Big Data - Portfolio Construction

GAUDIN Benjamin, MAILLARD Vincent

November 10, 2015

Abstract

We construct a mean-variance portfolio for a HNW client. Our universe is composed of swiss and U.S. stocks, as well as bitcoins and ETFs. From this large set of data, we pick between 40 and 100 instruments, chosen so that the portfolio is well diversified. In details, we build a correlation network and take stocks in each cluster. We then allocate the wealth to the instruments. Finally, we backtest our strategy.

1 Introduction

The more diversified a portfolio is, the less risk it bears. However, it is of practical impossibility to buy all the instruments in the world, because the amount of money an individual possesses is limited. Therefore, we must implement strategies that make us well diversified by choosing adequately a restricted number of stocks. For this purpose, we use a complex network which clusters correlated instruments.

A common approach to the problem of allocating wealth to financial instruments is to use the mean-variance optimization framework developed by Markowitz. The idea is to minimize the risk (the variance) given a target return. Named *modern portfolio theory*, it requires the portfolio builder to input the covariance matrix of the asset returns.

The problem we face is that real world data have outliers and missing values. It is then challenging to compute the covariance matrix. Moreover, a large set of data induces an error-maximisation effect. To prevent this, we use techniques such as matrix shrinkage and eigenvalue clipping, a technique coming from random matrix theory.

After picking our stocks and choosing their weight, we conduct a statistical exploration of data based on stylized facts, as well as the backtesting of our strategy. We find that the eigenvalue clipping method gives a higher Sharpe ratio than the matrix shrinkage. Both give a larger Sharpe ratio than the sample covariance matrix.

2 Data handling

The different sets of data that are used are extensive and have to be handled carefully. Indeed some values may be missing, or even worse, there may be some that are completely wrong inputs.

To handle missing values when it is needed, a linear interpolation of prices is used. This allows to compute returns without the missing values, which can prove to be useful when computing the portfolio worth. Moreover, the function that computes the portfolio weights requires that the returns are fully determined over the whole period. In cases where it not the case, an extrapolation of returns is used. This is done by simulating returns with the same mean and volatility with a student-t distribution.

To handle wrong inputs, the approach that is chosen is to winsorize the top and bottom half percent of returns. The advantage of this method is that it efficiently avoids possibly non-valid extreme values, and changes them to the top/bottom 0.5% quantile. The disadvantage is that it gets rid of one percent of the most extreme values whatever happens. Another option would be to put a simple threshold on returns. The down-side is that the determination of its value is completely arbitrary. For this reason, winsorizing is chosen.

Finally some products can have very scarce data. It would not be astute to include the latter instruments in our portfolio as their dynamics is difficult to assess and consequently unreliable. Therefore, we choose to ignore stocks for which less than 30% of the data is available.

While loading the data into our framework, the S&P500 is removed from the US stocks to avoid problem with the correlation network. Some data are also removed from the Bitcoin series due to the fact that multiple transaction took place during the same second.

Unfortunately, we cannot include bitcoins and ETFs in our portfolio. We don't include them because their time series begin when data from stocks end, so we have a problem to compute the correlation and to backtest a portfolio strategy that would include them. There is a second reason concerning bitcoins that makes us reluctant to take them into consideration. Indeed, currency investing is a zero-sum game but with volatility. As risk-averse agents, this is something we want to avoid. Investing in bitcoins is of speculative interest, and this is not what our client expects from us – although it can be a source a diversification, since is it not correlated with other products.

We conduct an analysis of the stocks based on stylized facts below, once we have chosen them.

3 Methodology

3.1 Correlation network

The correlation network provides a way of considering only a small fraction of the initial data, while still being diversified among the different risks. This is useful for a HNW individual who may not invest in too many assets. The correlation network offers a nice visual support that can be easily understood by non-financially literate people, and it can be used to explain the strategy to our client (see Figure 1 and 2).

To compute these correlation networks, the whole cleaned data sample is used. Though one may argue this is not consistent with the backtesting that is done later on, we decide that, since correlation is constant through time, it is worth using all the data at hand.

We have at our disposal two data sets: time series of swiss stocks and U.S. stocks. The first step of our method is to aggregate instruments in each sets according to their correlation. The correlation tree is built following the methodology given in [4]. The distance between two assets is given by

$$d(i,j) = \sqrt{2(1-\rho_{ij})}$$

The distance matrix is then easy to work with using the *igraphic* package which includes algorithms to spot community clusters. These clusters contain instruments that are highly correlated with each other, which means that their returns tend to move together. In our tree, a short distance between two instruments represent a high correlation. In each cluster, we pick the instrument that has the shortest distance to all the other members of the cluster.

