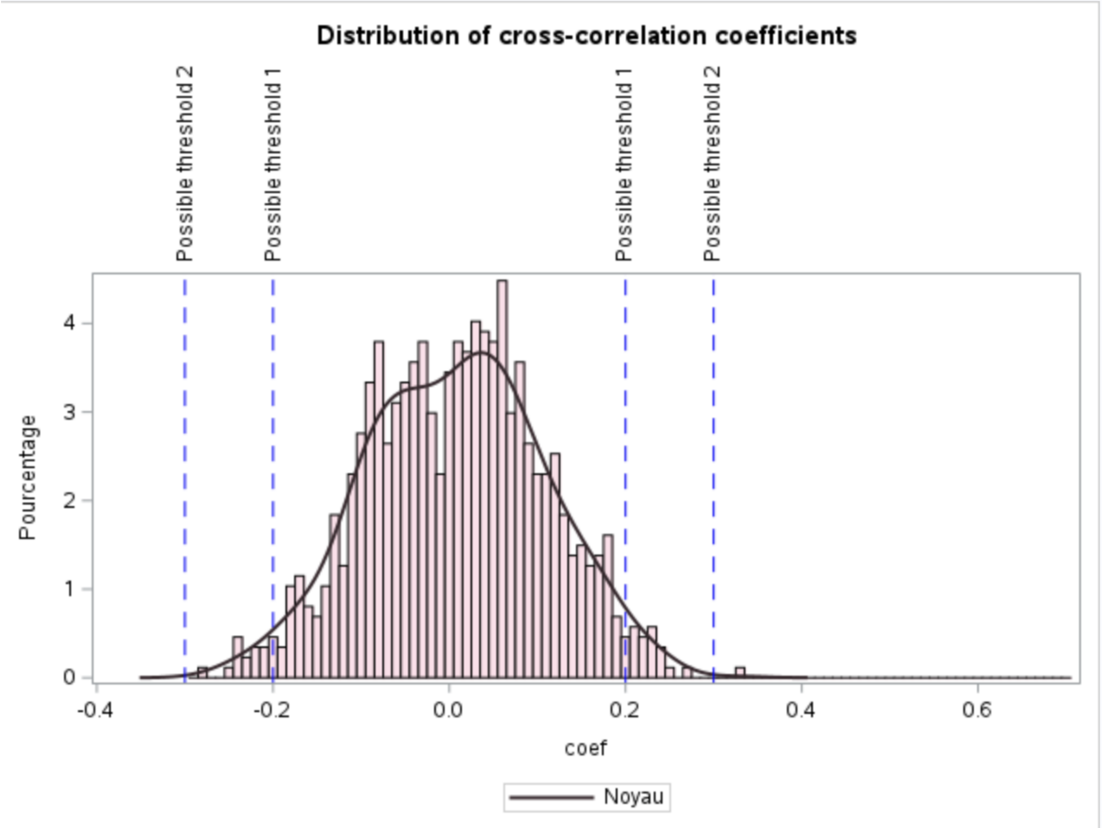
In order to improve our results, we try the same test but this time with an error threshold of 20%. Therefore, the rejection region for $H0$ changes from $QH(1) > 3.84$ to $QH(1) > 1.64$.

And we obtain the following results:

