International Journal of Pure and Applied Mathematics
Volume 115 No. 1 2017, 153-167
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: 10.12732/ijpam.v115i1.12

PRINCIPAL COMPONENT ANALYSIS FOR STOCK PORTFOLIO MANAGEMENT

Giorgia Pasini
Department of Computer Science
University of Verona
Strada le Grazie, 15-37134, Verona, ITALY

Abstract: In this paper the method of Principal Component Analysis is applied to three subgroups of stocks of the american index Down Jones Industrial (DJI) Average. While, the first and second group, are homogeneus, the third one contains heterogeneous stocks. Cumulative Variance and Kaiser's Rule are used to get the principal risk directions. The obtained results show how to optimize portfolios investments to derive the best returns and financial control.

AMS Subject Classification: 62H25, 62M10, 62P05, 62P20, 91B30, 91B84
Key Words: principal component analysis, cumulative variance, Kaiser's rule, portfolio management, stocks management, financial engineering

1. Introduction

The Principal Component Analysis is a method of multivariate analysis. The idea of Principal Component Analysis (PCA) is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, in order to maximize the variance of a linear combination of the variables. It is a method applied to data with no groupings among the observations and no partitioning of the variables into subsets y and x.

The elements, obtained applying this method, are the Principal Compo-

Received: May 9, 2017
Revised: June 10, 2017
Published: June 28, 2017

nents. The first one is the linear combination with maximal variance, the second one is the linear combination with maximal variance in the orthogonal direction to the first principal component and so on for the others. Moreover they are ordered sequentially with the first one explaining as much of the variation as it can.

Computation of the the principal components reduces to the solution of an eigenvalue and eigenvector problem for a positive semidefinite symmetric matrix.

In order to apply PCA, one can use either covariance and correlation for data because properties can be adapted to every case, as explained in [12]. It is worth to mention that the PCA method can be effectively used to increase the goodness of results obtained by exploiting different statistical methods. In particular the PCA approach is widely used in the emerging scenario of energy markets, where seminal statistical analysis have been already performed to forecast w.r.t. the unbalance previsions, see, e.g., [8]. Moreover PCA analysis could be used in interaction with rather sophisticated data analysis tools such those implemented in the regime switching type analysis of time series, as in the case, e.g., of [9, 10], or by using *neural networks*, resp. *statistical mechanics* approach, see, e.g., [11], resp. [7], or in the context of binary calssification, see, e.g. [6].

Two methods are used to choose the number of the principal components: the Cumulative Variance and the Kaiser's Rule.

The idea of the first one is choosing a cumulative percentage of total variation which the selected principal components should contribute, e.g. 80% or 90%. Hence the required number of principal components is the smallest value of m for which the percentage we chose is exceeded. The variance of the k-th principal component is l_k, and $\sum_{k=1}^{p} l_k = \sum_{j=1}^{p} s_{jj}$, the sum of the variances of the principal components, is equal to the sum ofthe variances of the elements of x. The definition of percentage of variation accounted for by the first m principal components' is therefore: $t_m = 100\frac{\sum_{k=1}^{m} l_k}{\sum_{j=1}^{p} s_{jj}} = 100\frac{\sum_{k=1}^{m} l_k}{\sum_{k=1}^{p} l_k}$, which reduces to $t_m = \frac{100}{p}\sum_{k=1}^{m} l_k$ in the case of a correlation matrix. One chooses a cut-off t^* somewhere between 70% and 90%, retaining m principal components, where m is the smallest integer for which $t_m \geq t^*$. The best value for t^* is generally smaller as p increases, or as n increases. For more details one can see [12].

The Kaiser's Rule, instead, is based on the size of variances of principal components; the idea is to retain only those principal components whose variances l_k exceed 1. Indeed, if all elements of x are independent, then the principal components are the same as the original variables and all have unit variances in the case of correlation matrix. Thus any principal component with variance

less than 1 contains less information than one of the original variables and so is not worth retaining. Moreover, if the data set contains groups of variables having large within-group correlations, but small between group correlations, then there is one principal component associated with each group whose variance is ≥ 1, whereas any other principal components associated with the group variances ≤ 1. Hence, the rule will generally retain one, and only one, principal component associated with each such group of variables.

This rule is usually used with correlation matrix, but it can be adapted also to the covariance matrix. Details are explained in [12].

