PCA as a Tool for Analyzing the Market

Efe Yigitbasi[1]

[1]*Boston University*

(Dated: May 5, 2015)

We have applied the method of Principal Component Analysis to two different markets, one consisting of stocks only, and one consisting of bonds only. We have tried to find the number of relevant dimensions for each of those markets, and inspected the eigenvectors corresponding to those dimensions. By using R^2 factor as a measure of goodness of our approximations by dimension reduction, we inspected the usefulness of PCA as a tool for analyzing different markets.

PACS numbers:

INTRODUCTION

A simple model for the market from the point of view of an investor is such that the market consists of N securities and an investment period τ. In general those securities can be a combination of bonds, commodities, mutual funds, currencies etc. One can build a time series out of the prices of such securities sampled at specific times. For a market with N such securities the prices up to a time t_0 will form a $T \times N$ dimensional matrix P_{t_0}, where T is the length of the time series for the price for one of the securities. The question is then, to determine the multivariate matrix $P_{t_0+\tau}$ which will give the prices of the securities at the end of the investment period.

This general procedure is not a simple one. We have to apply several approximations to be able to get a reasonable estimation for $P_{t_0+\tau}$. One of the approximations that can be applied is called Dimension Reduction. Although the multivariate matrix $P_{t_0+\tau}$ has N columns, in general those N securities are not independent. Therefore, by looking at the structure of the covariance matrix built by using the matrix P_{t_0} we can reduce the number of its dimension by projecting it onto a space which represents the most of the information in our matrix.

One method for applying Dimension Reduction is called Principal Component Analysis (PCA). The idea behind PCA is to find the subspace with most amount of information by looking at the eigenvalues of the covariance matrix constructed from P_{t_0}. In the following sections I'll go over the general procedure of constructing the covariance matrix from P_{t_0}, and then I'll apply PCA to two different covariance matrices that represents two different markets. Those markets are the following:

- 26 stocks from Dow-Jones Index.
- US National Treasury Bonds.

THEORY

The first step in analyzing the market is to choose our parameters. The matrix P_{t_0} is a matrix of prices, however the price of a security is not a useful parameter when we are trying build a model. If we consider our parameters to be stochastic variables, we require them to be time homogenous invariants. With this requirement the distribution of our invariants will not depend on time. This is essential for making future predictions by looking at past data. In order to find out if a variable satisfies the above requirement we can look at the lagged correlation of a given time series. For time homogenous invariants that distribution should be a circular cloud. This is easy to see by looking at two different time series for stocks. First time series to consider is the price of a stock P_t, the second one is the logarithmic (compounded) returns of the same stock $X_{t,t'}$. Where:

$$X_{t,t'} = \ln \frac{P_t}{P_{t-t'}} \tag{1}$$

It is easy to see that the price is not a time homogenous invariant, whereas compounded return is.

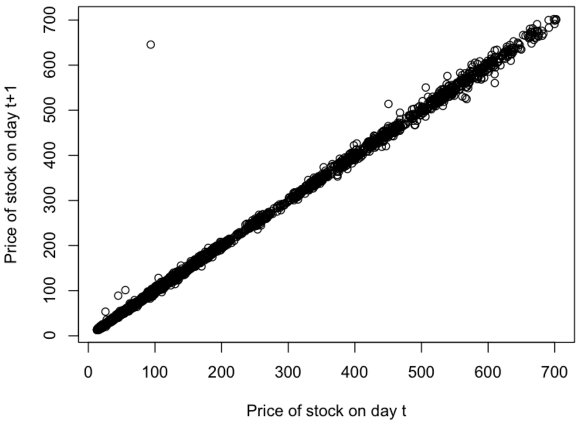

FIG. 1: Lagged correlation of the price of a stock.