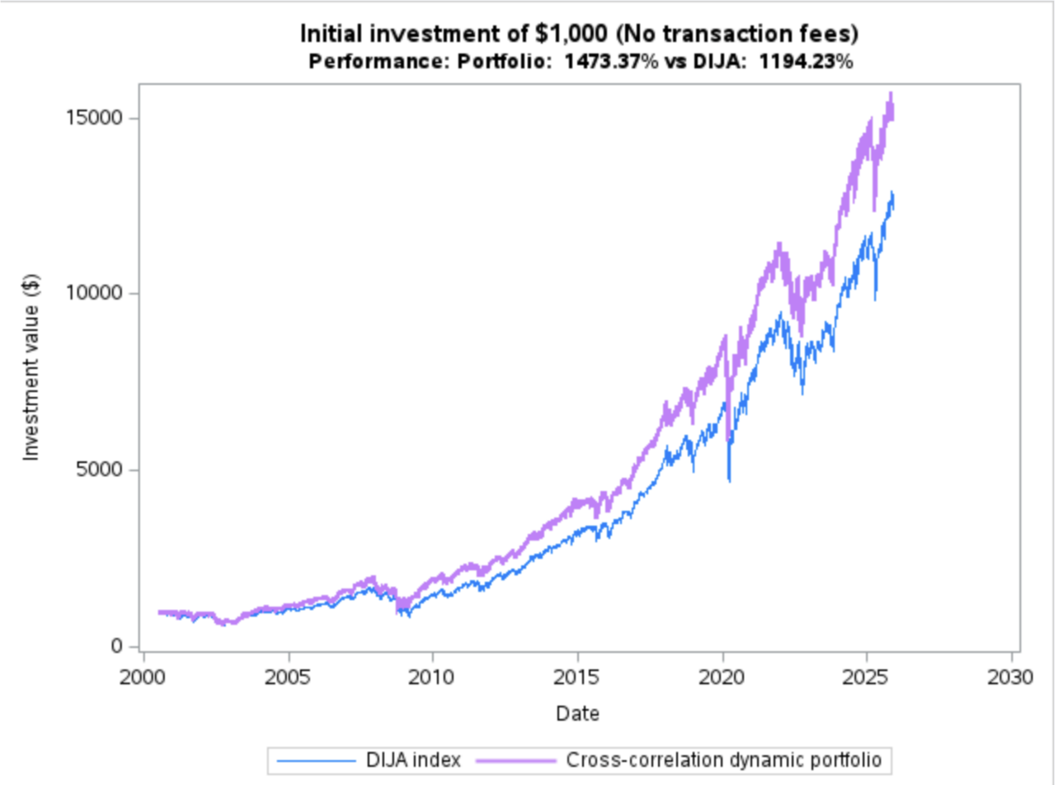


8.1.3 Cross-correlation : non parametric approach

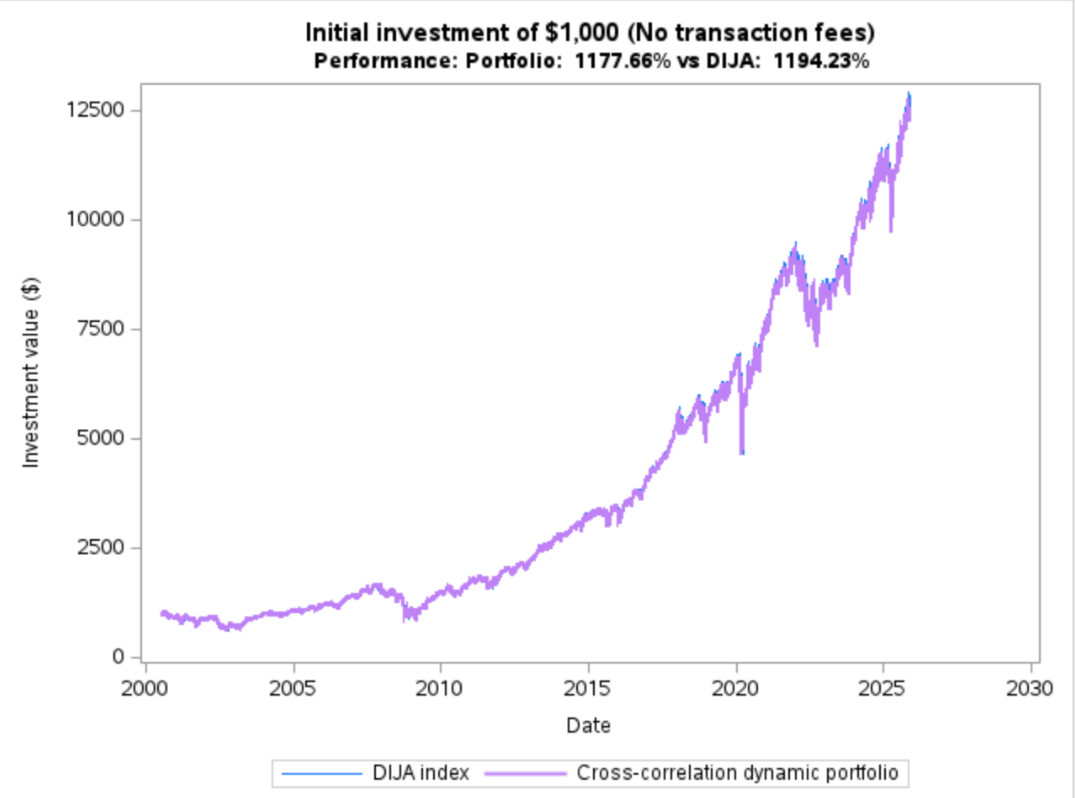
We now decide to choose a threshold such that if the absolute value of the cross-correlation is less than this threshold, then its place in the adjacency matrix will be 0, otherwise it will be 1.



