ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Financial big data

# Trading strategies backtesting

*Authors:*
Ridha Chahed
Haitham Hammami

*Supervisors:*
Dr. Damien Challet

Lausanne, January 2021

**Abstract**

**In this project we make use of the large amount of data of the stock prices in the *S&P 100* during the period of 2004-2008, in order to develop and backtest different trading strategies by performing robust estimate techniques for the covariance matrix of the stocks. Furthermore, we explore the pair trading arbitrage method and benchmark it from 2004 until the subprime crisis.**

# Table of contents

# 1 Data

## 1.1 Data set

We study financial data from the *S&P 100* for the period 2004-2008. We have at our disposal intraday prices, traded volumes and best bid and offer of 85 stocks with a total aggregated file size of 24GB. The trade data is structured as one file per stock and day totaling to 103 336 files with an average size of 200KB. The size of the data makes the analysis challenging but allows us to gain in statistical power.

## 1.2 Data wrangling

We start by extracting the data and aggregating it per company. We decide only to keep the data during the market trading hours New York time. For a same transaction we may sometimes have different entries with the same timestamp. We resample the data with 1 minute frequency, keeping the first occurrence in the given minute.

For the missing values, we apply a forward fill to keep the causal property of the time series data, but we still have the edge case of the first trading day in 2004 where we found two different possibilities. First, we have companies that have a lot of missing values. (more than 100 000) due to the fact that they are not presented at the beginning of 2004 but are rather listed in the S&P 100 way after. Six stocks are in this situation : Devon Strategy, Mastercard, Visa, Morgan Stanley, National Oilwell Varco and Philip Morris International, the latter is due to the fact that it got separated from its parent company Altria. We decide to remove these stocks. Secondly, we also have very few stocks with some missing trading price at the first minutes (maximum 8) of trading of the opening. For those stocks we apply a back fill breaking the causal relation for few minutes.

We also inspect the data and look for outliers. We found that several stock prices have big swings in their prices overnight (see Fig.6). Looking on the internet (on macrotends and split history) for each stock we discover two reasons. Mainly those huge differences are due to 2-1 splits. If an investor owns a number of shares of these companies stocks before the split, he will own twice as many shares after the split. So there is in fact no change in the total market value but only in the price. We decide to back-adjust the

prices before the ex-date of the split by dividing them by 2. Moreover, we also find that for some companies like Altria and Sprint. The difference is due to spin-off and mergers, we decide to remove those stocks.