2. Application to Stocks

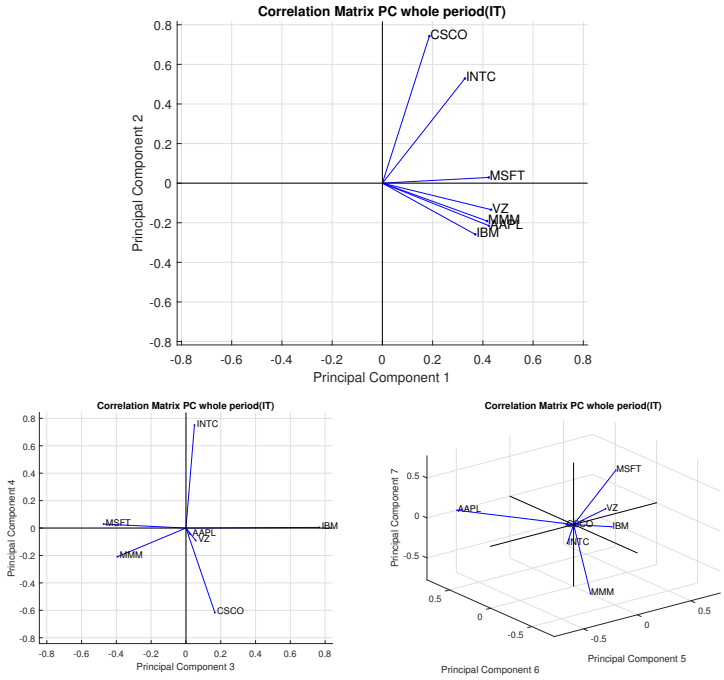
PCA is applied to three subgroups of stocks of the American Index DJI, Dow Jones Industrial Average. Data have been obtained from *Yahoo Finance* and three subgroups of them are considered: the first one contains technology and telecommunications stocks, the second one the financial and credit stocks and the last one mixes the other two groups. We would like to underline that a similar, in terms of the PCA analysis, approach can lead to the determination of a suitable index to measure the degree to which a time series departs from white noise with interesting financial applications as done in [1].

The informatics and telecommunications group is composed by these seven stocks: AAPL (Apple), CSCO (Cisco System), IBM (IBM), INTC (Intel), MMM (3M), MSFT (Microsoft), VZ (Verizon Communications). The financial and credit group, instead, is composed by these five stocks: AXP (American Express Company), GE (General Electric), GS (Goldman Sachs), JPM (JP Morgan and Chase), TRV (The Travelers Companies). Visa has been excluded from the last five ones because it became element of DJIA many years later than the other stocks, so it would not allow a complete analysis.

In the application of PCA three periods are distinguished: from 17/03/2000 to 13/01/2017 for the whole period, from 17/03/2000 to 31/12/2007 for the period before the crisis in 2008, and from 2/01/2009 to 13/01/2017, which constitutes the period after the crisis. Details about 2008 financial crisis can befound at *Financial Crisis 2007/2008*. Both the covariance and the correlation matrices can be used for application, but in this paper the second one is used, because the study of the risk direction is the goal of the work and correlation is the best method to do it. Indeed in this way data are standardized and there are not stocks with the highest variation standing out.

a) Technology and Telecommunications Stocks

Figure 1: PCA application to Technology and Telecommunications Stocks



Focus on the correlation matrix approach: applying the PCA method, first compute its values, then decompose the obtained matrix exploiting the SVD. The latter allows to get the eigenvalues and the eigenvectors matrices, respectively. In this section the whole period is distinguished from the periods before and after crisis in 2008.

To what concerns the whole period, results are resumed by the biplots in Figure 1:

The first principal component usually can be seen as the market component, i.e. that component in which stocks should have equal contribution; this is the reason for all the stocks should have the same sign. In the first biplot every stock has the same sign and moreover all the coefficients are positive. Instead the second principal component has both positive and negative coefficients. Looking at the second and third biplots one can not say very much about the other components, but it can be noticed that there are stocks, as MSFT ans IBM in the third principal component, with opposite directions; when this happens, it means that stocks have similar structures but in opposite ways.

The cumulative variance and the Kaiser's rule were applied in order to know how many components to retain; in the first method a percentage of 80% has been chosen. The result is that retaining two principal components is enough to explain most of the total variation.