For a threshold s=0.2, we obtain the following results:



And, for a threshold s=0.3, we obtain the following results:

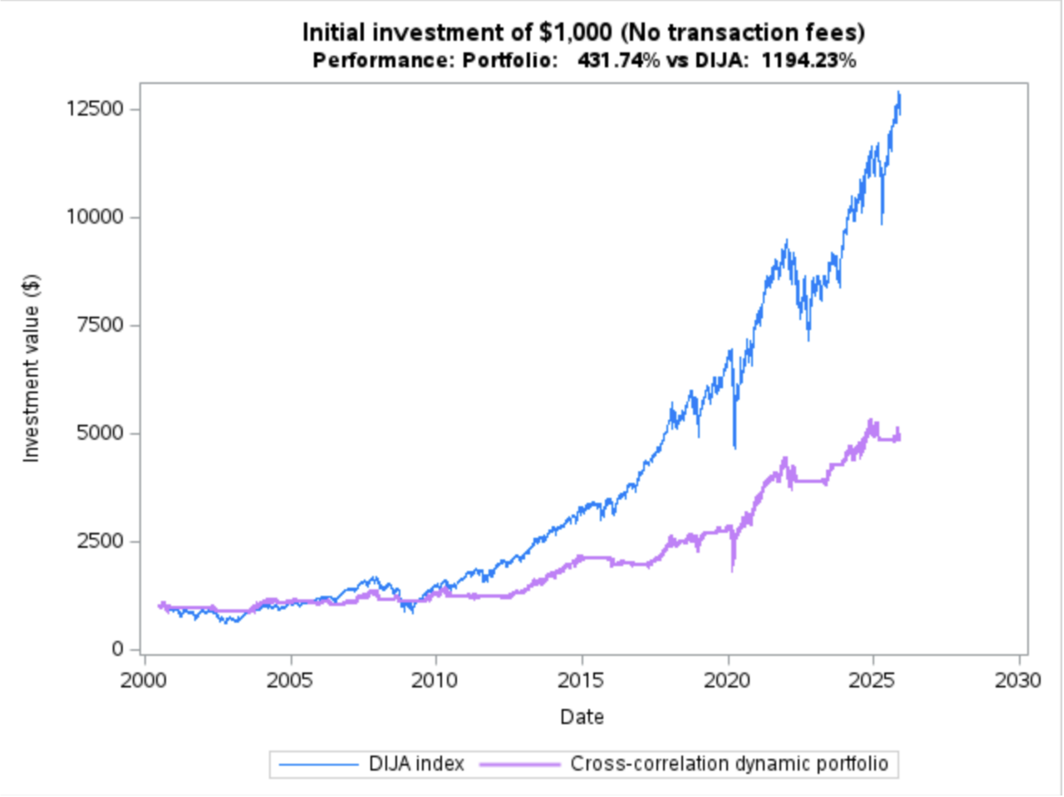


It is therefore important not to choose a threshold that is too high, otherwise even slightly correlated assets will be considered uncorrelated by the model, which will reduce the performance of our portfolio.

8.2 Not investing when portfolios are riskier than the index

We use the best method used so far, cross-correlation combined with the 20% threshold test.