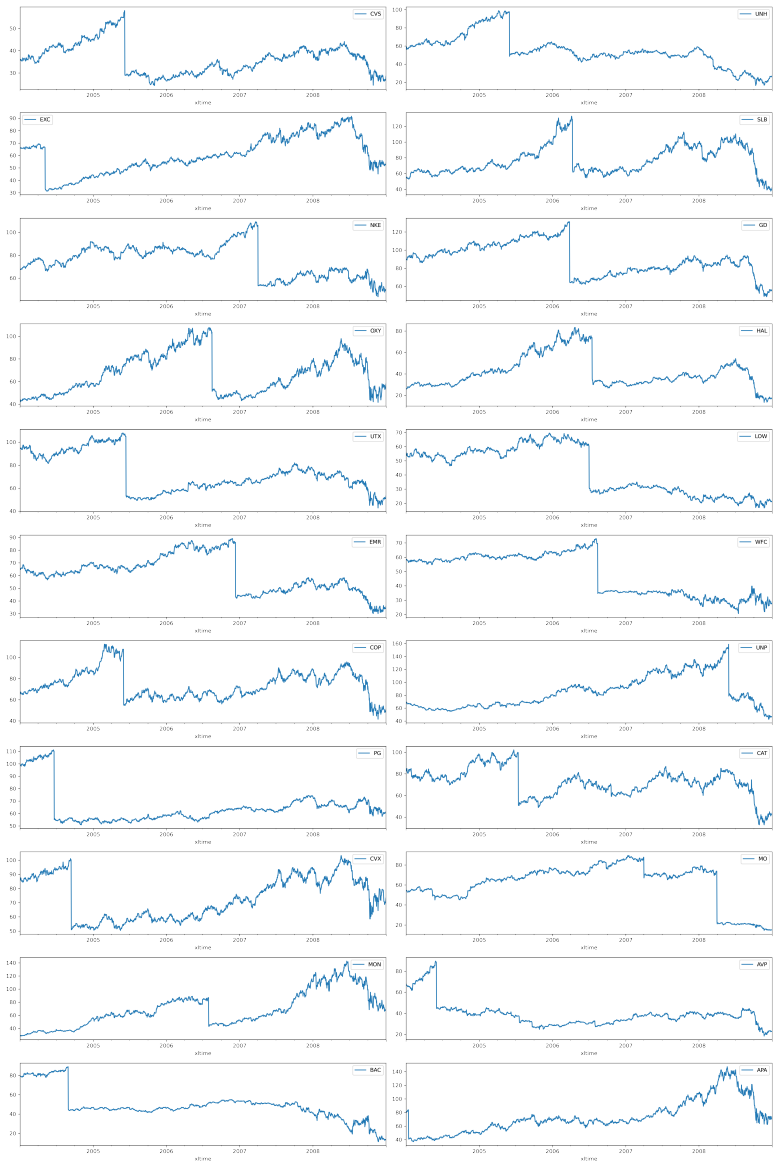



Figure 1: Overnight swings in some stock prices

# 2 Methods

## 2.1 A naive approach, hold the market

We begin with the first strategy, which is basically holding the equally weighted portfolio of the remaining 77 stocks. This strategy will serves as a baseline for other more complex strategies.

## 2.2 Value weighted portfolio

For this strategy, we apply a daily rolling calibration so we first resample our data to have only the closing price of each day. We also need the market capitalization of each stock, therefore we download the data we need from wrds by first downloading the companies information in order to create a mapping between ticker symbols and *permco* which is the permanent company identifier used in their database. Then we extract the daily shares outstanding for each company and compute its market cap by multiplying the number of shares by the daily and from that we get the portfolio weights by dividing by the sum of the companies' capitalization. We perform the rolling window calibration with this technique and we adjust the weights accordingly.

## 2.3 Mean variance approach

The Markowitz's mean-variance model [4] is a portfolio optimization model that determines the most efficient portfolio for a given risk aversion. Here we will construct the global minimum-variance portfolio, which has as this set of weights:

$$\mathbf{w} = \frac{\sum^{-1} \mathbf{1}}{\mathbf{1}^T \sum^{-1} \mathbf{1}} \tag{1}$$

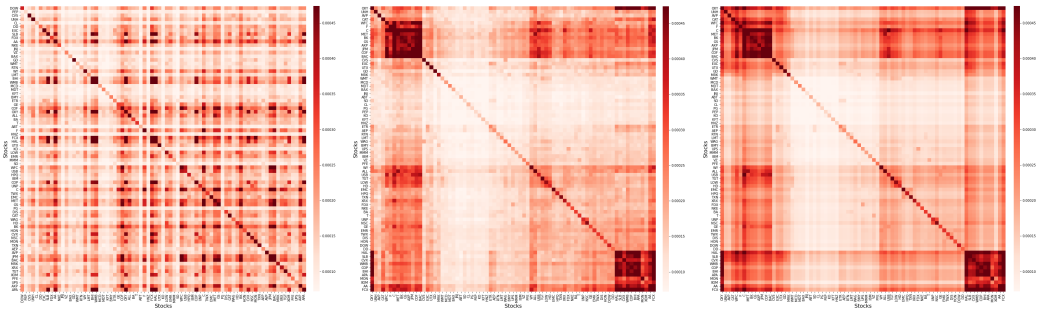
In order to compute the optimal weights, we first need to correctly estimate the covariance matrix, a tricky task as we will be performing a rolling calibration with 90 and 200 days as a window size, this limits the statistical significance of the estimation and give noise estimators. To overcome this problem, we implement the strategy using two different kinds of estimation:

### 2.3.1 Statistical estimation of covariance

As an initial attempt, we compute the covariance matrix with the standard formula $cov_{x,y} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N-1}$. We obtain the following matrix for the full sample in (Fig.2 a). As expected, it appears to be noisy. To better grasp the underlying structure, we reorder the matrix by applying a hierarchical clustering and average linkage (HCAL) [5]. The results can be seen in (Fig.2 b).

### 2.3.2 Robust estimation of covariance: BAHC

Bootstrapped Average Hierarchical Clustering (BAHC) [6] is a method based on computing bootstraps of the price return matrix and for each bootstrap, compute the HCAL filtered matrix [7]. This method yields an even more filtered matrix while maintaining the non-diagonal structures as we can see in (Fig.2 c).



(a) Unfiltered matrix (b) Reordered matrix with HCAL (c) Filtered matrix with BAHC

Figure 2: Covariance matrices

### 2.3.3 Implementation

For both strategies we compute the covariance matrix for each window and then derive the corresponding portfolio weights.