Therefore, an analysis about variance explained by components in different periods has been performed; it is resumed by Table 1. Then its variation and the median value it assumed in different periods can be seen in the plots in Figure 2.

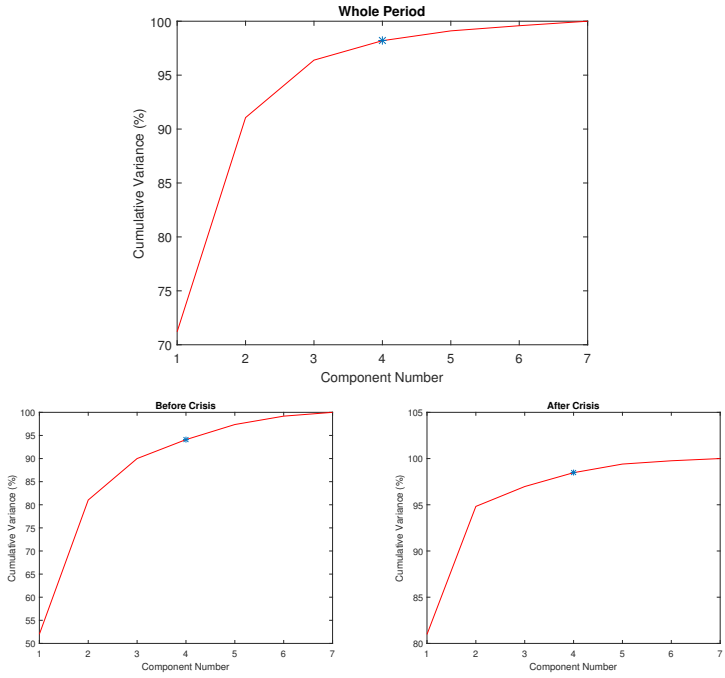
Table 1: Stocks variation in different three periods

Eigenvalues and Cumulative Variance						
	Whole Period		Before Crisis		After Crisis	
PC	Eigval	Cum Var(%)	Eigval	Cum Var (%)	Eigval	Cum Var (%)
1	4.9843	71.2044	3.6394	51.9918	5.6683	80.9764
2	1.3901	91.0624	2.0329	81.0328	0.9693	94.8228
3	0.3730	96.3911	0.6278	90.0015	0.1505	96.9726
4	0.1263	98.1954	0.2866	94.0964	0.105	98.4718
5	0.0637	99.1050	0.2286	97.3620	0.0653	99.4042
6	0.0334	99.5825	0.1271	99.1782	0.0249	99.7599
7	0.0292	100	0.0575	100	0.0168	100

Considering the same percentage of cumulative variance explained as before, 80%, for the whole period and the time before crisis, two principal components are enough. But for the period after the crisis, just an only principal component, which is the first one, has to be retained. It has to be noticed that the variance explained by the first principal component before 2008 was 51.9918%, and after 2008 it became higher: 80.9764%. Analogously, it can be observed a high increment in the variance value of the second component. Such a phenomenon also happens for the other ones, but its magnitude is rather small. Therefore, it is clear that the first two components, with a predominant role played by the first one, are responsible for almost the total variation. Hence the first two principal components, in particular the first one, absorbed a lot of variation after the crisis.

In the last part of this section the correlation matrix has been decomposed into eigenvalues and eigenvectors matrices. Then the second one has been considered: every row of the matrix corresponds to an asset contained in the original investment universe, instead every column represents a principal portfolio. Every coefficient of each column corresponds to a long position if it is positive, to a short position if it is negative. The eigenvalue corresponding to

Figure 2: Cumulative variance in different three periods



every column is the proportion of total portfolio (original portfolio) standardized variance that can be attributed to that factor. Consider the whole period: the first principal portfolio has an eigenvalue which explains the 71% of the total variance, hence it should reproduce very well the original portfolio, as can be see in the plots in Figure 3. The original portfolio is constructed with the strategy $1/N$, in which every stock is considered having equal contribution. In the plot the portfolio had at the beginning has been compared to the first principal portfolio and the year 2008 hs been pointed out with a red $*$. Both portfolios are influenced by data related to this year, with a particular decrease of the return for the PCA-portfolio. This is because it can be seen as the market component, so it reproduces rather accurately the new trend.

