When dealing with multiple stocks

In finance, a covariance matrix can be a useful tool to estimate the cross correlations between different stocks. However, historical data contains random noise, which can alter the underlying information. Here we illustrate how Random Matrix Theory can be used to filter a diagonalisable matrix and how information about the market is carried inside its eigenvectors.

Correlation matrix construction

Let's consider N stocks with T+1 historical prices. First of all, we calculate the return of stock k=1,2,…,N as $R_k(t) = \ln(P_{t+1}^k/P_t^k)$, where P_t^k is the t-th price (with t=1,2,…,T+1) of stock k.

Then, we define the normalised return

$$r_k(t) = \frac{R_k(t) - \mu_k}{\sigma_k}$$

in which μ_k and σ_k are respectively the mean and the standard deviation of $\{R_k(t) | \, t = 1,2,\dots,T\}$.

Each $r_k(t)$ term can be rearranged into a TxN matrix A, whose elements are $a_{ij} \coloneqq r_j(i)$

$$A = \begin{pmatrix} r_1(1) & \cdots & r_N(1) \\ \vdots & \ddots & \vdots \\ r_1(T) & \cdots & r_N(T) \end{pmatrix}$$

So that, we can define the NxN correlation matrix C as

$$C = \frac{1}{T-1} A^T A$$

where A^T is the transpose of A, so that each element C_{ij} of C is

$$C_{ij} = \frac{1}{T-1} \sum_{t=1}^{T} r_i(t) r_j(t)$$

Since we used normalised returns, $-1 \leq C_{ij} \leq 1$ where $C_{ij} = 1$ means perfectly positive linear correlation between the i-th and the j-th stock. Moreover, C is symmetric and has ones on the main diagonal.

The following problem arises when constructing C from historical data:

- Using a finite number of historical prices T leads to the surge of random noise inside the correlation matrix;
- On the other hand, longer series of data could lead to the identification of non-stationary correlations between stocks, which were spurious and/or are no longer present in the market.

One solution is to filter the covariance matrix from noise.

Random Matrix Theory

Let's consider a TxN matrix B whose elements are $B_{ij} \sim iid(0,1)$ (that means, independent random variables with null mean and unitary variance). We can define a random NxN correlation matrix G as

$$G = \frac{1}{T-1} B^T B$$

We proceed with the diagonalization of G, finding its eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_N\}$ (ordered such that $\lambda_i \leq \lambda_{i+1}$) and eigenvectors $\{\bar{\alpha}_1, \bar{\alpha}_2, \dots, \bar{\alpha}_N\}$ (such that their norm is unitary $\bar{\alpha}_i{}^T \bar{\alpha}_i = 1$). Because G is a symmetric real-value matrix, it has N real eigenvalues.

Thus G can be represented as

$$G = (\bar{\alpha}_1 \quad \cdots \quad \bar{\alpha}_N) \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{pmatrix} (\bar{\alpha}_1 \quad \cdots \quad \bar{\alpha}_N)^T$$

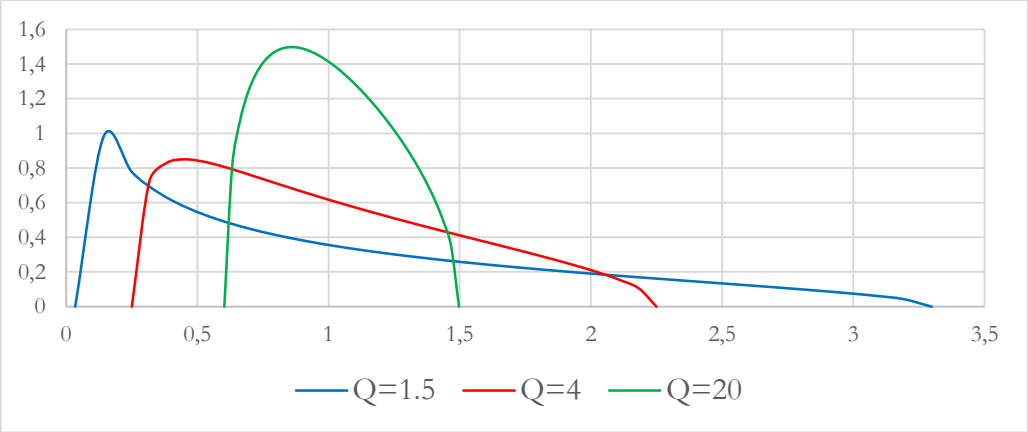
Properties of G are well defined in the random matrix theory, in particular in the limit $T \to +\infty,\ N \to +\infty$ such that $Q \coloneqq T/N > 1$ is constant, the probability distribution of the eigenvalues $\rho(\lambda)$ is