## 2.4 Pair trading, a mean reversion strategy

Mean reversion is the idea that a process is randomly distributed around its mean, so an exceptionally good performance is very likely to be followed by

performances that are closer to the average. One famous strategy relying on this idea is the pair trading strategy.

Pair trading is an arbitrage strategy where two assets are traded. Those two assets are chosen such that they tend to move together. If a detachment between the two stocks is observed the trader holds a long position on one stock while shorting the other with the expectation that the spread will revert to its mean. The strategy is composed of two windows of actions, first we need to filter and select pairs of stocks and then we need to apply the strategy.

### 2.4.1 The pair selection and estimation phase

The first question that needs to be answered is how do we measure that two assets 'move together' ? When we think of comparing the movements of two processes, the first metric that comes to mind is correlation that will measure the degree of linear relation. Unfortunately, correlation presents several downfalls. It has a short memory and it fails to capture the magnitude of the movements so the two stocks may move similarly in their sense but with totally different magnitude. Cointegration helps overcome this problem as it tells whether the distance between the 2 stocks remains the same over time.

Mathematically, the pair of assets $X_t$ and $Y_t$ is said to be cointegrated if :

- They are I(1) processes, by differencing them we get stationary series.
- Their linear combination $Z_t = Y_t - \alpha X_t - b$ is stationary

The Engle and Granger procedure allows us to test if two assets are cointegrated:

1. We first need to estimate the cointegrating relation. So we regress $X_t$ over $Y_t$ using an ordinary least squares regression.
2. We then test the residual for stationarity using the Augmented Dickey Fuller (ADF) test.

Some precision on the Augmented Dickey Fuller test.

$\Delta Z_t = \beta + \sigma Z_{t-1} + \delta \Delta Z_{t-1} + \epsilon_t$

- Null hypothesis $X_t$ and $Y_t$ are not cointegrated: $\delta = 1$

- Alternative hypothesis $X_t$ and $Y_t$ are cointegrated : $\delta < 1$

We use a t-stat for the delta.

With cointegrated assets, we can expect a mean-reverting behaviour. When the spread is positive we should therefore go long with the stock $X_t$ and short $Y_t$ (and it is the opposite when the spread is negative).

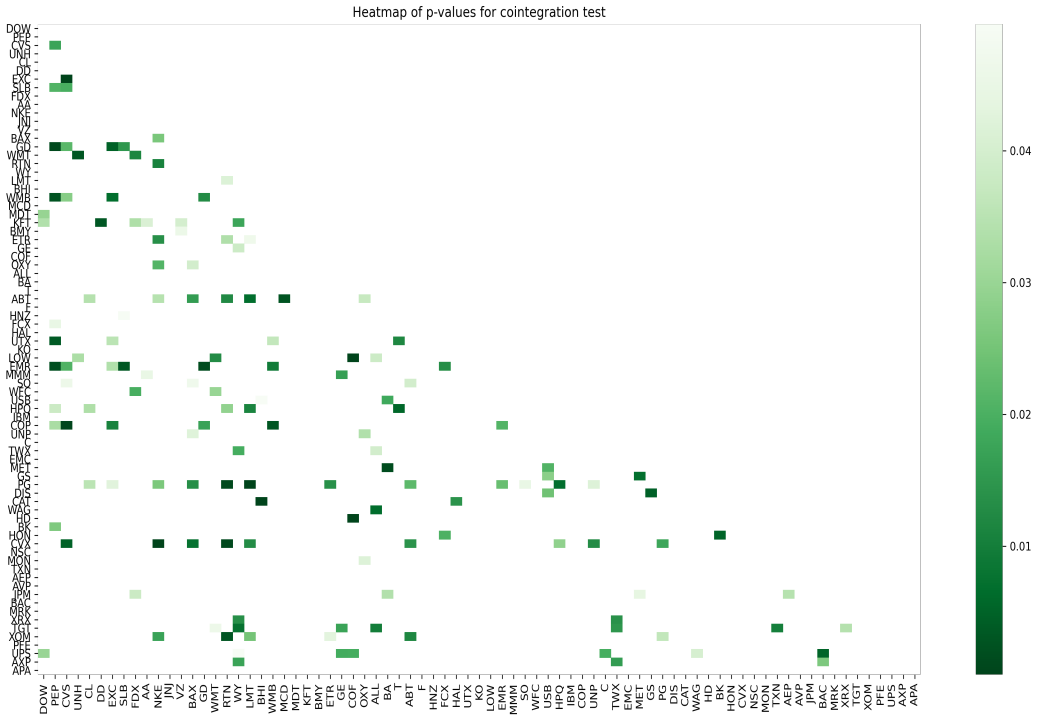



Figure 3: Heatmap of p-values for cointegration test

### 2.4.2 The strategy implementation phase

Entering the strategy window we should have selected our cointegrated pairs and for each pair we have an estimated $\alpha$ and $\beta$.