A difficulty we encounter is that when the set of data is very large, the network gets extremely messy. Because of that, we cannot merge all the data together. We choose to build a network for each set of data. We can argue that this is a relevant decision for handling exposure to currency risk (U.S. stocks trade in dollars, swiss stocks trade in CHF). By knowing how much exposure to a currency a portfolio has, one can better assess its risk.

This results in the choice of 54 instruments: 13 in swiss stocks and 41 in U.S. stocks.

3.2 Stylized Facts

We conduct an analysis of the chosen stocks by empirically determining whether they have the patterns of stylized facts that are typical to stock dynamics. In particular, we look at fat tails of returns, asymmetry of returns, aggregate Gaussianity, leverage effect, autocorrelation of returns and autocorrelation of absolute return. The correlation between daily returns and volatility is -0.48%, so it is negative. In Figure 3, we can see that kurtosis and skewness are important in daily returns, but fade away when the time span increases to monthly returns. In Figure 4, we compute the p-value testing the hypothesis that the autocorrelation between two

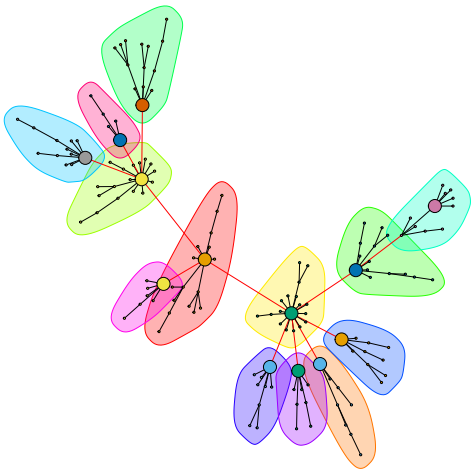

Figure 1: Correlation network swiss stocks

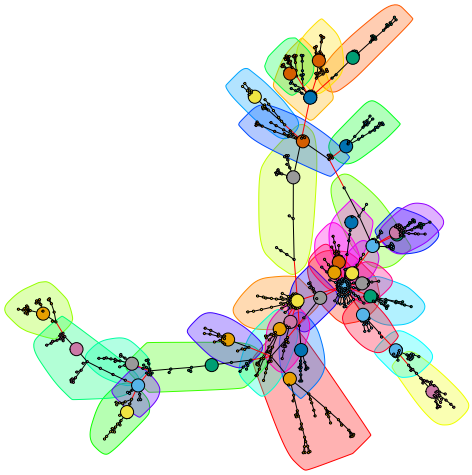

Figure 2: Correlation network U.S stocks

days of return (absolute return) is absent. We see that the hypothesis is rejected for absolute return, and that it is not rejected for returns for most stocks. This probably comes simply from randomness. We find that the data follows well the usual stylized facts. We would still need to show to ACF for returns to be convinced of that, along with volatility clustering and slow decay of autocorrelation of absolute returns. We don't do that here, by lack of place, but we show these features for our portfolio below.

3.3 Weights

At this stage, we have our instruments. The third step is about correctly allocating a fraction of wealth to each of the chosen instruments.

We use the mean-variance framework to allocate the wealth. We argue that on average, agents act like mean-variance optimizers (minimizing risk and maximizing returns), so this is the optimal behavior that we want to replicate for our client.

Singularity of the sample correlation matrix

Handling a large set of data makes one face the curse of dimensionality. The sample correlation matrix may be close to singularity (i.e. have eigenvalues close to zero). Because we use the inverse of the correlation matrix in the choice of the instrument weights, this will lead to over-allocation and error-maximization. To prevent this, we examine two techniques.

The first one is the sample correlation matrix shrinkage. We consider a highly

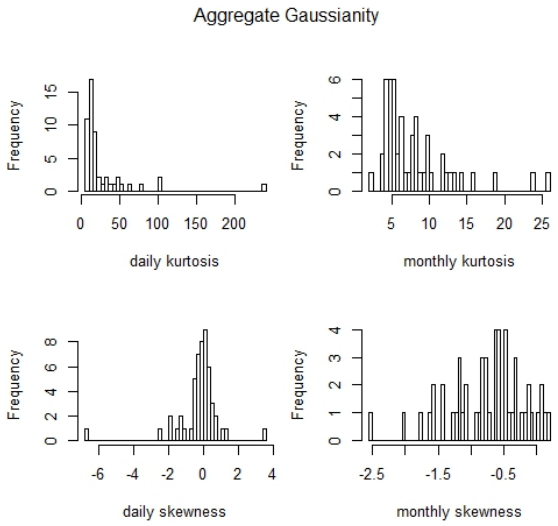

Figure 3: Kurtosis and skewness are distributed closer to zero as time span increases.