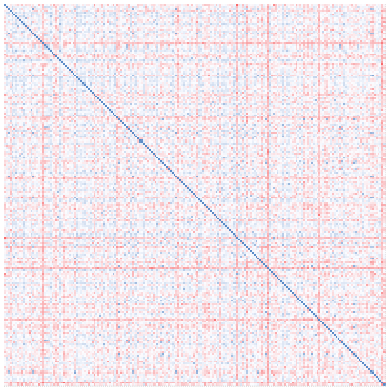
$$\rho(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda}$$

where

$$\lambda_{min}^{max} = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}$$

Here are some examples of $\rho(\lambda)$ with different Q.

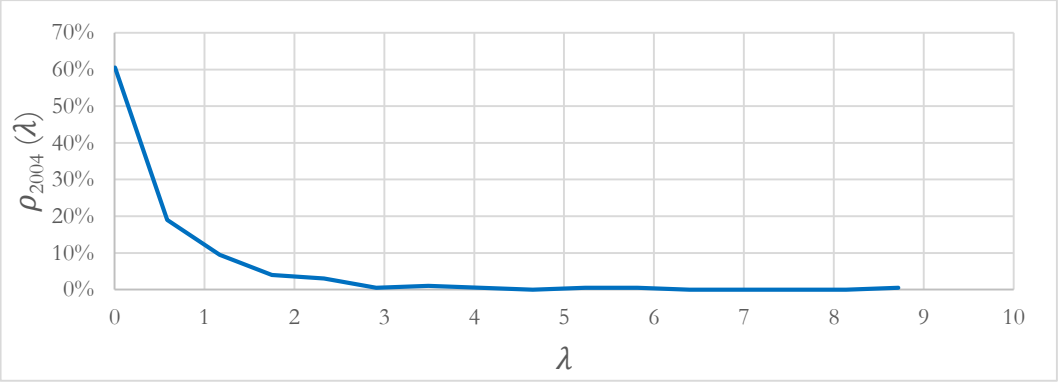




In order to analyse the amount of noise in our correlation matrix C, we diagonalize it and we look at the probability distribution of its eigenvalues. In this example, we took the 200 most capitalised stocks in the S&P 500 index and their normalised revenues in 2004 (N=200, T=250, Q=1.25). Its correlation matrix C_{2004} can be seen here, were blue dots corresponds to 1 and red dots corresponds to -1.

Filtering

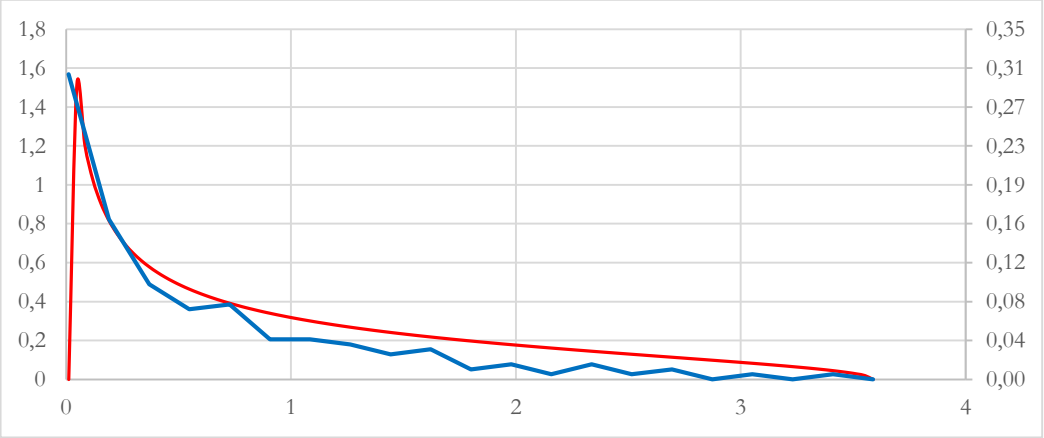
Its eigenvalues probability distribution $\rho_{2004}(\lambda)$ with $0 \leq \lambda \leq 10$ (all eigenvalues fall in this interval, except the greatest one which is $\lambda_{1000} = 50.6929$)