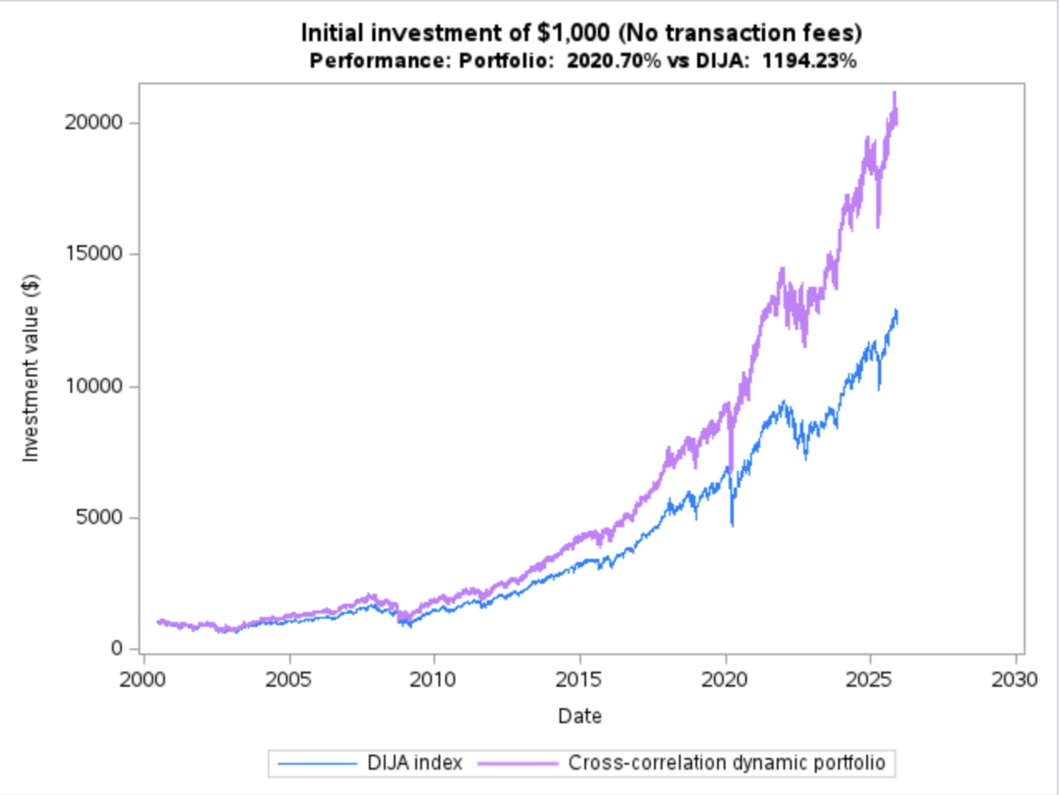
We obtain the following results:



We note that the risk of portfolio loss is lower than or equal to that of the index only over a few periods, so this investment method, although very low risk, is also very ineffective.

8.3 Investing in the index when portfolios are riskier than the index

We obtain the following results:



As seen above, a few periods have a risk of loss lower than or equal to that of the index, so our portfolio mostly tracks the index, but still performs better thanks to the periods with interesting risk of loss.

9 Conclusion

In conclusion, we observe that the best performance corresponds to the strategy of investing in a portfolio even if it is riskier than the index. This strategy is indeed riskier than the others, but it allows for much higher returns, and when we use cross-correlation and its parametric test at the 20% threshold, we are able to significantly outperform the index.

Despite the encouraging results of our model, we note that it has limits. Firstly, the network structure is calculated based on past data. In the event of a sudden market shock, correlations tend to converge towards 1, meaning that clustering will no longer function optimally at the very moment when it could be most relevant. We observed in fact that during the 2008 and 2020 crises, our portfolio continued to lose value at the same rate as the index.

Here, we only consider dependencies between linear assets, and therefore ignore all non-linear dependencies, which would be interesting to try in order to complete our model.

Furthermore, we are not taking transaction costs into account here. It would be interesting to consider brokerage fees, for example, which could completely change the results.

We could also generalize our model to other financial indices such as the MSCI.

Finally, the model invests equally in all selected assets, which is not optimal. Replacing equal weighting with equal contribution risk allocation, for example, could improve our results.

10 Annexe

10.1 Bibliography

References

[1] Régis Bourbonnais and Virginie Terreza. Analyse de séries temporelles cours et exercices corrigés - applications à l'économie et à la gestion. 5, 2022.

[2] Philippe De Peretti. Measures of association, 2025. Lecture notes, Master 1 in Econometrics and Statistics.

[3] Khalid El Himdi and Roch Roy. Tests for noncorrelation of two multivariate arma time series. *Statistical Society of Canada*, 25(2):233–256, 1997.

[4] Giorgio Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007.

[5] Christophe Hurlin. Un test simple de l'hypothèse de non-causalité dans un modèle de panel hétérogène. *Revue économique*, 56:799–809, 2005.