Using the last day of the estimation window we compute the Hedge ratio $H = \alpha \frac{X_{t_N}}{Y_{t_N}}$. The hedge ratio represents the amount that we need to invest in the stock X for each dollar invested in Y.

The strategy consists of computing the spread and if it diverges too much from it's empirical mean we need to act in consequence. To facilitate the notation, we compute and use z-score with the empirical mean and the empirical standard deviation.

The strategy takes 3 parameters:

- Investment threshold
- Exit threshold
- Loss threshold

Investment strategy:

- If z-score $>$ Investment threshold : Short the spread, go long with X and short Y. For each dollar invested in Y we need to invest H on X.
- If z-score $<$ Investment threshold : Long the spread, short X and go long with Y.

One parameter that we add in order to invest is the spread trend of the past 5 days. Indeed, we would like to make sure that the spread is reverting to its mean before taking a position.

Exit strategy:

- |z-score| $<$ Exit threshold : The spread is close to its mean we can exit the position.
- |z-score| $>$ Loss threshold : Cut the losses and exit the position.

# 3 Results

As discussed above, we have implemented seven different trading strategies. The recap is in table 1

| Trading Strategy | Window Size | Step Size |
|---|---|---|
| Hold the market | - | - |
| Value Weighted | 1 day | 1 day |
| Mean Variance unfiltered 90 | 90 days | 1 day |
| Mean Variance unfiltered 200 | 200 days | 1 day |
| Mean Variance with BAHC 90 | 90 days | 1 day |
| Mean Variance with BAHC 200 | 200 days | 1 day |
| Mean reversion | 90 days | 90 days |

Table 1: Portfolio calibration parameters

For the first six strategies, we create a portfolio in which we invest one dollar at the beginning then we measure each portfolio's performance during the whole period. The results are shown in figure 5.
As we can see, the value weighted portfolio under-performs compared to the equally weighted one right until the market crash, and outperforms it right after it. This is a good demonstration of the role of dynamic allocation for limiting losses in catastrophic events.
For the unfiltered mean variance portfolios, we see that the performance is quite volatile depending on the window size; the earnings are rather poor when using 90 days whereas using 200 days makes the portfolio comparable to the first two. We don't find this issue for the BAHC portfolio as it proves itself to be robust to the change of the window size, even though it is far from beating the baseline. We also remark that the BAHC 90 is the least volatile in the daily returns in Fig 4.