As you can see, it does not show much noise because it is taken from daily prices rather than intra-day quotations.

However, we can look closely. For our time series where $Q = 1.25, \lambda_{max} = 3.5889$ and $\lambda_{min} = 0.0111$, so we focus on the interval $\lambda_{min} \leq \lambda \leq \lambda_{max}$, plotted in the following graph.

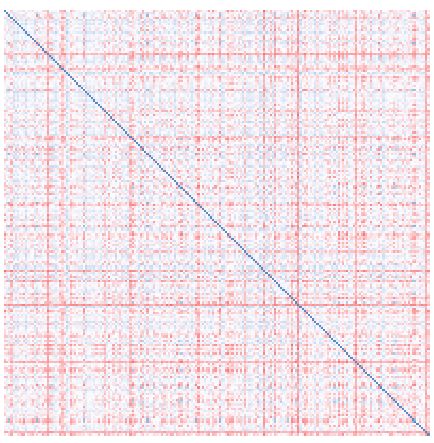
The red line shows the probability function of a random matrix (scaled on the left axis), while the blue line is the probability function ρ_{2004}, obtained from the historical data (scaled on the right axis).



Although they have different values, they share a common shape which suggest the presence of some random oscillations influence on the matrix.

The eigenvalues in the interval $[\lambda_{min}, \lambda_{max}]$ are the ones from λ_{50} to λ_{194}, so we substitute them with zeros, obtaining the set of filtered eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_{49}, 0, \dots, 0, \lambda_{195}, \dots, \lambda_{200}\}$. From this set we build the filtered NxN diagonal matrix Λ_{filt}

$$\Lambda_{filt} \coloneqq diag(\lambda_1, \lambda_2, \dots, \lambda_{49}, 0, \dots, 0, \lambda_{195}, \dots, \lambda_{200})$$



Using the eigenvector matrix $P_{2004} = (\bar{\alpha}_1, \bar{\alpha}_2, \dots, \bar{\alpha}_N)$, where $\bar{\alpha}_i$ is the eigenvector of C_{2004} corresponding to the i-th eigenvalue and with unitary norm, we can construct the filtered correlation matrix C'_{2004}

$$C'_{2004} \coloneqq P_{2004} \Lambda_{filt} P^T_{2004}$$

We impose also the elements of C'_{2004} on its main diagonal to be equal to 1. This is to preserve the fact that $Tr(C_{2004}) = Tr(C'_{2004}) = N$ (each stock must be perfectly positively correlated to itself).

On the left is the matrix obtained (blue means 1 and red means -1 as usual).

Portfolio construction

The difference between C_{2004} and C'_{2004} comes in hand (and can be tested) when we want to construct optimal portfolios of stocks. Here, we'll use Markowitz mean variance formulation. We call μ_i and σ_i with i=1, 2, … ,N the mean and the standard deviation of the revenues of the i-th stock calculated on the considered historical data (in our case, 2004).

They can be rearranged in the following Nx1 column vectors

$$\bar{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_N \end{pmatrix} \quad \bar{\sigma} = \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_N \end{pmatrix}$$

After that, we define the Nx1 column vector $\bar{w}$, whose elements w_i are weight of the i-th stock in the portfolio. Our portfolio will have expected mean return μ_P and variance σ_P^2

$$\mu_P = \sum_{i=1}^{N} w_i \mu_i \qquad \sigma_P^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i \sigma_i C_{ij} \sigma_j w_j$$

where C_{ij} are the elements of the covariance matrix. For a given μ_P, we want to minimize σ_P^2 (since the risk is quantified by the variance of the investment) under the constraint $\sum_{i=1}^{N} w_i = 1$. This task can be easily done and we will not spend time on it.