[6] Eric Sigward. Introduction à la théorie des graphes. 2002.

[7] Jean-Michel Zakoian. Modèles garch et à volatilité stochastique. *Université de Montréal*, 2007.

10.2 SAS code

10.2.1 Common code for every method

```
proc IML;
*Importation base de donn es des Log-Rendements;
use S.LR_DIJA;
read all var {
    log_return_3M
    log_return_American_Express
    log_return_Amgen
    log_return_Amazon
    log_return_Apple
    log_return_Boeing
    log_return_Caterpillar
    log_return_Chevron_Corp
    log_return_Cisco_Systems
    log_return_Coca_Cola
    log_return_Goldman_Sachs
    log_return_Home_Depot
    log_return_Honeywell_Internation
    log_return_IBM
    log_return_Intel
    log_return_Johnson_Johnson
    log_return_JPMorgan_Chase
    log_return_McDonald
    log_return_Merck_Co
    log_return_Microsoft
    log_return_NIKE
    log_return_NVIDIA
    log_return_Pfizer
    log_return_Procter_Gamble
    log_return_Sherwin_Williams
    log_return_Travelers_Companies
    log_return_UnitedHealth_Group
    log_return_Verizon_Communication
    log_return_Walt_Disney
    log_return_Walmart
} into X;
close S.LR_DIJA;

*Remplacer les 0 par la moyenne du jour pr c dant et du jour suivant;
N = ncol(X);
LR = X;
do j = 1 to N;
    do t = 2 to nrow(X)-1;
        if X[t,j] = 0 then do;
            if X[t-1,j] ^= . & X[t+1,j] ^= . then
                LR[t,j] = (X[t-1,j] + X[t+1,j]) / 2;
```

```
      end;
   end;
end;

*Mod le GARCH(1,1) sur tous les rendements;
LR_GARCH = j(nrow(LR), N, .);

do i = 1 to N;
    asset = LR[,i];
    create g_temp var {"asset"}; append; close g_temp;

    submit;
          proc autoreg data=g_temp noprint;
            model asset = / garch=(p=1,q=1);
            output out=out_g resid=eps cev=ht;
        run;
    endsubmit;

    use out_g; read all var {eps ht} into tmp; close out_g;
    LR_GARCH[,i] = tmp[,1] / sqrt(tmp[,2]);
end;

* Log-Rendement journalier de l'indice;
DIJA = LR[,+] / N;

* Fonction de calcul de la cross-corr lation;
start cross_cor(xi, xj);
    xi = xi [2: nrow (xi) ,];
    xj = xj [1: nrow (xj) -1 ,];
    cov = j (nrow(xi),1);
    do t =1 to nrow(xi);
        c = (xi[t])*(xj[t]);
        cov [t] = c;
    end ;
    cov_emp = sum (cov) / nrow (xi);
    stdi = std(xi);
    stdj = std(xj);
    c_cor = cov_emp/(stdi*stdj);
    return (c_cor);
finish cross_cor ;


* Fonction de calcul de Granger ;
start Granger(x,y);
    idxOK = loc( (x ^= .) & (y ^= .) );
    if ncol(idxOK) < 10 then return(0);

    x = x[idxOK];
    y = y[idxOK];

    x_one_lag = x[1:nrow(x)-1];
    y_one_lag = y[1:nrow(y)-1];

    y_t = y[2:nrow(y)];

    constant = j(nrow(y_t), 1, 1);

    Z  = constant || x_one_lag || y_one_lag;
    Z0 = constant || y_one_lag;

    bad = loc(Z = .);  if ncol(bad) > 0 then return(0);
    bad = loc(Z0 = .); if ncol(bad) > 0 then return(0);

    NC = ginv(Z‘*Z)  * Z‘  * y_t;
    C  = ginv(Z0‘*Z0) * Z0‘ * y_t;

    SCR_NC = (y_t - Z*NC)‘*(y_t - Z*NC);
    SCR_C  = (y_t - Z0*C )‘*(y_t - Z0*C );

    denom = SCR_NC / (nrow(y_t)-3);
    if denom <= 0 then return(0);

    statistics = (SCR_C - SCR_NC) / denom;
    chi2_1 = 2.71;
    coef = (statistics > chi2_1);
    return(coef);
finish Granger;