The same relation applies to the price of a bond and the difference of yield to maturity for the same

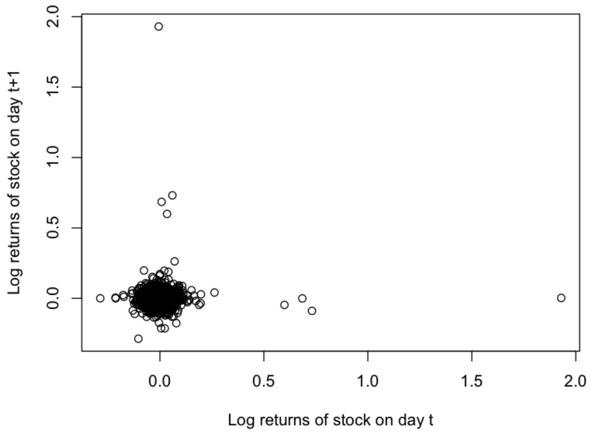

FIG. 2: Lagged correlation of the compounded returns of a stock.

bond. Once again the price of the bond is not an invariant, however the difference in yield rates is. Assume that the price of a bond at time t which will expire at time $t+v$ is given by Z_t^{t+v}, the yield rate and the difference in yield rates are given by:

$$Y_t^v = -\frac{1}{v}\ln Z_t^{t+v} \tag{2}$$

$$X_{t,t'}^v = Y_t^v - Y_{t-t'}^v \tag{3}$$

In general, for a market with both bonds and stocks the multivariate matrix that is important would be the combination of the invariants defined above. Therefore, for a market with N securities we will have a $T \times N$ matrix X which includes the invariants for all our securities.

The next step in a general market analysis is to analyze the parameter X_{t_0} up to time t_0, and make predictions for a later time $X_{t_0+\tau}$. Finally one should convert the parameter $X_{t_0+\tau}$ to the prices of the securities. Here I will only focus on the analysis of X_{t_0} by using PCA.

First we have to construct the covariance matrix using X. The covariance matrix for N securities is an $N \times N$ matrix given by the following equation where T is the number of rows in the matrix X:

$$C = \frac{1}{T}X^TX \tag{4}$$

The idea behind PCA is to find the eigenvalues and eigenvectors of C and try to extract information from them. Pictorially one can plot the location-dispersion ellipsoid by using C, and the eigenvectors of C are the principal axes of this ellipsoid. If all of our N securities are perfectly independent of each other, the location-dispersion ellipsoid will be spherical. However if there are some dependent securities, the ellipsoid might have a pancake-like distribution. For such a market we can apply PCA to find the irrelevant directions in our ellipsoid, and reduce the dimension of the ellipsoid by projecting it out of the irrelevant directions.

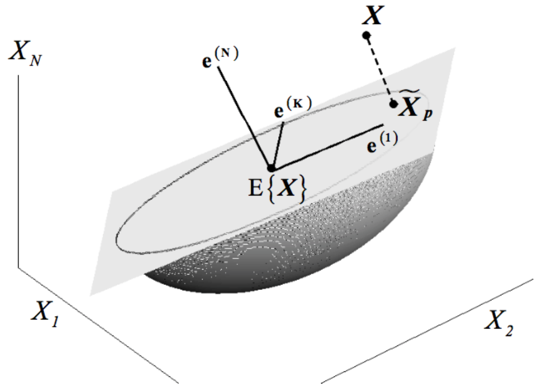

FIG. 3: Location-dispersion ellipsoid and the principal axes.

This approximation is only a good one if the principal axes that we are projecting onto correspond to nearly the entire variation in our system. A good way to measure this is comparing the relative size of the first k eigenvalues starting from the largest one, to the sum of all the eigenvalues. The parameter R^2 is defined as following:

$$R^2 = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{N}\lambda_i} \tag{5}$$

If $R^2 \approx 1$ for $k < N$ we can project the ellipsoid onto the subspace given by the first k eigenvectors without much error.

In the following sections, I will try to look at a market that consists of stocks, and a market that consists of bonds to see if we can get some dimension reduction by using PCA.

STOCK MARKET

Assume that our market consists of 26 stocks which are all in Dow-Jones Industrial Average Index. We take daily close price data from: 1/1/1990 up to: 4/20/2015. We calculate the compounded returns and the covariance matrix of compounded returns, by using the definitions that are given above. The eigenvalues and the R^2 value that we obtain by using the first k eigenvalues is given in Fig. 4 and Fig. 5.