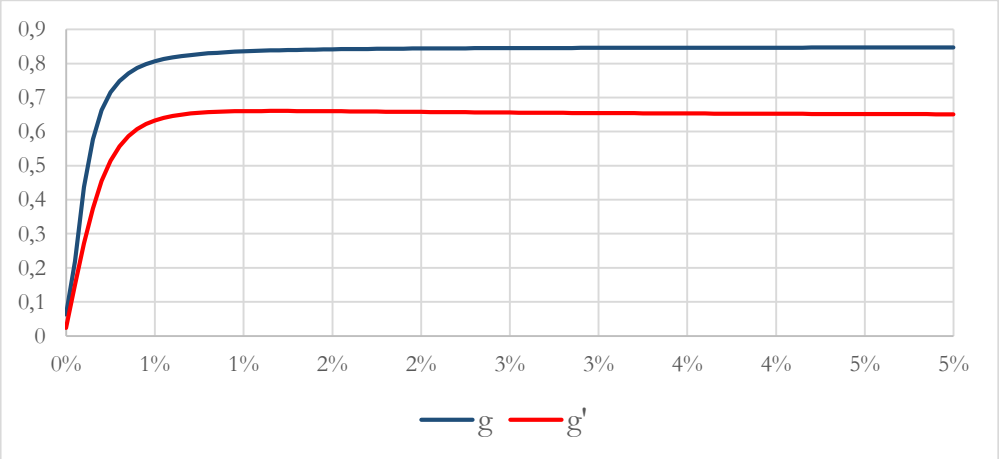
To test the effectiveness of the filtration, we calculate the $\bar{w}_{2004}(\mu^t)$ portfolio with target revenue μ^t using C_{2004} (in this case, with $0\% \leq \mu \leq 5\%$). Then, we use $\bar{w}_{2004}(\mu^t)$ to simulate the investment of those portfolios during 2005, sing historical data. We are able to calculate $\mu_{2005}(\mu^t)$, the actual mean revenue we would have earned if we invested the $\bar{w}_{2004}(\mu^t)$ portfolio in 2015. Moreover, it will be useful to define $\sigma_P(\mu^t)$, the expected volatility of portfolio $\bar{w}_{2004}(\mu^t)$.

In the same way, we calculate $\bar{w}'_{2004}(\mu^t)$, $\mu'_{2005}(\mu^t)$ and $\sigma'_P(\mu^t)$ using C'_{2004} instead of C_{2004}.

In order to quantify how much the estimates diverged from the expectation, we calculate the ratios

$$g(\mu^t) = \frac{|\mu_{2005}(\mu^t) - \mu^t|}{\sigma_P(\mu^t)} \qquad g'(\mu^t) = \frac{|\mu'_{2005}(\mu^t) - \mu^t|}{\sigma'_P(\mu^t)}$$

We can plot them in function of μ^t.