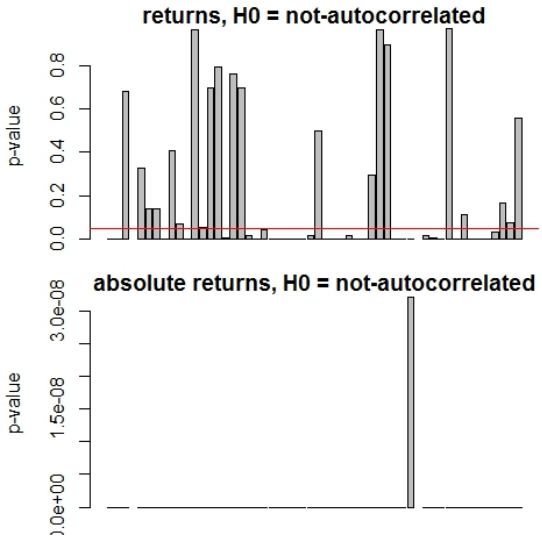

Figure 4: Absolute returns are correlated, not returns. Lag 1 day.

structured correlation matrix estimator with constant coefficients of correlation, which constitutes our target. There exists an optimal linear combination of the sample matrix and the target matrix. This has the advantage of lowering the mean squared error (it lowers the variance at the expense of introducing a bias) and removing the singular characteristic of the sample matrix.

The second one is eigenvalue clipping described in [1]. For large random matrices, it is known from random matrix theory that the distribution of the eigenvalues converges to the Marcenko-Pastur distribution. In our sample matrix, the information contained in eigenvalues situated within the Marcenko-Pastur distribution cannot be disentangled from pure randomness. Because these eigenvalues emerge only from noise, we reset their values to their average. This allows for the increase of near-zero eigenvalues and thus eliminates the singularity of the matrix.

The eigenvalue clipping method we implemented mainly relied on finding Q. Indeed, once this is done on can get λ_{max} and average smaller eigenvalues

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

In the event $T \to \infty$ and $N \to \infty$ (the dimensions of our matrix), then $Q = \frac{T}{N}$. Clearly this would be an approximation in our case. The other method that was implemented was fitting the density of the eigenvalues to evaluate Q

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

This is what has been used for our portfolio computations. A fit of the plot can be seen on 5 for the swiss stocks.

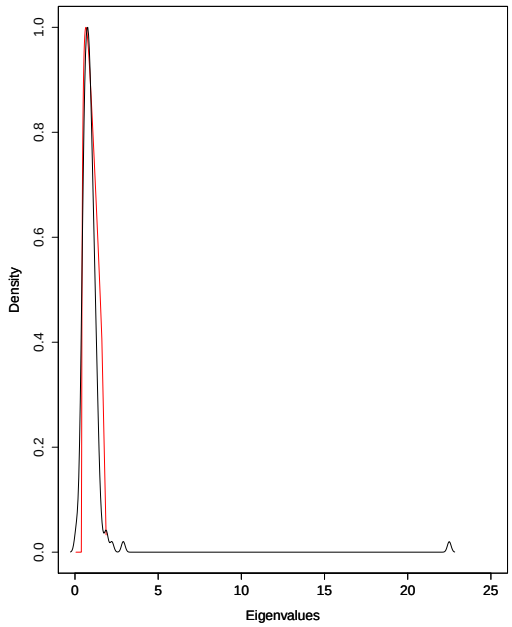

Figure 5: *Fit of the density of the eigenvalues for the swiss stocks correlation. In this case,* $Q = 7.37$ *instead of* $Q = \frac{T}{N} = 36.2$ *that we would get without the fit.*

Constrains

We can now use our freshly built correlation matrix in the mean-variance framework to obtain the weights. We have to be careful to put constrains on the magnitude of weight that a single instrument can bear. We want to avoid investing too much of the wealth into one asset. To that end, our weights may not exceed 10% of the whole portfolio. Even if it behaved very well in the past, the future remains uncertain, and we want to avoid potential large losses from an unexpected change in the dynamics of a particular asset.

3.4 Portfolio features

It is well known that stock returns share common stylized facts across markets and across time. We want to verify that our portfolio has these properties. In particular, we look at the absence of autocorrelation, volatility clustering, slow decay of autocorrelation in absolute returns (see Figure 6), heavy tails, skewness, aggregate Gaussianity (see Figure 7), and leverage effect.