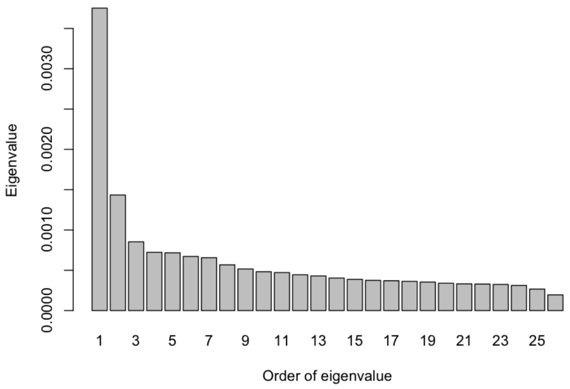

FIG. 4: Eigenvalues of the stock market.

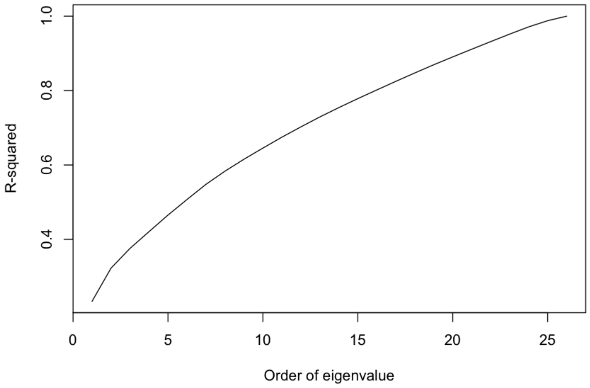

FIG. 5: R^2 if we include the first k eigenvalues.

As we can see, apart from the first one or two eigenvalues, all the eigenvalues have similar magnitudes. Therefore the value of R^2 does not become close to unity before including nearly all the eigenvalues. This means that the location-dispersion ellipsoid does not have any irrelevant directions, and we cannot reduce the number of dimensions of this market without introducing significant errors.

However, we might be able to get some information by looking at the eigenvectors that correspond to the first few eigenvalues. If we plot the contribution from all the stocks to the first eigenvector (Fig. 6) we see that it has similar contributions from all the stocks in our market. This principal axis is usually called the market mode. The market, and the prices of all the stocks in it tend to change in the same way, and this is by far the biggest contribution in the change of price of any individual stock.

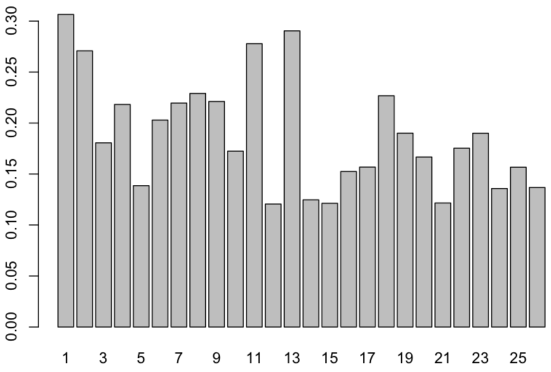

FIG. 6: Contribution of companies to the eigenvector with the largest eigenvalue.

Another information that we can get from the eigenvectors is the relations between different stocks. These relations will usually be between companies in the same industry, or similar sectors. If we look at the eigenvalue that corresponds to the third largest eigenvalue (Fig. 7), we can see that the contributions from all the stocks are close to zero except for two huge contributions from stocks labeled as 11 and 18. Those two companies are Intel, and Microsoft respectively, which are in similar sectors. The other contributions to this eigenvector mainly come from company 1 which is Apple, company 17 which is Merck & Co.,and company 20 which is Pfizer where the last two companies are both pharmaceutical companies.

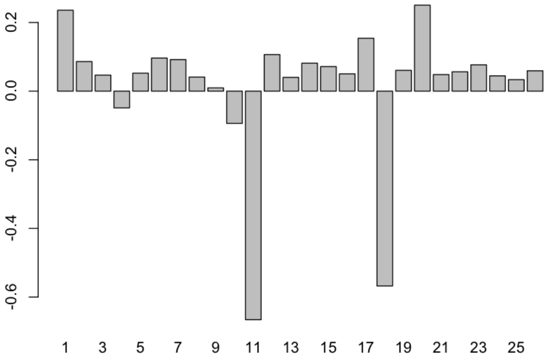

FIG. 7: Contribution of companies to the eigenvector with the third largest eigenvalue.