But in the study of the correlation matrix it has been found that if one considers the whole period, needs just two principal components to explain the total variation. In particular looking at the plot of the return for the second principal portfolio, it can be seen that it reproduces the trend of the initial one, but not so accurate as in the case of the first principal portfolio as one can expect. Its return is not higher than the original portfolio return, but it is

Figure 3: Comparison between initial portfolio and first and second principal portfolios

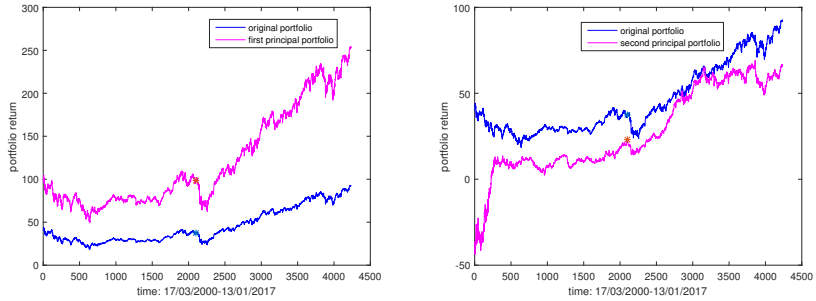
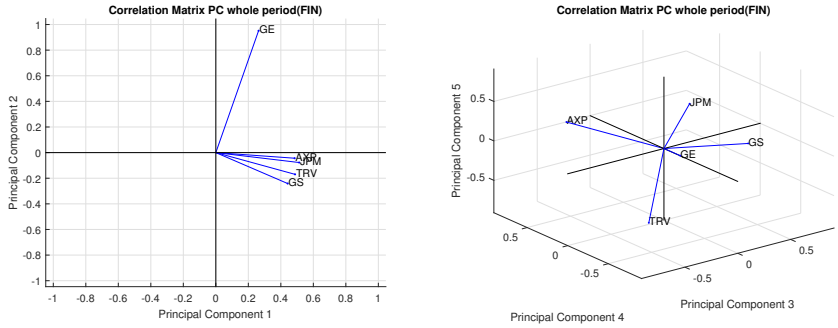


Figure 4: PCA application to Credit and Financial Stocks



anyway positive except for a little period at the beginning. It is influenced by the crisis, too, and it can be seen a decreasing corresponding to the year 2008, marked with $*$.

Summing up results, it can be seen that the two aforementioned components allow for the best description of the variance. It follows that investing in such components turns to be preferable to better control financial risk also getting higher returns.

b) Financial and Credit Stocks

Apply PCA to correlation matrix, standardizing data of the five financial and credit stocks. Three different periods are distinguished as done for technology and telecommunications stocks: whole period, before and after 2008 crisis. Results of the PCA application are resumed in biplots in Figure 4.

It can be seen that all the coefficients of the first principal component are

positive and recall that the reason is that it represents the market component in which every stock should have the same contribution. The second principal component coefficients are all negative, except one: GE. In the second biplot there is a star shape and this means that there are coefficients with opposite directions that have similar behaviors and structures, but in opposite way.

In order to know how many components are enough to explain total variation, one needs the cumulative variance or the Kaiser's rule. Give a look to Table 2. Consider the whole period: first of all look at the cumulative variance.