In Figure 6 the plot of the daily returns shows clearly that volatility clusters in some periods of time. It is also significantly autocorrelated, a pattern that can be seen in the volatility autocorrelation function (ACF) and the absolute return ACF (which decays slowly).

On the contrary, the ACF of returns show the absence of autocorrelation – all values are below the significance interval.

In Figure 7, the excess kurtosis decreases as the time scale increases. This effect

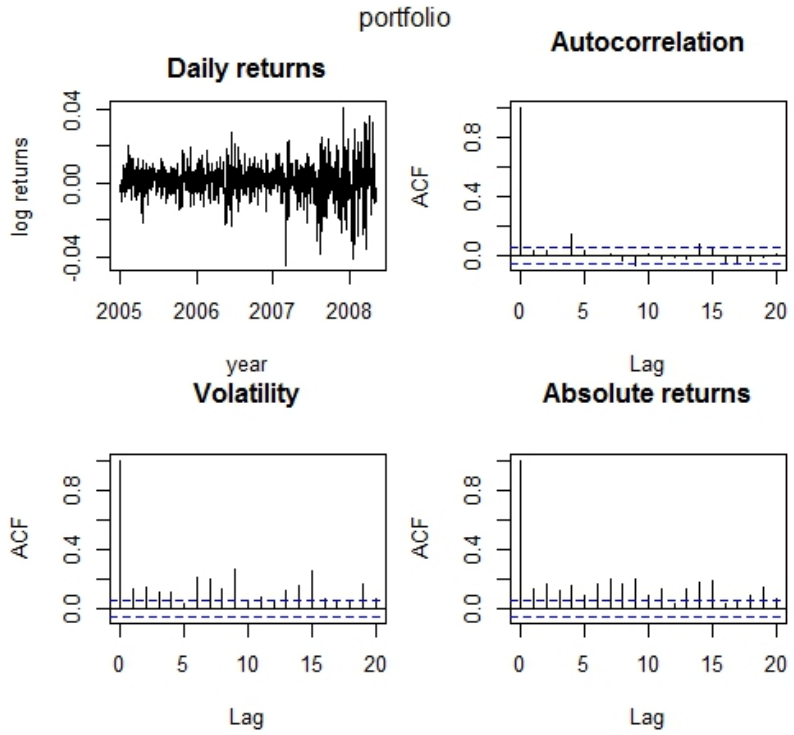

Figure 6: *Daily returns and autocorrelation functions of returns, volatility and absolute returns for our portfolio.*

is a little bit more difficult to see for the skewness. Hence the distribution goes closer to a normal distribution for larger time scales, although conducting a Jarque-Bera test for normality rejects the hypothesis of normality for all time scales.

The correlation between returns and volatility is negative: -10.5%. This is conformed to the leverage effect.

3.5 Backtesting

Taking everything that has been done before, we can backtest the portfolio on the historical data we have. This is done by restructuring the portfolio each month and computing weights based on a rolling window of five years. Results can be seen on figure 8.

Since the historical prices are only available for stocks that haven't failed, there is a strong bias with our data. This and the fact that we chose to run our backtesting prior to the 2008 crisis explains the fact that our portfolio is so successful in the first years. A better way to evaluate its performance is to compute its sharp ratio and compare it with the different strategies we have.

Using the sample correlation matrix yields the lowest sharp ratio. However, using the method with the shrinkage estimator gives us a 9% improvement on the same simulation, and the method with the eigenvalue clipping a 11% improvement. This results is consistent with simulations on other time periods.

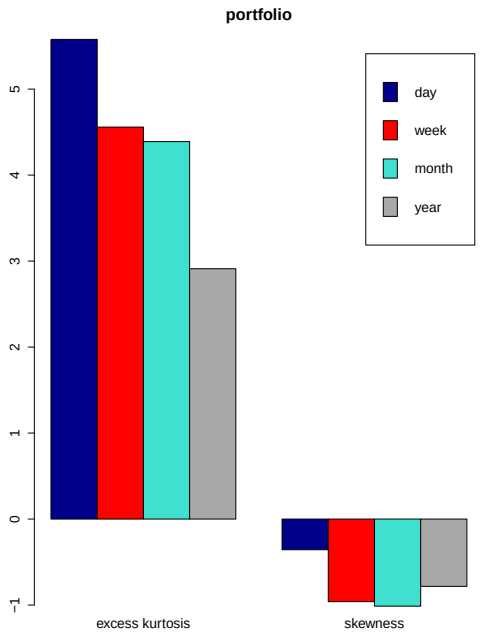

Figure 7: *Heavy tails (excess kurtosis) and asymmetry (skewness) fade away when the time scale is increased (aggregate Gaussianity).*

Conclusion

We find that more advanced methods for computing the covariance matrix give a higher Sharpe ratio than the sample covariance matrix, with the eigenvalue clipping method dominating the shrinkage matrix. This is however specific to our set of data and this assumption must be further tested.

