Ukrainian Catholic University

Faculty of Applied Sciences

Business Analytics Programme

Portfolio theory: curse of dimensionality, Monte-Carlo and eigen-vesting

Linear Algebra final project report

Authors:

Yevheniia Mitriakhina

Yarka Lyba

14 May 2019

Contents

Abstract

Modern portfolio theory is a concept from finance to describe ways of diversifying and allocating assets in a financial portfolio to maximize the portfolio's expected return. Using modern portfolio theory an investor bundles different types of investments together, so that when some of the securities fall in value another rise. So, in that case, the return of each asset is less important than the behavior of the entire portfolio. American economist Harry Markowitz first introduced Modern portfolio theory in 1952 what led to him being awarded the Nobel Prize in Economics in 1990. Since that time many researchers have developed theories on portfolio optimization and risk management.

1 Introduction

Financial markets are a great field for testing statistical ideas, because forces that drive them are unpredictable and to large extent random. The number of participants is large, and we have access to historical data from Yahoo which allow us to estimate empirical correlation matrix. Our aim is to compose an optimal portfolio from stocks included into SP 500 index. We discuss various methods of cleaning large correlation matrices, implement some of them in practice and compare in- and out-of-sample return and volatility of resulting portfolios. As the result of the work we want to understand what are the best approaches and if they are applicable for the real world at all.

2 Review of the literature

Portfolio theory was founded by Harry Markowitz in 50's [1]. The earliest groundwork in the Random Matrix Theory comes not from finance, but from physics and signal processing. We studied more recent works focusing mainly on financial correlation matrices, namely [3], [4], [5], [6], [7]. Our code can be found here [8]

3 Problem Setting

As an input we get daily close prices for assets of many companies, for example, Google, Facebook and Tesla. Each of the assets has its own dynamics of growth or descent. An investor has some amount of money that he will spend on buying assets. Task is to compose the portfolio (set of assets), that will maximize the return given some level of risk tolerance. Portfolio is composed by setting weights on each type of assets. For example, if the weight for Apple is 0.4 then 40 percent of our money we spend on Apple assets. We will use different approaches for finding optimal weights.

3.1 Notions

R - matrix of historical normalized returns.
C - true correlation matrix.
$\hat{C}$ - estimator of the correlation matrix
E - empirical correlation matrix.
r - vector of expected returns.
w - vector of weights assigned to each asset of the optimal portfolio.

3.2 Problem formulation

Let form the portfolio for n assets, let M be the amount of money we want to invest. For each asset $i = 1, 2, 3, ..., n$ there is some weight w_i, so that $w_i \cdot m_i$ is amount of money we invest into asset i. We can not invest more money then we have (short selling and leverage not allowed), so $\sum_{i=1}^{n} w_i = 1$. Let r_i be the return for asset i, then $\sum_{i=1}^{n} w_i \cdot r_i$ is return of the portfolio. Find such values for w_i, that make biggest return at given level of risk.

4 Examining the data

We use trade data from Yahoo Finance. To get that data we use R package "BatchGetSymbols". It gives the opportunity to download and organize financial data for multiple tickers. It gives a lot of values, but for the project we will use only close prices - the final prices at which a securities were traded on a given trading day.

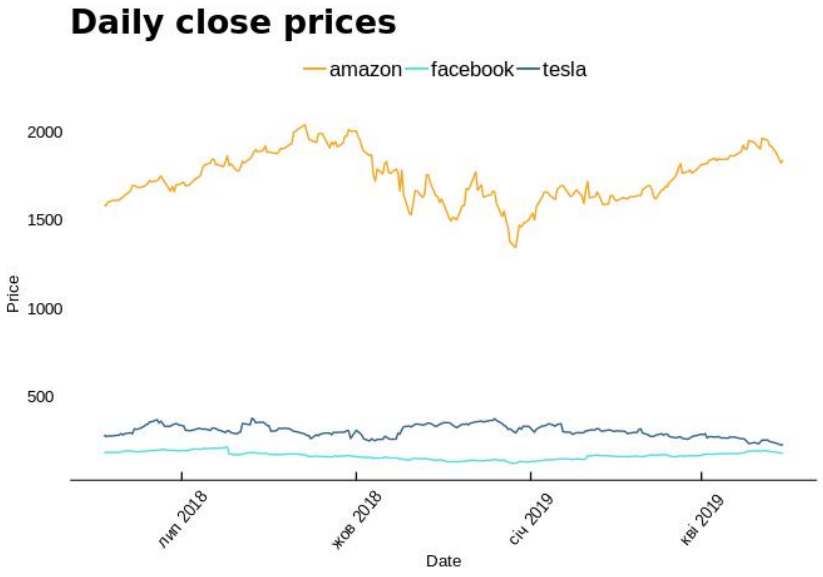


Figure 1: Daily close prices.