Table 2: Stocks variation in different three periods

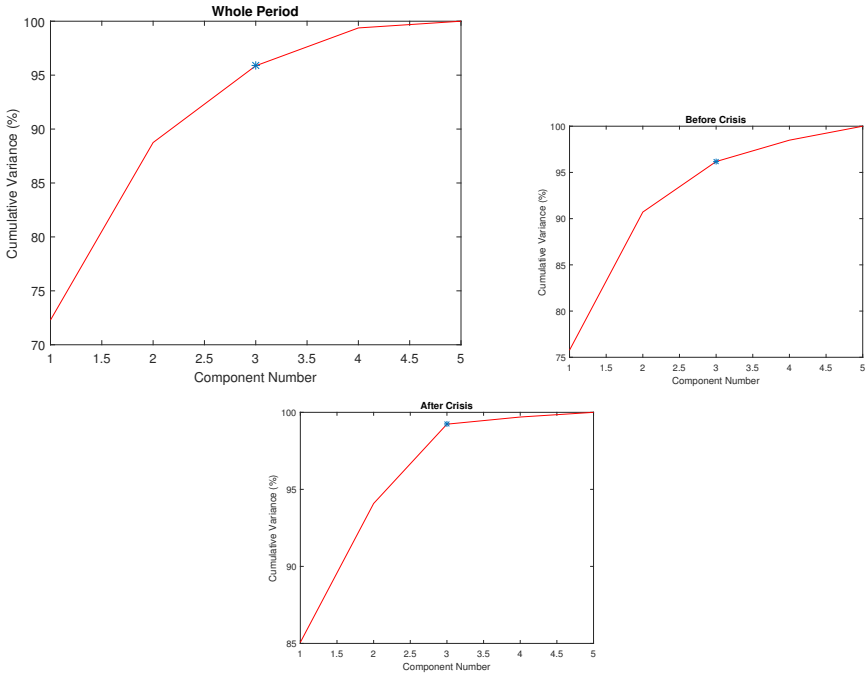
Eigenvalues and Cumulative Variance						
	Whole Period		Before Crisis		After Crisis	
PC	Eigval	Cum Var (%)	Eigval	Cum Var (%)	Eigval	Cum Var (%)
1	3.6147	72.2932	3.7885	75.7698	4.2535	85.07
2	0.8228	88.75	0.7471	90.7114	0.4507	94.0832
3	0.3559	95.8689	0.2733	96.1773	0.2575	99.2324
4	0.1754	99.3770	0.1156	98.4895	0.0232	99.6962
5	0.0311	100	0.0755	100	0.0152	100

If a percentage of 80% is fixed the number of components to retain is two. Now focus on the Kaiser's rule: it establishes to retain all the components whose eigenvalues are greater than 1. In this case there is an only eigenvalue, the first one, that satisfies the rule. However, the second eigenvalue is really near to 1, hence this is a particular case studied by Jolliffe in [12]. He suggests to retain those components whose eigenvalues are greater than 0.7 to not loose too information and not forget any important components in case of sample with small dimensions. For more details see [12]. Hence, in this case, retaining two components is convenient, observing that one gets the same result got before in discussion of the cumulative variance.

About the three different periods, there are some considerations based on the cumulative variance changing. As seen for technology and telecommunications stocks analysis, the crisis influenced very much the components. The first principal component is the most affected element: its cumulative variance, 51.9918% before 2008, becomes 80.9764% after that year; this means it absorbed most of the variation after the crisis. Also other components are influenced by the same event but not so much as the first one, as we can see in Table 2. Cumulative variance behaviour can be seen in plots in Figure 5.

Eigenvalues and eigenvectors matrices of the correlation matrix have been already found; now take the eigenvectors in order to study the principal port-

Figure 5: Cumulative variance in different three periods



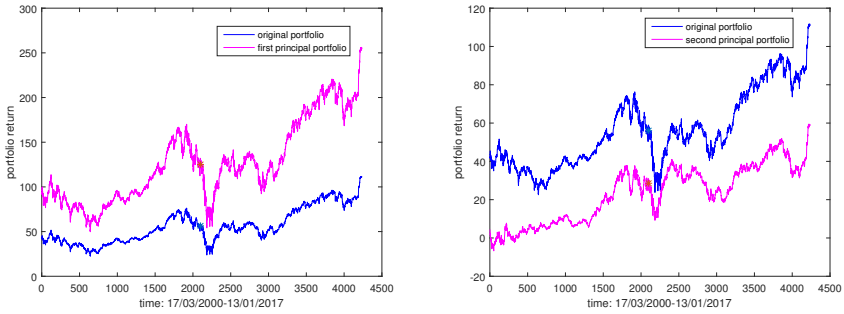
folios, as done for the other group of stocks. For this analysis consider again the whole period and the original portfolio constructed with the method $1/N$. Since two components are enough to explain most of the variation from the analysis, also in this case the best strategy will be to invest in the (first) two principal portfolios.

The first principal portfolio is the element that describes in the best way the original portfolio, because it has the highest eigenvalue that explains most of variation: 72.2932%. Also in this case the red $*$ in the plots in Figure 6 correspond to the crisis in 2008 and it can be observed how it influenced the returns of the portfolios.

c) Mixed Stocks Analysis

In this section previously seen time series are taken into account to compute the correlations and order them in a matrix. Then, applying PCA, eigenvalues and eigenvectors matrices are obtained. Again, three different periods are distinguished: whole period, before and after crisis. First of all consider the biplots of principal components in Figure 7.