In this work, selecting a proper set of data is of paramount importance, to be able to quantify the relations among them. We unfortunaltely had to disregard some of them due to scarcity. This is why the field of big data is so important: it allows one to take into account more and more diverse products as the data available further expands.

There are a few suggestions that we would like to present in order to further improve our investment strategy:

We assumed that the weights are continuous variables, that is that we can invest any fraction of wealth into an instrument. In reality, we deal with discrete variables: stock prices are integers. This is a fact we must take into account in our model.

In addition of reallocating the portfolio regularly, we could also set triggers when the optimality and the riskiness of the portfolio deviates too far away from our original plan. This will allow the strategy to promptly react to extreme events and avoid large losses.

The portfolio we built is solely composed of stocks. Being fully invested in stocks

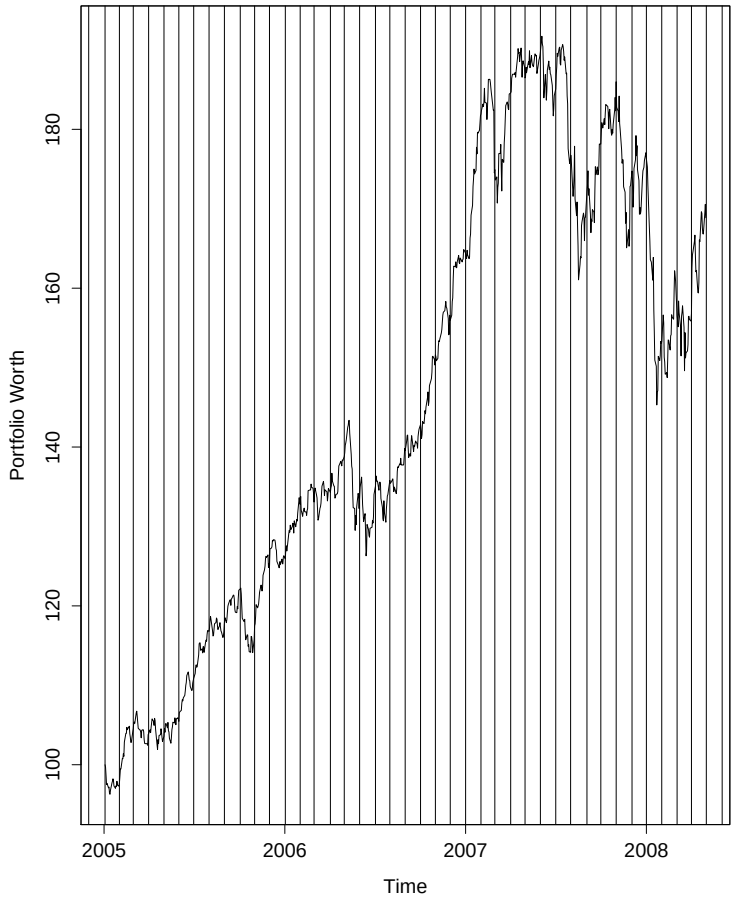

Figure 8: *Backtesting of our portfolio on the US and swiss stocks. The black lines indicate when the portfolio is restructured.*

is never good. Because of the client's medium-to-high appetite for risk, we would advise an allocation to investment grade bonds between 25% and 35% of the wealth. We could also insert alternative investment instruments (funds of funds, structured products, etc.)

We must also pay attention to the exchange rate between the denomination of the client's account and the currency of foreign placements. Particularly, if the client invests and lives in CHF, we will buy bonds denominated in CHF to avoid currency volatility, whose risk is not compensated (it is a zero-sum game). We would rather shift all the risk of the portfolio to the stock part. Investing in U.S. is safe, but we should take greater care if we aim to invest in economically more unstable countries.

References

[1] Laurent Laloux, Pierre Cizeau and Marc Potters, Jean-Philippe Bouchaud, *Random matrix theory and financial correlation* (2000)

[2] Olivier Ledoit, Michael Wolf, *Honey, I Shrunk the Sample Covariance Matrix* (2003)

[3] Rama Cont, *Empirical properties of asset returns: stylized facts and statistical issues* (2000)

[4] R.N. Mantegna, *Hierarchical structure in financial markets* (1999)