To track the dynamics of change we calculate daily returns: $p_i(t) = \frac{Price_i(t) - Price_i(t-1)}{Price_i(t-1)}$

	fb	tsla	amzn
2	-0.0037401	-0.0333227	-0.0025607
3	0.0168661	0.0147631	0.0129379
4	-0.0051899	-0.0043717	0.0007553
5	-0.0054321	0.0035991	0.0044166
6	0.0044344	0.0176080	0.0016893
7	0.0103908	0.0280518	0.0074526
8	0.0219001	-0.0239613	0.0029110
9	0.0115237	0.0249008	0.0073146

Figure 2: Daily returns of the companies

5 Mean-variance portfolio

5.1 Basics on mean-variance portfolio theory

The idea of the mean-variance portfolio theory is to choose weights for assets that will compose portfolio with minimal volatility. Volatility is a statistical measure of the dispersion of returns for a given security or market index. In most cases, the higher the volatility, the riskier the security.
Let R be the matrix of historical normalized returns of i assets and r_i - rate of return for the i_{th} asset.
Let E be the empirical correlation matrix constructed out of R and w - vector of weights.
To form the rate of return of the portfolio we sum all returns multiplied by weights.

$$return = \sum_{i=1}^{n} w_i \cdot r_i$$

It leads to the most important point for the further calculations. $return$ is a random variable that has variance $w^T E w$. Variance is the risk of the portfolio, so it has to be minimized with respect to the constraint that sum of the weights should be equal to zero.

5.2 Finding the solution

To find the weights we have to solve the quadratic program. Lets define the function:

$$F(w, l) = w^T E w + l(w^T e - 1),$$

where e is the vector of ones. The first one part of the function represents the risk that has to be minimized, the second one - the constraint about the sum of weights equal to 1. Covariance matrix is symmetric. We may assume that E is positive definite as the only reason for semi-definiteness is perfect correlation of two variable what is impossible for finance data.
The minimum occurs at (w, l), where

$$\nabla F(w, l) = \mathbf{0}^T.$$

Taking the derivatives we get two equations:

$$w^T \cdot \mathbf{1} = 1$$

and

$$E \cdot w = -l \cdot \mathbf{1}$$

.
We take the inverse of E and get the solution:

$$w = -l \cdot E^{-1}1$$

To eliminate l we use the first equation:

$$w = \frac{-l \cdot E^{-1}\mathbf{1}}{1} = \frac{-l \cdot E^{-1}1}{w^T\mathbf{1}} = \frac{-l \cdot E^{-1}\mathbf{1}}{(-l \cdot E^{-1}1)^T\mathbf{1}} = \frac{E^{-1}\mathbf{1}}{(E^{-1}1)^T\mathbf{1}}$$

We rely on R to do quadratic optimization for us, so using library [2] one has to set covariance matrix, vector of returns, and the matrix of constraints.