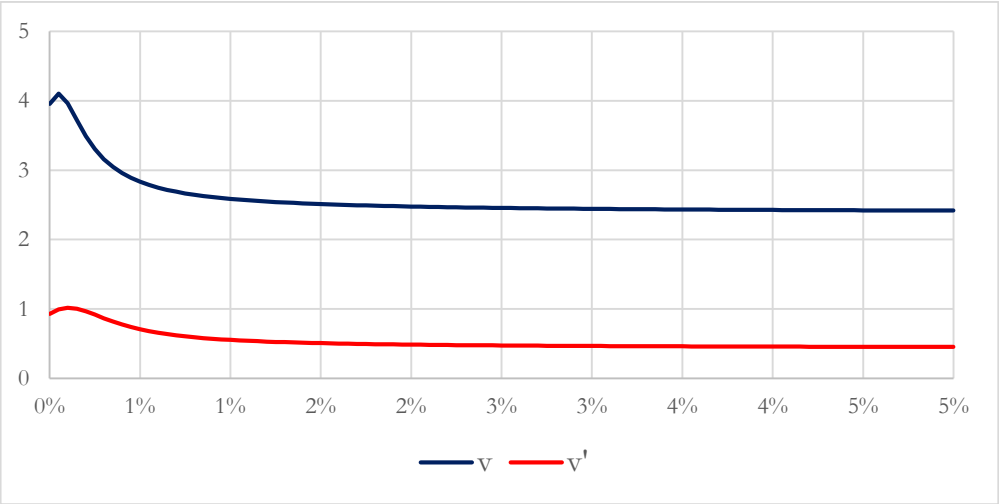


As you can see, the values obtained from the filtered covariance matrix are closer to the expectations than the ones obtained from the original data. Moreover, the difference between $g(\mu^t)$ and $g'(\mu^t)$ increases with the expected mean revenue.

We now want to know which one of the two covariance matrices better predicts future volatility. We define $\sigma_{2005}(\mu^t)$ and $\sigma'_{2005}(\mu^t)$ as the standard deviations of the returns obtained by investing $\bar{w}_{2004}(\mu^t)$ and $\bar{w}'_{2004}(\mu^t)$ in 2005. In order to measure the error in the estimation, we use

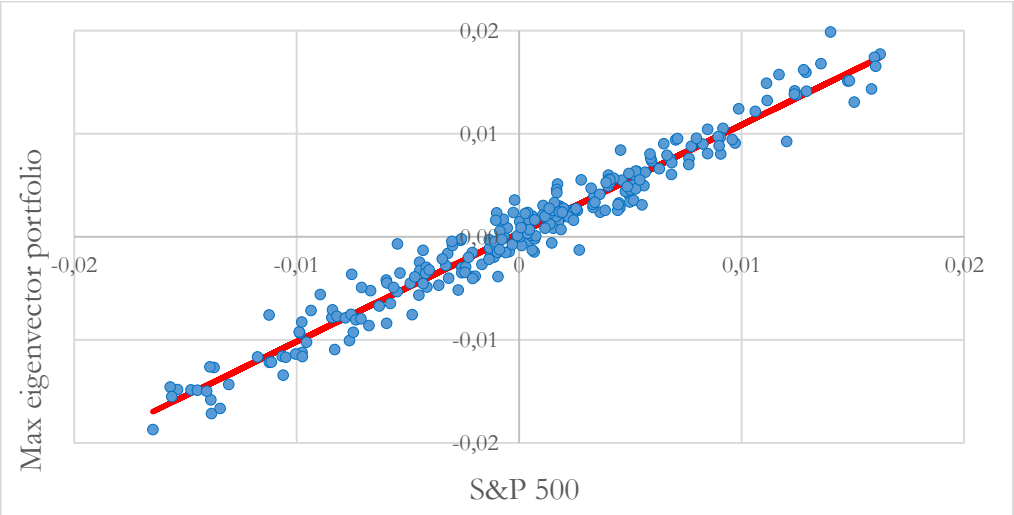
$$v(\mu^t) = \frac{\sigma_{2005}(\mu^t) - \sigma_P(\mu^t)}{\sigma_P(\mu^t)} \qquad v'(\mu^t) = \frac{\sigma'_{2005}(\mu^t) - \sigma'_P(\mu^t)}{\sigma'_P(\mu^t)}$$

which are the percentage displacement of the actual sigma from the predicted one. The filtered matrix C'_{2004} is better again.



Eigenvectors analysis

We can deduce more information from the decomposition of C_{2014} into its eigenvalues and eigenvectors. Between the eigenvalues which do not fall in the interval $[\lambda_{min}, \lambda_{max}]$ (where noise is), λ_{200} is interesting, because is at least one order bigger than the others. Its corresponding eigenvector $\bar{\alpha}_{200}$ witnesses influences which are common to all stocks, or the collective reaction of the market to external stimuli. We can use the components of $\bar{\alpha}_{200}$ as weights and compare the revenues of this portfolio with revenues of the S&P 500 in 2004 (we must not forget that the 200 stocks we are considering account for more than 70% of the S&P 500 total capitalization). In the following graph, the revenues of the $\bar{\alpha}_{200}$ portfolio are plotted against S&P 500. The values are close to the red regression line: we find a correlation coefficient of 0.978 between the two series.



