École Polytechnique Fédérale de Lausanne

Log Returns

Milestone 1

Etienne Salimbeni, Carlo Musso, Filippo Salmina

7th April 2022

Context

Good visualization is a very powerful tool to gain insights into large data sets. For this reason, we decided to engage in an ongoing real-world project which would highly benefit from an ad-hoc visualization tool. We contacted an EPFL associated professor Damien Challet with the main affiliation from the Centrale Supélec School in Paris.

Prof. Damien's finance data-driven lab is working on reviewing the state of the art of clustering financial time series and the study of their correlations on an interaction network. The goal is to build an interactive visualization tool to study correlations, networks and clustering in financial markets.

Interactive visualization tool for correlation of financial markets using graph theory

Dataset

The dataset provided by prof. Challet consists in a wide table of **log returns**. Namely, each column represents a different stock (ex. Amazon, Apple, Tesla, etc.), and each entry of the corresponding column has the value of the log return of day i.

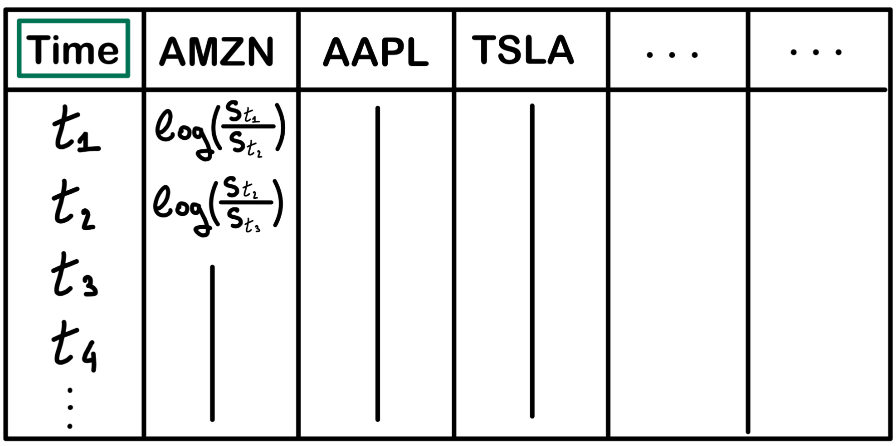
In more detail

$$r = \frac{S_t}{S_{t+1}}$$

the return (r) is the ratio between the stock price (S) at time t (market opening) and time $t+1$ (market closing). The **log return** is the return value in logarithmic scale.

$$log\_r = log(r)$$

A broad, comprehensive representation of the data set is given in the below diagram:



Note - currently prof. Challet provided us with a small but still significant sample of the whole dataset. Yet, soon the application back-end will be able to communicate with the finance data-driven lab's server to provide the entire dataset's availability.

Problematic

Correlation between stock prices gives powerful insights into the evolution of financial systems in interesting periods: during a financial crisis, political instability, or significant economic development of a country. Correlation tells us a lot about how stocks are linked, such as how industries are dependent on or independent of crisis or development in specific economic sectors. The notion of stock independence has numerous applications: an instrumental one is portfolio differentiation for risk minimization.

The goal of this visualization is to **represent the evolution in time of correlation, hierarchies and clustering of various financial stocks**. Visualization is done through graphs and basic graph theory principles, focusing on per-sector analysis. In our visualization, we **represent stocks as nodes of a graph and their corresponding correlation as the weight of the edges connecting them**.

Anyone interested in markets and stock valuation in current times and given historical moments would gain powerful insights into how stock prices interact, both from a fine-grained and a high-level-sector perspective.

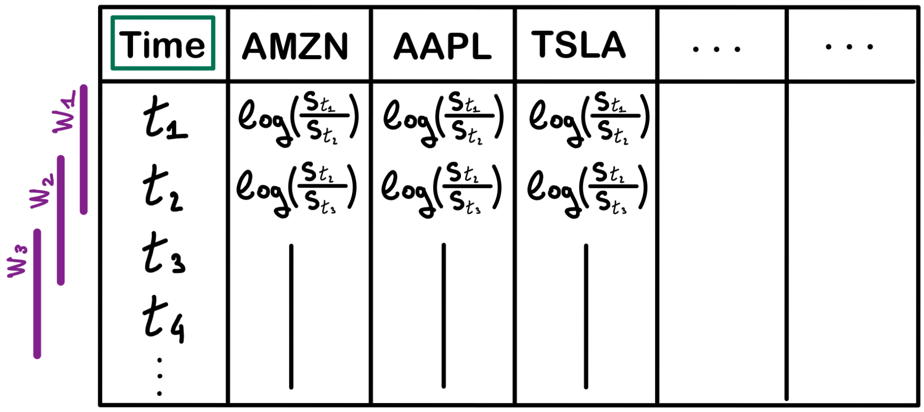
Exploratory Data Analysis - Data processing

Correlation Matrix

In order to obtain the correlation values of each pair of stocks, we select two stocks (S_1 and S_2) and a window (W_1) of size n on the data set (where n is the number of days which the window will represent) and compute the Pearson correlation in log returns between the two selected stocks in the window time-frame. If we repeat this process for each pair of stocks in our data set, we can structure our **correlation data-type** as a correlation matrix as follows:

	AMZN	AAPL	TSLA	...
AMZN	1	corr(AMZN,AAPL)	corr(AMZN,TSLA)	
AAPL	corr(AMZN,AAPL)	1	corr(TSLA,AAPL)	
TSLA	corr(AMZN,TSLA)	corr(TSLA,AAPL)	1	
⋮				

However, one time window is not that insightful. In order to represent the correlation evolving with time we must compute k of these matrices and thus have k windows on the data-set (e.g. W_1, W_2 and W_3). Windows are selected from the dataset as shown below:



More correlation matrices will enable us to visualize the evolution of correlation between stocks in desired periods (e.g. during a financial crisis, political instability or economic development).

Distance Matrix

The final graph however will not directly represent the notion of **correlation** but it's proportional inverse: **distance**. The notion of distance is straightforward when visualizing data on a graph. Data-processing in the distance-matrix step transforms the correlation matrix into a distance matrix using **distance co-variance**.

	AMZN	AAPL	TSLA	...
AMZN	0	dist(AMZN, AAPL)	dist(AMZN, TSLA)	
AAPL	dist(AMZN, AAPL)	0	dist(TSLA, AAPL)	
TSLA	dist(AMZN, TSLA)	dist(TSLA, AAPL)	0	
⋮				

Visualizing the graph

In order to easily visualize this graph, we compute the minimum spanning tree (MST). An MST always exists since the graph is fully-connected. The graph will thus be constructed by **nodes represented by stocks** and edges with **weights representing the distance co-variance between each pair of stocks**. Our web application enables users to navigate the graph and gain further statistics on a selected node.

Related Work

Most of the work done in this sector from which we took inspiration both on data processing and graph representation comes from:

- *A review of two decades of correlations, hierarchies, networks and clustering in financial markets* by Gautier Marti, Frank Nielsen, Mikołaj Bińkowski and Philippe Donnat.
- *Non-parametric sign prediction of high-dimensional correlation matrix coefficients* by Christian Bongiorno and Damien Challet