BOND MARKET

In this section we will look into a bond market. Assume that our market consists of several bonds with different maturities. We take the zero coupon bond prices for each maturity, and calculate the difference in yield rates. Then we construct the covariance matrix as explained above. The covariance matrix between two bonds with maturity v and $v + p$ is in the form:

$$C(v, p) = Cov(X^v, X^{v+p}) \tag{6}$$

In general for bond markets, this matrix has the following properties:

$$C(v + dv, p) \approx C(v, p + dv) \tag{7}$$
$$C(v, 0) \approx C(v + \tau, 0) \tag{8}$$
$$C(v, p) \approx C(v + \tau, p) \tag{9}$$

The above equations mean that our covariance matrix is only a function of one parameter, which is the difference in maturities of the two bonds. It is also a real, symmetric matrix which is smooth in this one argument, mostly diagonal, and constant along its diagonal. These special matrices are called a *Toeplitz Matrix.*

The infinite dimensional case of a *Toeplitz Matrix* is called a *Toeplitz Operator* and the solution for its eigenvalues and eigenvectors is given in Fig. 8.

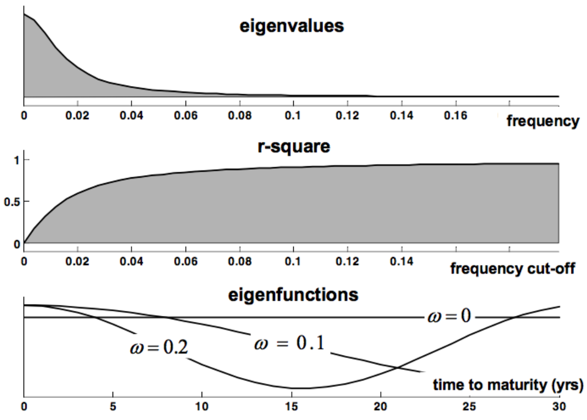

FIG. 8: Eigenvalues, R^2, and first three eigenvectors of a Toeplitz Operator.

Apparently our stock market example did not have the above given properties, and we did not see a similar pattern for the stock market. However our bond market example might have a similar eigenvalue and eigenvector composition even though they are finite dimensional. The first thing that we can check is the representation of the covariance matrix of a bond market to a stock market. As we can see in Fig. 9 the covariance matrix of our stock market is not a smooth function at all. However, it can be seen in Fig. 10 that for the US Treasury Bond Market the covariance matrix looks like a smooth two dimensional function. The other properties other than smoothness are not really visible, therefore we have to find the eigenvalues for ourselves and see if they have the same form as a *Toeplitz Operator.*

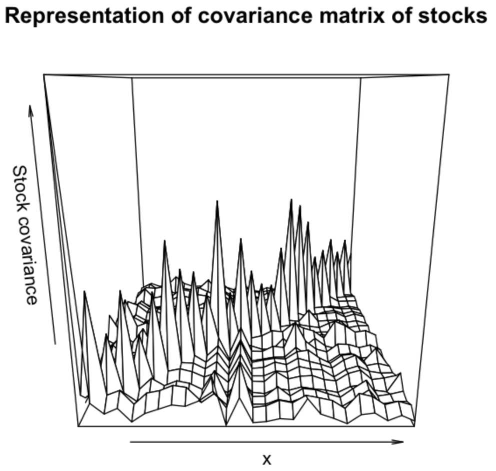
Representation of covariance matrix of stocks

FIG. 9: Representation of the covariance matrix of the stock market.