Figure 6: Comparison between initial portfolio and first and second principal portfolios



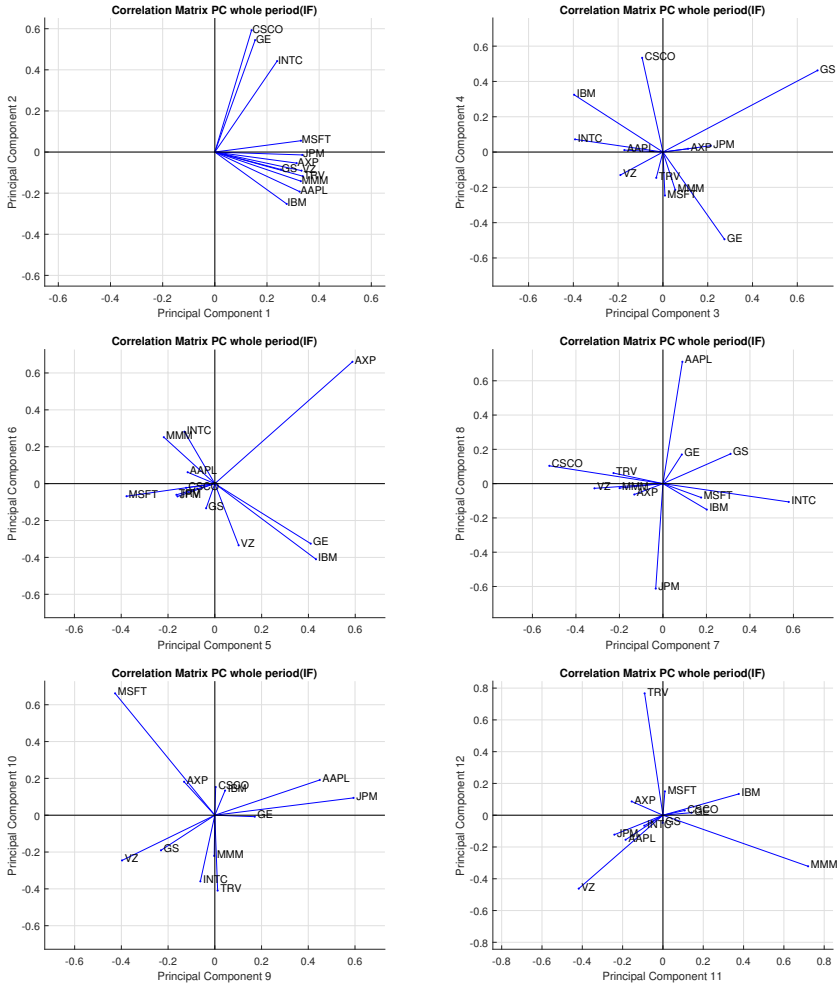
The first principal component should describe the market and its coefficients are, in fact, all positive. In this case all the stocks should be considered having equal contribution. The second principal component has either positive and negative coefficients. All the other biplots have a star shape. They show us stocks having same structures and behaviors but in opposite ways.

About how many components to retain, cumulative variance and Kaiser's rule are applied. Fix the same percentage as before for every single group of stocks: 80%. Hence in this case (the first) two components satisfy it. About the Kaiser's rule, taking the eigenvalues greater than 1 is enough. Also for this case, the number of components to retain is two.

Considering the periods before and after 2008, resumed in Table 3, the cumulative variance increases very much after that year. The most influenced element is the first one, indeed its cumulative variance changes from 51.9738% to 81.1339%. Also in this case, the increasing of cumulative variance happens also for the other components, but its magnitude is small. By the way, the first two principal components are the elements that absorbed most of the variation of the crisis. The cumulative variance behavior can be seen in the plots in Figure 8.

In this last section consider data of whole period. The first principal portfolio has the highest eigenvalues, so it should reproduce very well the initial one as we can see in the first plot. The original portfolio is constructed with the strategy $1/N$, in which every stock is considered having equal contribution. In the plots in Figure 9 the red symbol * points out the year of the crisis, 2008. Previous plots show that, in correspondence of *, there is a significant decrease, which implies that the returns of each portfolio is highly influenced by it. In

Figure 7: PCA application to Mixed Stocks



particular, the PCA-portfolios are the most affected.