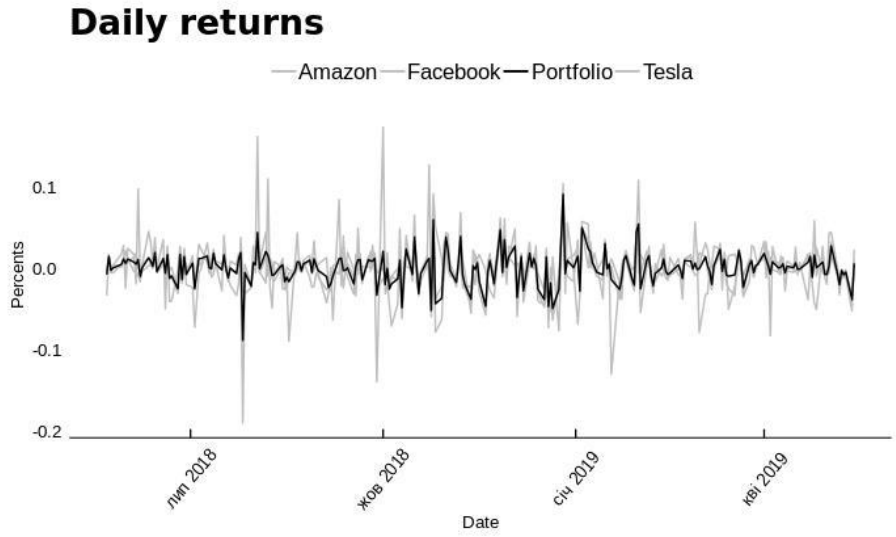


Figure 3: Returns of seperate assets vs. returns of the portfolio

The black one line represents portfolio, grey lines are separate assets.

6 Covariances in portfolio theory

In order to determine the optimal portfolio, according to Markowitz approach, one has to invert the correlation matrix. Since this has, as a rule, a number of small eigenvalues, any measurement error will get amplified and the resulting portfolio will be putting much weight on eigenvalues which are likely coming from random noise. We attempt to characterize the difference between E and C, and discuss how well (or how badly) one may reconstruct C from the knowledge of E in the case where N and T become very large but of the same order with their ratio q = N/T not vanishingly small; this is called the large dimension limit, or else the "Kolmogorov regime". We have data available for N = 447 companies for about T = 2000 trading days, so the covariance matrix is of size 447*447. We consider this matrix to be large enough for random effects to be present. Note, that as we normalized returns as

$$r_{i,t}^{`} = \frac{r_{i,t} - \bar{r_i}}{\sigma_i}$$

terms correlation matrix and covariance matrix from now on can be used interchangeably.

Estimation of covariance matrices

The empirical correlation matrix is obtained as simply Pearson coefficients

$$E_{i,j} = \frac{1}{T} R R^*_{i,j}$$

We examine the distribution of eigenvalues of the empirical correlation matrix:

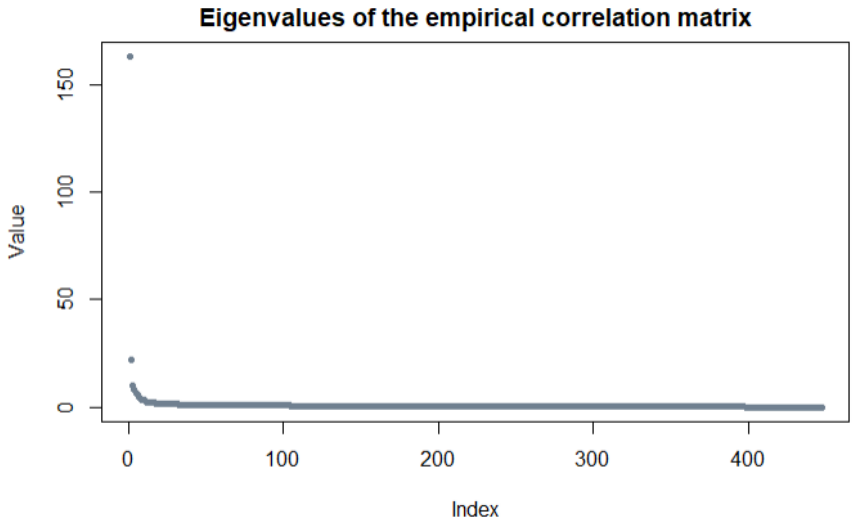


Figure 4: Distribution of the eigenvalues of empirical correlation matrix

One can easily see the largest eigenvalue is much larger than all the other eigenvalues. This implies that most stocks follow a dominant correlation mode: there is a tendency of a positive mutual

correlation of all stocks in the market. From the financial point of view, it is intuitive to understand because all companies operate in the same economy and therefore, are influenced by many common macro factors. There are a few eigenvalues that are larger than mean, but less than the market mode. Following the same intuition we can assume that they describe 'sector modes' - mutual correlations between assets from one sector of economy.

6.1 Eigenvalues clipping

There are also a lot of eigenvalues clustered near 0, which we consider to be coming from noise. To obtain reliable estimator of $\boldsymbol{C}$ one must set a threshold value λ_t, with eigenvalues lying above it considered informative, and those lying below - random. Then smallest eigenvalues are replaced by constant γ such that

$$Tr(\hat{C}) = Tr(E)$$

.

$$\hat{\lambda_k} = \begin{cases} \lambda_k & \text{if} \lambda_k > \lambda_t, \\ \gamma & \text{otherwise} \end{cases}$$

This method is known as ***eigenvalues clipping*** [7].