* Boucle de gestion des portefeuilles dans le temps;
portefeuilles_finaux = {};

do k =1 to nrow(LR) by 22*2;
    mat_in_sample ={};
    mat_out_sample={};

    *condition de sortie de boucle;
    if k + 22*8 > nrow(LR) then leave;

    *matrice des log-rendements avec GARCH pour la p riode in_sample (6mois);
    mat_in_sample = LR_GARCH[k : k +22 *6  ,];

    *matrice des log-rendements pour la p riode out_sample (2mois);
    mat_out_sample = LR[ k + 22 *6+1 : k + 22*8  ,];


    *Corr lation crois e;
    C = j(N, N, .);
    do i = 1 to N;
        do j = 1 to N;
            if i ^= j then C[i,j] = cross_cor(mat_in_sample[,i], mat_in_sample[,j]);
        end;
    end;

    *test param trique  5%;
    T = nrow(mat_in_sample);
    QH = j(N, N, .);
    do i = 1 to N;
        do j = 1 to N;
            if i ^= j then QH[i,j] = (T##2 / (T-1)) * (C[i,j]##2);
        end;
    end;
    A_CC = j(N, N, 0);
    do i = 1 to N;
        do j = 1 to N;
            if i ^= j then do;
                if QH[i,j] > 3.84 then A_CC[i,j] = 1;
            end;
           end;
        end;

/* Changer le commentaire de place en fonction de la m thode choisie */
/*
    * Granger;
    A_CC = j(N, N, 0);
    do i = 1 to N;
        do j = 1 to N;
            if i ^= j then A_CC[i,j] = Granger(mat_in_sample[,i], mat_in_sample[,j]);
        end;
    end;
*/


       *degr s;
       d_in={};
       d_out={};
       d_bi={};
    d_in  = A_CC[+,];
    d_out = A_CC[,+]`;
    d_bi  = vecdiag(A_CC*A_CC`);

    *Triangles th oriques;
    T_in={};
    T_out={};
    T_cyc={};
```

```
T_mid={};
T_in  = d_in  # (d_in  - 1);
T_out = d_out # (d_out - 1);
T_cyc = d_in  # d_out - t(d_bi);
T_mid = T_cyc;

*Triangles observ s;
t_in_p={};
t_out_p={};
t_cyc_p={};
t_mid_p={};
t_in_p  = vecdiag(t(A_CC)*A_CC*A_CC);
t_out_p = vecdiag(A_CC*A_CC*t(A_CC));
t_cyc_p = vecdiag(A_CC*A_CC*A_CC);
t_mid_p = vecdiag(A_CC*t(A_CC)*A_CC);

*clustering local;
CC = j(N,4,0);
idx1={};
idx2={};
idx3={};
idx4={};
idx1 = loc(T_in  > 0);  if ncol(idx1)>0 then CC[idx1,1] = t_in_p [idx1] / T_in[idx1];
idx2 = loc(T_out > 0);  if ncol(idx2)>0 then CC[idx2,2] = t_out_p[idx2] / T_out[idx2];
idx3 = loc(T_cyc > 0);  if ncol(idx3)>0 then CC[idx3,3] = t_cyc_p[idx3] / T_cyc[idx3];
idx4 = loc(T_mid > 0);  if ncol(idx4)>0 then CC[idx4,4] = t_mid_p[idx4] / T_mid[idx4];
*print(CC)[colname={"CC_IN" "CC_OUT" "CC_CYC" "CC_MID"}];

*moyennes;
m_CC ={};
m_CC = CC[:,]`;
*print(t(m_CC))[colname={"m_IN" "m_OUT" "m_CYC" "m_MID"}];

*clustering global;
S ={};
d_tot ={};
tri ={};
idx ={};
S = (A_CC + t(A_CC)) > 0;
d_tot = S[,+]`;
tri = vecdiag(S*S*S) / 2;

Cg = j(N,1,0);
idx = loc(d_tot > 1);
if ncol(idx)>0 then Cg[idx] = tri[idx] / (d_tot[idx] # (d_tot[idx]-1));

CC_global={};
CC_global = mean(Cg);
```

```
*print(CC_global);

*Downside du BenchMark (indice DIJA);
DIJA_in = mat_in_sample[, +] / N;
neg_DIJA_in = DIJA_in[loc(DIJA_in < 0)];
if isempty(neg_DIJA_in) then downside_risk_DIJA_in = 0.001;
else downside_risk_DIJA_in = sqrt(mean(neg_DIJA_in##2));

seuil = t(do(0.01, 1, 0.01));
resultats = j(4, 2, .); /* Col 1: Ratio, Col 2: Seuil */

 do t = 1 to 4;
     max_ratio_t = -1e99;
     best_s_t = .;

     best_abs_sortino = -1e99;
     best_abs_s = .;

     do s = 1 to nrow(seuil);