Studying the correlations, the result was that two components were enough to explain most of the variation magnitude. Therefore is natural to consider also the second principal portfolio. It can be seen that it reproduces the trend of the original portfolio but not well as the first one, even if its return is positive almost everywhere. This is also influenced by 2008, but not so much as the first principal portfolio. The reason is that the first element is the market component and it is reproduces exactly the behavior of the real market. Since results show

Table 3: Stocks variation in different three periods

Eigenvalues and Cumulative Variance						
	Whole Period		Before Crisis		After Crisis	
PC	Eigval	Cum Var(%)	Eigval	Cum Var (%)	Eigval	Cum Var (%)
1	8.2574	68.8114	6.2369	51.9738	9.7361	81.1339
2	1.8913	84.5718	3.4359	80.6059	1.2701	91.7182
3	0.7080	90.4716	0.8611	87.7814	0.4615	95.5640
4	0.4817	94.4859	0.3914	91.0430	0.1694	96.9755
5	0.3168	97.1259	0.2988	93.5332	0.1252	98.0191
6	0.1361	98.2604	0.2297	95.4470	0.0964	98.8225
7	0.0798	98.9254	0.1886	97.0188	0.0641	99.3571
8	0.0445	99.2965	0.1488	98.2587	0.0287	99.5962
9	0.0291	99.5388	0.0981	99.0758	0.0185	99.7505
10	0.0241	99.7400	0.0631	99.6012	0.0127	99.8562
11	0.0201	99.9074	0.0301	99.8523	0.0100	99.9397
12	0.0111	100	0.0177	100	0.0072	100

that the portfolios studied in this section describe in the best way the variance, the conclusion is that investing in these two give the best way to get financial control and high returns.

Conclusions

Results prove that the first element of PCA can be seen as the market component and it has been proved in first biplots of evey studied case. In fact the coefficients of the first PC were always all positive. It could happen also that all the coefficients are negative; in this case one can rotate the components and get the same result. In this paper it has not been necessary because every stock in first principal component was positive.

About the method used to choose the number of the components, every time a percentage of 80% for cumulative variance has been chosen. Instead the value for Kaiser's rule has been modified in the second application because in this particular case dimension of the sample was very small and maybe too much information could be lost. Furthermore in the first and last analysis, taking the eigenvalues greater than 1 was a good choice because sample dimension was

Figure 8: Cumulative variance in different three periods

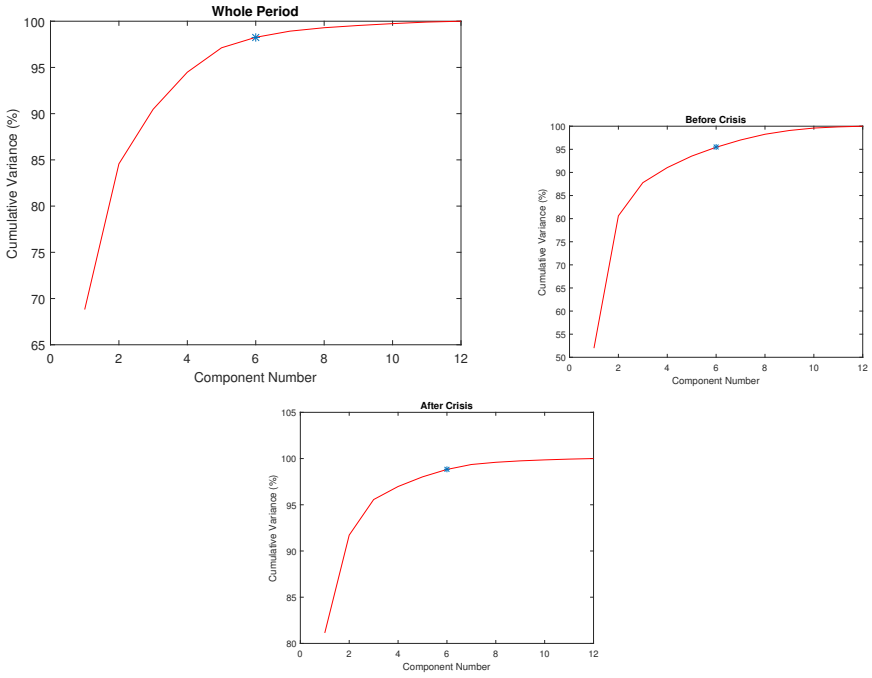
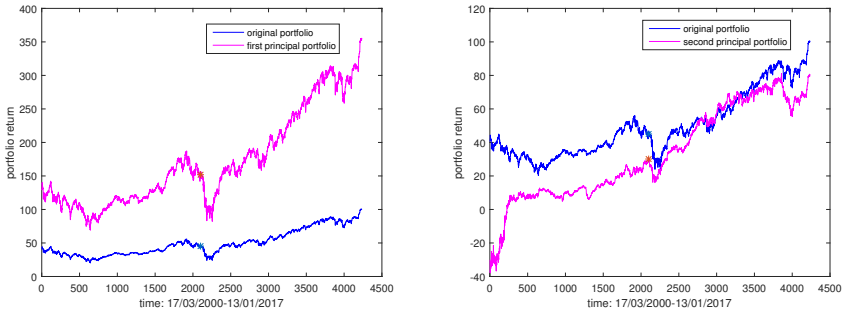