What about the other eigenvectors? So as $\bar{\alpha}_{200}$ represents the market as a whole, the other biggest eigenvectors outside the noise interval characterise families of stocks. If we look at the biggest 10 components of each eigenvector, we can find that they are part of the same one or two groups of similar companies. This division comes from the fact that the profitability of

analogous business depends mainly on the same variables, so their market value behaves similarly. In particular, $\bar{\alpha}_{199}$ is composed by Information Technology companies, $\bar{\alpha}_{198}$ is composed by real estate investment trusts, $\bar{\alpha}_{197}$ is also composed by real estate, $\bar{\alpha}_{196}$ is composed by healthcare and insurance companies and, eventually, $\bar{\alpha}_{195}$ is composed by consumer discretionary and consumer staples companies. Here is the list of the companies.

$\bar{\alpha}_{199}$	
Applied Materials Inc	Information Technology
Intel Corp.	Information Technology
Texas Instruments	Information Technology
EMC Corp.	Information Technology
Cisco Systems	Information Technology
QUALCOMM Inc.	Information Technology
Dollar Tree	Consumer Discretionary
Yahoo Inc.	Information Technology
Automatic Data Processing	Information Technology
Amazon.com Inc	Consumer Discretionary
Paychex Inc.	Information Technology
Stryker Corp.	Health Care
Microsoft Corp.	Information Technology
Adobe Systems Inc	Information Technology
International Bus. Machines	Information Technology

$\bar{\alpha}_{198}$	
AvalonBay Communities, Inc.	Real estate
Simon Property Group Inc	Real estate
Public Storage	Real estate
Equity Residential	Real estate
General Growth Properties Inc.	Real estate
Prologis	Real estate
Welltower Inc.	Real estate
Wal-Mart Stores	Consumer Staples
American Electric Power	Utilities
Home Depot	Consumer Discretionary

$\bar{\alpha}_{197}$	
Welltower Inc.	Real estate
Prologis	Real estate
Texas Instruments	Information Technology
Public Storage	Real estate
General Growth Properties Inc.	Real estate
AvalonBay Communities, Inc.	Real estate
Simon Property Group Inc	Real estate
Equity Residential	Real estate
Corning Inc.	Industrials
Cisco Systems	Information Technology

$\bar{\alpha}_{196}$	
Aetna Inc	Health Care
Marsh & McLennan	Insurance
Aon plc	Insurance
Chubb Limited	Insurance
Anthem Inc.	Health Care
United Health Group Inc.	Health Care
CIGNA Corp.	Health Care
Prudential Financial	Insurance
MetLife Inc.	Insurance
Humana Inc.	Health Care

$\bar{\alpha}_{195}$	
Target Corp.	Consumer Discretionary
AutoZone Inc	Consumer Discretionary
TJX Companies Inc.	Consumer Discretionary
Costco Co.	Consumer Staples
Ross Stores	Consumer Discretionary
Nike	Consumer Discretionary
Aon plc	Insurance
Wal-Mart Stores	Consumer Staples
The Travelers Companies Inc.	Insurance
Lowe's Cos.	Consumer Discretionary
V.F. Corp.	Consumer Discretionary
Chubb Limited	Insurance
L Brands Inc.	Consumer Discretionary
Dollar Tree	Consumer Discretionary
Valero Energy	Energy

In conclusion, we found that the random matrix theory is a useful tool to dampen the noise inside a correlation matrix. Because portfolios built with the filtered matrix showed improvements in their prediction of mean return and risk, the filtered matrix seems to be more reliable than the raw one. Moreover, eigenvalues and eigenvectors from the correlation matrix can be used to deduce which portfolio follows the market reaction to external stimuli and how the market divides itself in groups of similar businesses.