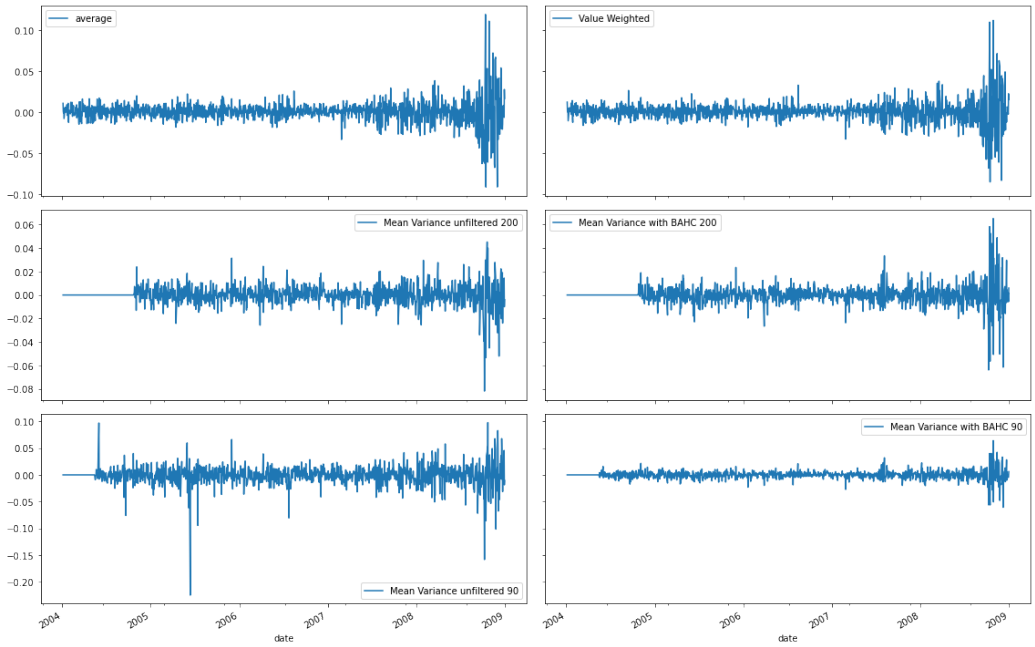



Figure 4: Portfolio values

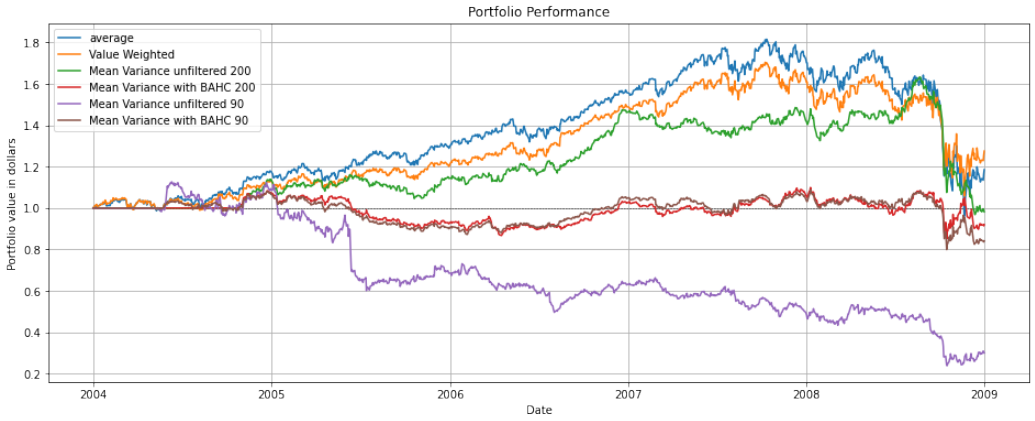



Figure 5: Portfolio daily return

For the pair trading strategy the evaluation is a bit different as it's an arbitrage strategy. We have almost a null dollar market exposition as when one stock is sold, a proportional amount of the other stock is purchased (using

the hedge ratio estimation).

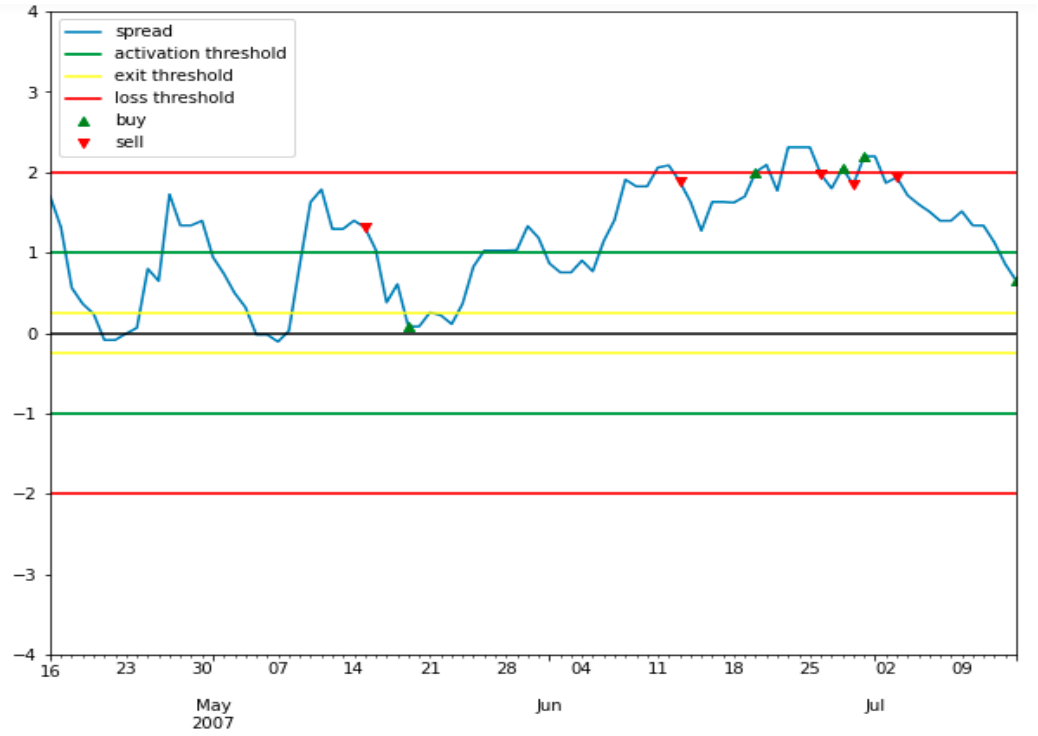



Figure 6: Backtesting during a period of 90 days of a pair trading strategy for the pair CVS and Pepsico.