The density function for eigenvalues is given by Marchenko-Pastur distribution at the limit $N \to \infty$ and $T \to \infty$ and having $Q = N/T$ fixed and greater than 1

$$p_c(\lambda) = \frac{1}{2Q\pi\sigma^2} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda}$$

where λ_{max} and λ_{min} are given by $\lambda max, min = \sigma^2(1+Q \pm 2\sqrt{Q})$. Nextly, we fit Marchenko-Pastur pdf to the data to make sure if the theoretical distribution satisfyingly describes the smallest eigenvalues.

Predicted $\lambda_{max} = 1.42$ is almost 100 times smaller than actual λ_{max}, so our hypothesis of purely random nature of the data is indeed inconsistent and it also can be seen from the plot above(black line is not the best fit). Due to that fact and sampling error and we can adjust parameters Q and σ so that they better describe small eigenvalues. The best fit we could get is represented by the red line. λ_{max} corresponding to the adjusted parameters equals 0.348 so we dismiss them and replace with constant. Estimator $\hat{C}$ of the true correlation matrix C is constructed as

$$\hat{C} = \sum_{i=1}^{N} \hat{\lambda}_i u_i u_i^*$$

where u_i is the eigenvector of E corresponding to eigenvalue λ_i.

6.2 Rotationally invariant estimator

Eigenvalues clipping deals with the smallest eigenvalues being underestimated, but [4] suggests that exists systematic bias about the largest eigenvalues. To deal with both overestimation of the largest

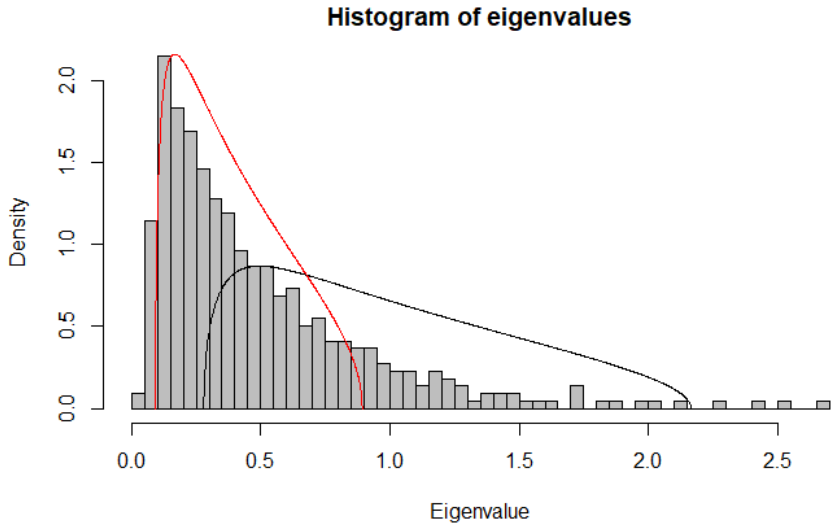


Figure 5: Density of eigenvalues. Black line shows Marchenko-Pastur pdf with parameters Q = 0.224 and $\sigma^2 = 1$. Red line represents the best fit with Q = 0.263 and $\sigma^2 = 0.39$

eigenvalues and underestimation of the smallest ones we use ***rotationally invariant*** estimator based on the optimal shrinkage function of the form

$$\lambda_k^{\hat{RIE}} = \frac{\lambda_k}{|1 - q + qz_k s(z_k)|^2}$$

where z_k is the complex value $z_k = \lambda_k - i\nu$ and

$$s(z_k) = \frac{1}{N} \sum_{\substack{j=1 \\ j \neq k}}^{N} \frac{1}{z_k - \lambda_j}$$

as it was introduced in [7]. The rotational invariant hypothesis on an estimator assumes that one does not have any prior knowledge of the structure of the true eigenvectors, and therefore that the best one can do is to keep the eigenvectors u_i of E untouched.
We set parameter $\nu = \sqrt{N}$. For not so large N such choice of ν creates downward bias which can be calculated explicitly as

$$\Gamma_k = \sigma^2 \frac{|1 - q + qz_k g_{mp}(z_k)|^2}{\lambda_k}$$

where $g_{mp}(z)$ is the Stieltjes transform of Marcenko-Pastur density given by

$$g_{mp}(z) = \frac{z + \sigma^2(q-1) - \sqrt{z - \lambda_n}\sqrt{z - \lambda_+}}{2qz\sigma^2}$$