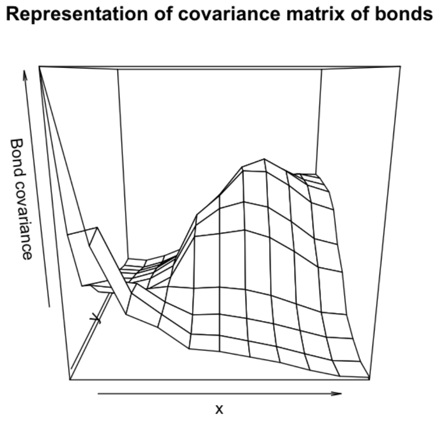
Representation of covariance matrix of bonds

FIG. 10: Representation of the covariance matrix of the bond market.

US Treasury Bond Market

In this bond market example we will look into the bond market of US Treasury Bonds. We take the daily prices from: 2/9/2006 up to: 4/20/2015, calculate the difference in yields and the covariance matrix. The eigenvalues and the R^2 value that we obtain by using the first k eigenvalues is given in Fig. 11 and Fig. 12.

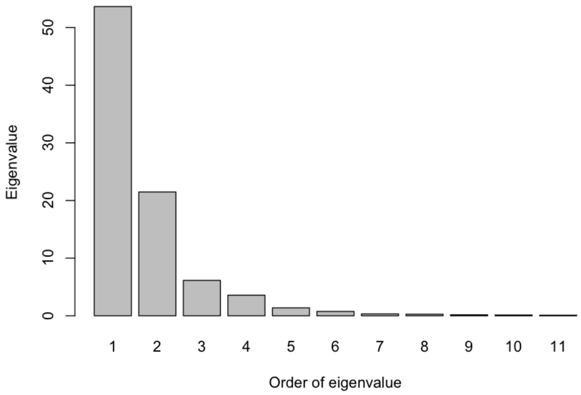

FIG. 11: Eigenvalues of US bond market.

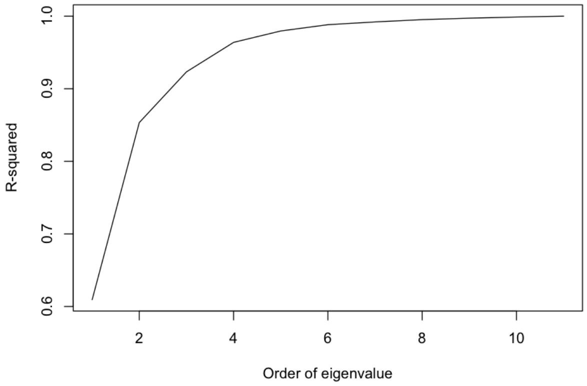

FIG. 12: R^2 if we include the first k eigenvalues.

We can see that the eigenvalues are completely dominated by the first four eigenvalues, and that the R^2 goes above 0.95 if we include the first four eigenvalues. This means that the US Treasury bond market is not really 11 dimensional as our invariants suggest, but it is in fact close to being four dimensional. Looking at the contributions to the first four eigenvectors, we can see that the eigenvectors resemble the eigenvectors of the *Toeplitz Matrix* solution (Fig 13). However due to the covariance matrix being not an exact *Toeplitz Matrix* we can see that the eigenvectors are not exactly the oscillatory solutions we see in the exact case.

Now the last step is to actually reduce the number of dimensions by projecting our invariants in the direction of our eigenvectors. By doing this we are basically mixing our bonds in such a way that our invariant matrix now consists of our eigenvectors instead of our bonds. By doing this we can see the different things happening in our market more easily than the previous case. By looking at the time series of the first three eigenvectors, we can see that there are some points where the values change dramatically (Fig. 14, Fig. 15, Fig. 16). In this case the location of these jumps (around day 500 in our time series) happens to be around the year 2008. We can see that the market has some unusual price changes around that period.

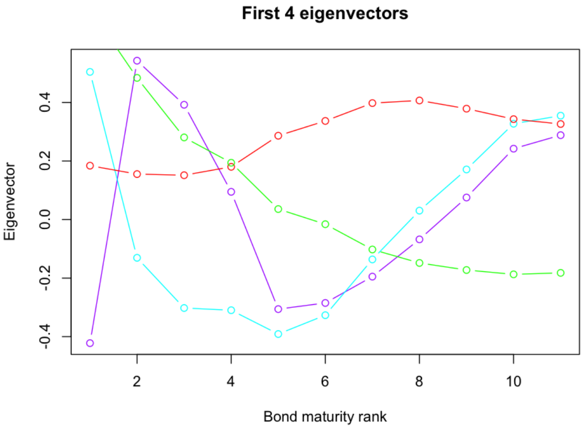

FIG. 13: Contributions to the first four eigenvectors.