The strategy is divided in two parts with a calibration window followed by a back testing window.

We have several parameters to set that control our trading strategy notably the activation threshold = 1, the loss threshold = 2, the exit threshold = 0.25 and the window size of the spread trend = 5 days. We can see the results in figure 2. A winning pair being a pair where the accumulated profit and loss is positive.

| Start backtesting | Pnl ($) | Winning pair | Losing pair |
|---|---|---|---|
| 2004-04-01 | 19 | 138 | 105 |
| 2004-06-30 | -25 | 60 | 73 |
| 2004-09-28 | -61 | 99 | 116 |
| 2004-12-27 | 33 | 191 | 148 |
| 2005-03-27 | -55 | 94 | 101 |
| 2005-06-25 | -72 | 64 | 86 |
| 2005-09-23 | 10 | 127 | 83 |
| 2005-12-22 | -60 | 113 | 104 |
| 2006-03-22 | 62 | 136 | 99 |
| 2006-06-20 | 55 | 116 | 84 |
| 2006-09-18 | -62 | 106 | 88 |
| 2006-12-17 | 2 | 117 | 101 |
| 2007-03-17 | -13 | 74 | 64 |
| 2007-06-15 | 55 | 121 | 103 |
| 2007-09-13 | -166 | 183 | 160 |
| 2007-12-12 | 0 | 112 | 85 |
| 2008-03-11 | -127 | 87 | 77 |
| 2008-06-09 | -501 | 156 | 195 |
| 2008-09-07 | -441 | 132 | 141 |

Table 2: Pair trading backtesting results with a window of 90 days

We can see that the arbitrage method performs very poorly during the 2008 crisis. By modifying the parameters we get different profit and losses for the pre-crisis period but consistent big losses starting from September 2007. It's difficult to have robust parameters that work well for each pair. We can also observe that the holding horizon, that is the expected time for the spread to revert, depends on each pair, but this isn't taken into account here. Our conclusion is that each pair should be considered individually since aggregating them yields low returns and high volatility.

# 4 Conclusion and discussion

In this project we got familiar with handling large amounts of financial data and implementing several trading algorithms. We have concluded that in

order to build a robust trading strategy, we need robust estimates of the data. In our case it's the covariance matrix that suffers from the noisy data, which is why it is essential to perform filtering techniques on it, and BAHC proves itself to be the right tool for the task.
Despite the solid results we achieved, we are aware of several gaps that our work suffer from. First of all the choice of stocks that we are working with implies a survivorship bias. Indeed all the stocks were and have remained in the *S&P 100* during the 5-years-period meaning that they have relatively survived the crash ( compared to, for example, The Lehman Brothers who was also in the index in 2008 ). We also find a discrepancy in the choice of the window size, as the performance of the Mean Variance portfolio and the pair trading strategy are heavily affected by it. Moreover, these strategies don't take into consideration the transaction costs, which can also alter the weight allocation of the portfolios.

# References


[1] Patrick McSharry. *Efficient Pair Selection for Pair-Trading Strategies*. Oxford.

[2] Ramos-Requena, J.P.; Trinidad-Segovia, J.E.; Sánchez-Granero, M.Á. *Some Notes on the Formation of a Pair in Pairs Trading*. Mathematics 2020, 8, 348.

[3] Ernest P. Chan. *Algorithmic Trading : Winning strategies and their rationale*.

[4] George Pennachi *"Theory of Asset Pricing"* chapter 2. 2007

[5] C. Bongiorno, S. Miccich‘e, and R. N. Mantegna, "Nested partitions from hierarchical clustering statistical validation," (2019), arXiv preprint arXiv:1906.06908.

[6] C. Bongiorno and D. Challet, "Covariance matrix filtering with bootstrapped hierarchies" (2020) arXiv:2003.05807.

[7] D. Challet, *Financial Big Data lecture slides*. EPFL 2020