where λ_n is the smallest eigenvalue and

$$\lambda_+ = \lambda_n \Big(\frac{1 + \sqrt{q}}{1 - \sqrt{q}} \Big)^2$$

and σ^2 is estimated standard deviation defined as

$$\sigma^2 = \frac{\lambda_n}{(1-\sqrt{q})^2}$$

Finally, RI estimator is

$$\hat{\lambda}_k^{RIE} = max(1, \Gamma_k)\hat{\lambda}_k^{RIE}$$

This method is implemented in separate file which can be run parallel to the main ($random\_matrix\_and\_financial$ after the data preparation part.

6.3 Comparison of the resulting portfolios

After finding all three covariance matrices with quadratic optimization we calculate in-sample returns as

$$R = \sum_{i=1}^{N} R_i w_i$$

We also calculate out-of-sample return for the test period of 40 days in the same way to compare performances of portfolios.

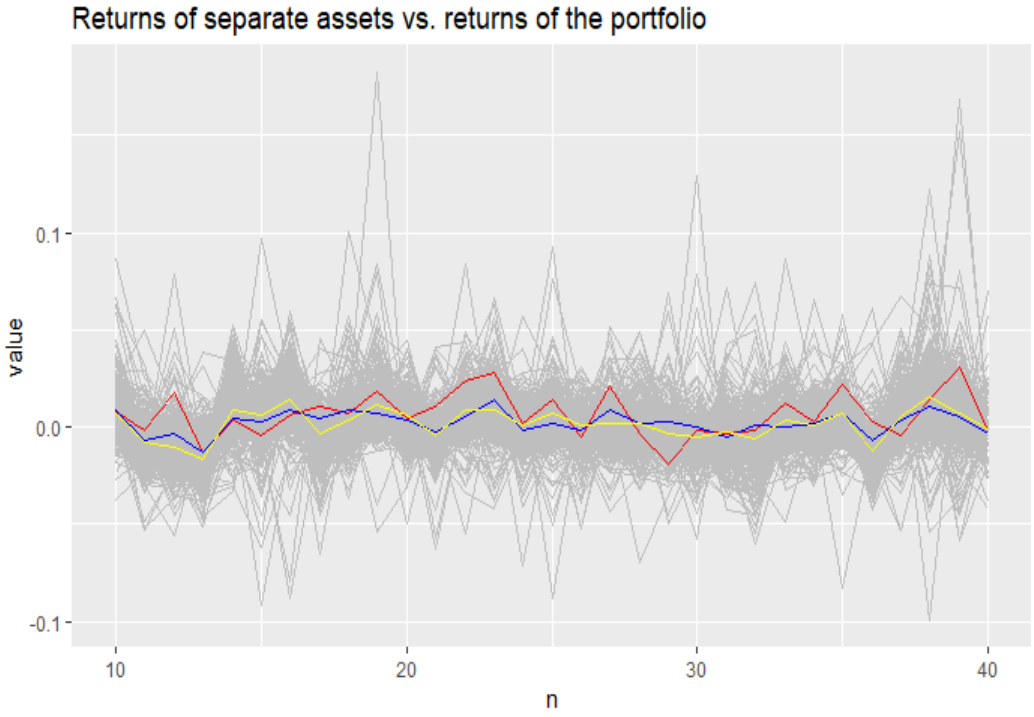


Figure 6: 30 days close-up to returns of separate assets(grey) and portfolios. Red line represents in-sample portfolio returns after eigenvalues clipping. Blue - composed with empirical correlation matrix. Yellow - after RI estimation.

One can observe that total variance of all three portfolios is less than those of separate assets as it was expected. It as also clear, that portfolios put weight on drastically different assets, because total returns tend to be directed in different directions.