          indices = loc(CC[, t] < seuil[s]);
          if ncol(indices) < 3 then continue;

          Rp_in = mat_in_sample[, indices][, :];
          neg_in_idx = loc(Rp_in < 0);

          if isempty(neg_in_idx) then risk_in = 0.0001;
          else risk_in = sqrt(mean(Rp_in[neg_in_idx]##2));

          sortino_in = mean(Rp_in) / risk_in;

          if sortino_in > best_abs_sortino then do;
               best_abs_sortino = sortino_in;
               best_abs_s = seuil[s];
          end;

          if risk_in <= downside_risk_DIJA_in then do;
               if sortino_in > max_ratio_t then do;
                    max_ratio_t = sortino_in;
                    best_s_t = seuil[s];
               end;
          end;
     end;

     if max_ratio_t = -1e99 then do;
          max_ratio_t = best_abs_sortino;
          best_s_t = best_abs_s;
     end;
```

```
        resultats[t, 1] = max_ratio_t;
        resultats[t, 2] = best_s_t;
    end;

    sortie = resultats || t(1:4);
    call sort(sortie, {1});

    final_ratio = sortie[4, 1];
    final_seuil = sortie[4, 2];
    final_cc = sortie[4, 3];


    indices_finaux = loc(CC[, final_cc] < final_seuil);

    portefeuille_f = mat_out_sample[, indices_finaux];
    portefeuille_m_jour = portefeuille_f[, :];

    portefeuilles_finaux = portefeuilles_finaux // portefeuille_m_jour;
end;

create portefeuille var {"Log_return"};
append from portefeuilles_finaux;
close portefeuille;

debut_DIJA = 22*6 + 1;
nb_jours_disponibles = nrow(DIJA) - debut_DIJA + 1;
nb_jours_portefeuille = nrow(portefeuilles_finaux);
fin_indice = debut_DIJA + min(nb_jours_disponibles, nb_jours_portefeuille) - 1;
DIJA_final = DIJA[debut_DIJA : fin_indice];
create BenchMark var {"Log_return_DIJA"};
append from DIJA_final;
close BenchMark;

quit;

/*Graphique de comparaison entre nos portefeuilles et l'indice (benchmark) pour un
investissement   quipondr   initial de 1000$ sans co t de transaction*/
*DIJA sans les premiers 6mois d'analyse;
data DIJA_out_sample;
    set S.LR_DIJA;
    if _n_ > 6*22;
run;
*base de donn es avec les log-rendement journalier du DIJA et de notre portefeuille;
data comparaison;
    merge DIJA_out_sample(keep=Date)
          portefeuille(rename=(Log_return=LR_Portefeuille))
          BenchMark(rename=(Log_return_DIJA=LR_DIJA));
run;
```

```
*Investissement de 1000$ dans notre portefeuille et dans le DIJA;
data valeur_1000;
    set comparaison;
    retain Portefeuille_Value 1000 DIJA_Value 1000;

    Portefeuille_Value = Portefeuille_Value * exp(LR_Portefeuille);
    DIJA_Value = DIJA_Value * exp(LR_DIJA);

    label Portefeuille_Value = "Valeur Portefeuille ($)"
          DIJA_Value = "Valeur Indice DIJA ($)";
run;

*Affichage des r sultats;
proc sgplot data=valeur_1000;
    title "Initial investment of $1,000 (No transaction fees)";

    series x=Date y=DIJA_Value / lineattrs=(color='#0080FF' thickness=1)
                                 legendlabel="DIJA index";
    series x=Date y=Portefeuille_Value / lineattrs=(color='#C97EFD' thickness=2)
                                         legendlabel="Cross-correlation dynamic portfolio"

    xaxis label="Date";
    yaxis label="Investment value ($)";
run;
```