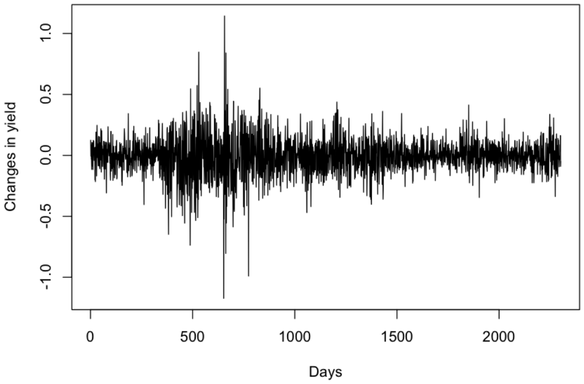

FIG. 14: Time series of the first eigenvector.

CONCLUSION

I have applied the method of Principal Component Analysis to an example of a stock market and an example of a bond market. For both cases the method is useful in having a better understanding of the market. However the information gathered from PCA is different for the two examples that are given here.

For the stock market PCA does not help in reducing the number of relevant dimensions in our market.

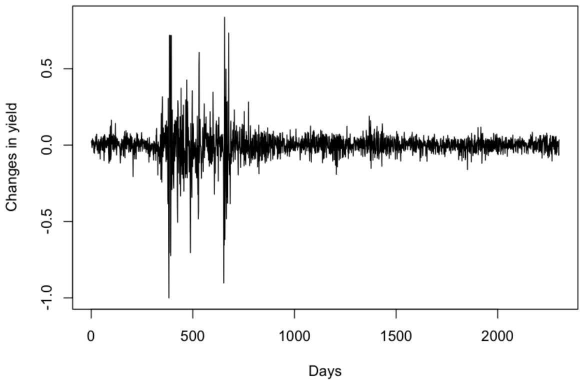

FIG. 15: Time series of the second eigenvector.

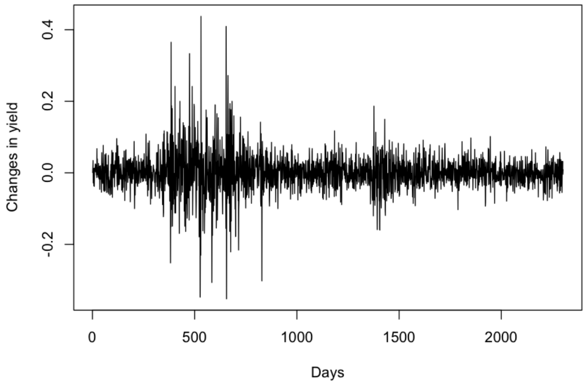

FIG. 16: Time series of the third eigenvector.

It is possible that for markets with greater number of stock than 26 that I have used here, PCA might give some dimension reduction, however it was not the case here. The usefulness of PCA in stock market example was to get an understanding of underlying connections between the prices of stocks of different companies.

For the bond market, PCA helped us identify the relevant directions in the location-dispersion ellipsoid. For the US Treasury Bond Market, the number of relevant dimensions was 4 out of a total number of 11 different bonds. This is a huge dimension reduction. Also the plot of eigenvectors shows that the changes in yield rates of bonds can be analyzed with a few simple functions. By far the greatest contributors to the R^2 factor is the first two eigenvalues. The effect of the first two eigenvectors can be understood to be a constant shift in the bond yield rates (from first eigenvector), and a skew (from the second eigenvector). These contributions are easy to understand, visualize and use when trying to understand the behavior of the market. The effect of a financial crisis is clearly visible on the time series of the most important eigenvectors. This information can also be useful in understanding the overall behavior of the bond market during financially unstable periods.

[1] Risk and Asset Allocation, Meucci, Attilio