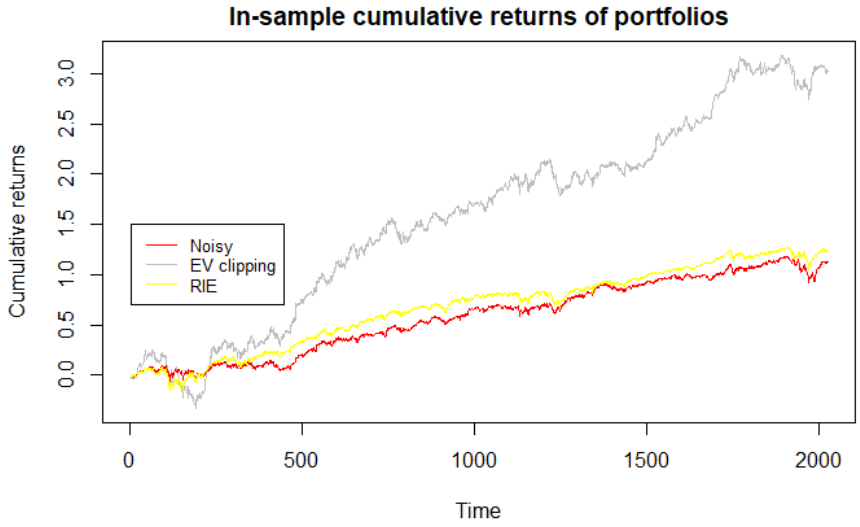


Figure 7: Portfolio obtained from EV clipped correlation matrix significantly outperforms both our cleaned and RI-estimated ones

The in-sample result is not of great interest for us, but from the plot we can see that EV-clipped portfolio is more profitable than both based on empirical correlations and on rotationally invariant estimator.

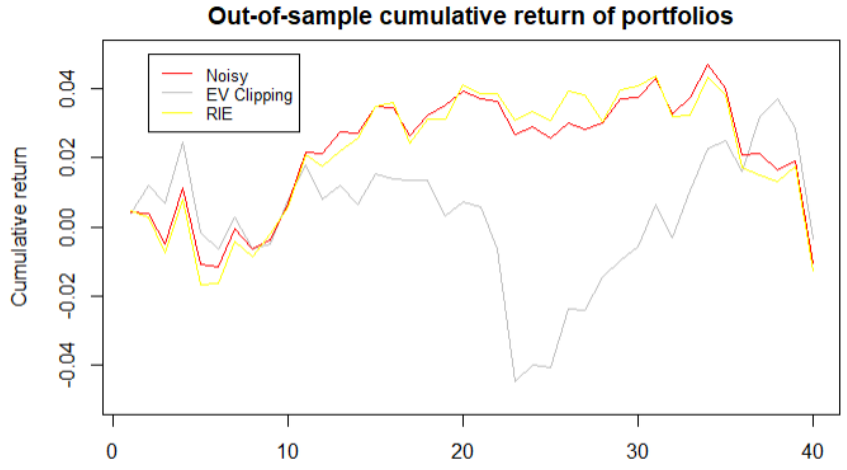


Figure 8: Out-of-sample performance

The out-of-sample performance of EV-clipped portfolio on the test period of the last 40 trading days is the most volatile with $\sigma_{EV} = 0.0124$. The unfiltered and RI-estimated are almost tied up with $\sigma_{noisy} = 0.0092$ and $\sigma_{RIE} = 0.0099$. EV-clipping did not prove to be reliable way of estimation

correlation matrix, as it has shown the worst performance on the test data, while being the best on the training sample.

6.4 Can we trust the correlation matrix and how to break the curse of dimensionality?

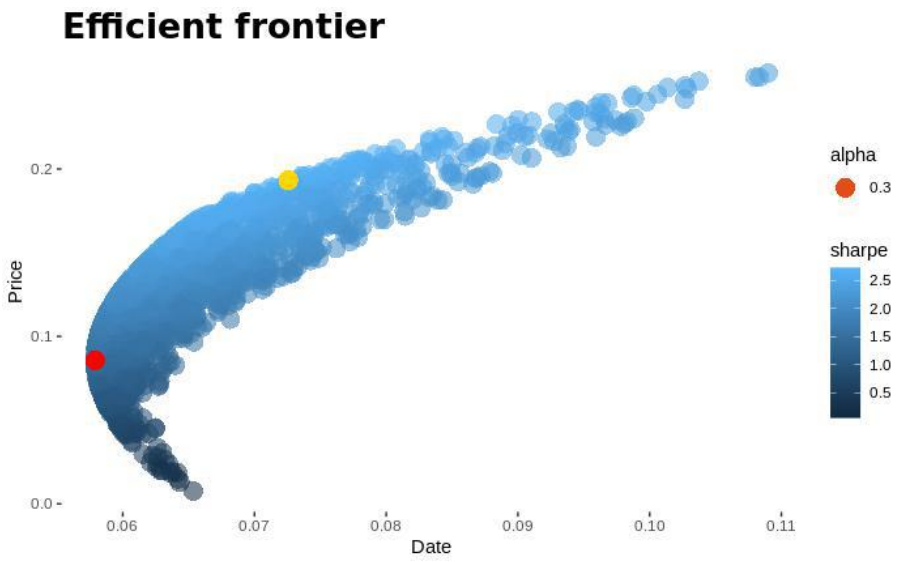
We can see that sometimes noisy portfolio beats our cleaned one. Of course, the covariance matrix is time dependent regardless of the filtering, and the filtering changes overtime. It's possible for the unfiltered portfolio to beat the filtered portfolio on occasion, but if it does, it's probably due to random chance. In summary, we have seen that the smallest eigenvalues of the correlation matrix, on which the Markowitz solution puts the most weight, are absolutely dominated by the measurement noise. Consequently, to choose the most efficient portfolio one needs a better estimator $\tilde{C}$, than empirical correlation matrix. $\tilde{C}$ might account for the underestimation of small eigenvalues and overestimation of large eigenvalues. Eigenvalues clipping approach which we used did not show steady better performance than simple unfiltered portfolio. Probably, this cleaning overlooks the fact that the large empirical eigenvalues are overestimated. However, RIE, which is designed to correct this two-sided systematic bias is not showing significant improvements over considering just the empirical correlation matrix with $q << 1$. For the future research we propose pre-filtering of the data(i.e. shrinking the number of assets to consider) and avoiding the dimensionality trap.

7 Monte-Carlo simulation

Monte Carlo Simulation is a mathematical technique that generates random variables for modeling risk or uncertainty of a certain system. Monte Carlo simulation is often used in finance, engineering, and physics. Monte Carlo simulation produces distributions of possible outcome values. This approach is much more easier and does not require any calculations. In our case, we use uniform distribution. All values have equal chances to occur. We define the minimum and maximum as 0 and 1.

We have generated 2000 random vectors of weights and formed different portfolios. They all where stored into the data frame with two additional values: Sharpe ration and volatility.

We plot all the portfolios and construct the efficient frontier. The efficient frontier is the set of optimal portfolios that offer the highest expected return for a defined level of risk or the lowest risk for a given level of expected return. Portfolios that lie below the efficient frontier are sub-optimal because they do not provide enough return for the level of risk. We select two portfolios - with smallest volatility(red) and biggest Sharpe ration(yellow). Portfolio is chosen depending on the fact, how much risk-averse the investor is.



8 Summary

We reviewed the basics of portfolio theory, starting from Markowitz's optimal portfolio with constrained quadratic optimization to ways to improve the performance of this approach by filtering the correlation matrix. The results obtained with random matrix theory methods are not quite robust in our case: we did not reach significant drop in out-of-sample volatility. All of the approaches have a significant drawback - we are relying on the historical data which implies the strong dependence of risk estimators on the period chosen, whilst the markets are affected by random factors and we cannot presume stationarity of returns. All in all, the financial time series are the least stable, so the martingale hypothesis does not hold mostly. To put it briefly, portfolio theory is not a finished field, and there is still a lot of work to do.

References

[1] Harry Markowitz, *Portfolio selection*

[2] *https://cran.r-project.org/web/packages/quadprog/quadprog.pdf*

[3] Jean-Philippe Bouchaud, Marc Potters, *Theory of financial risk and derivative pricing: from statistical physics to risk management*

[4] Jean-Philippe Bouchaud, Marc Potters, Laurent Laloux *Financial Applications of Random Matrix Theory: Old Laces and New Pieces*

[5] Joel Bun, Jean-Philippe Bouchaud, Marc Potters *Cleaning large Correlation Matrices: tools from Random Matrix Theory*

[6] Laurent Laloux, Pierre Cizeau and Marc Potters *RANDOM MATRIX THEORY AND FINANCIAL CORRELATIONS*

[7] Joel Bun, Jean-Philippe Bouchaud, Marc Potters *Cleaning correlation matrices: A new cleaning recipe that outperforms all existing estimators in terms of the out-of-sample risk of synthetic portfolios*

[8] `https://github.com/mitryahina/random-matrix-theory-in-estimating-financial-correlations`