Figure 9: Comparison between initial portfolio and first and second principal portfolios



greater. It is interesting notice that also the percentage one has to choose for cumulative variance can be modified: usually one can take a value from 70% to 90%. The choice depends on data one has. It is worth to mention that such type of analysis do not cover the investor against hidden financial risks, as those

related to deteriorated financial positions of the company's stocks involved in the PCA selection procedure, or against possible default contagion over financial networks, see, e.g., [3], etc. About the method of PCA, it is very useful if one wants to perform a market analysis or diversify the risk, because it keeps in evidence the best directions in which one could invest. Following this way one should get the best portfolio returns, controlling as best as possible the risk. Instead the application of PCA is not a good method if one wants to know how many stocks to retain in a portfolio and this result is given by application of PCA to covariance matrix, that will be the subject of a forthcoming paper.

References

[1] Aguilar, R. R., Cruz-Aké, S., Martínez, F. M., *The Mahalanobis Distance between the Hurst coefficient and the Alpha-Stable parameter: an early warning indicator of crises*, International Journal of Pure and Applied Mathematics Volume 110 No. 2, 283-310, 2016.

[2] *Yahoo Finance*. URL: http://www.wallstreetoasis.com/financial-crisis-overview, March 2017, 06.

[3] Benazzoli, C., Di Persio, L., *Default contagion in financial networks*, International Journal of Mathematics and Computers in Simulation, 10, pp. 112-117, 2016.

[4] *Financial Crisis 2007/2008*. URL: http://www.wallstreetoasis.com/financial-crisis-overview, March 2017, 06.

[5] Bilodeau, M., Brenner, D., *Theory of Multivariate Statistics*. Springer Texts in Statistics, Springer New York, 1999.

[6] Boulesteix, A.-L., *A note on between-group PCA*, International Journal of Pure and Applied Mathematics, Volume 19 No. 3, 359-366, 2005.

[7] Crescimanna, V., Di Persio, L., *Herd Behavior and Financial Crashes: An Interacting Particle System Approach*, Journal of Mathematics, 2016, art. no. 7510567, 2016.

[8] Di Persio, L., Cecchin, A., Cordoni, F., *Novel approaches to the energy load unbalance forecasting in the Italian electricity market*, Journal of Mathematics in Industry, 7 (1), art. no. 5, 2017.

[9] Di Persio, L., Frigo, M., *Maximum likelihood approach to markov switching models*, WSEAS Transactions on Business and Economics, 12, pp. 239-242, 2015.

[10] Di Persio, L., Frigo, M., *Gibbs sampling approach to regime switching analysis of financial time series*, Journal of Computational and Applied Mathematics, 300, pp. 43-55, 2016.

[11] Di Persio, L., Honchar, O., *Artificial neural networks architectures for stock price prediction: Comparisons and applications*, International Journal of Circuits, Systems and Signal Processing, 10, pp. 403-413, 2016.

[12] Jolliffe, I.T. (1986). *Principal Component Analysis*, Second Edition. Springer Texts in Statistics, Springer New York, 2002.

[13] Wolfang Härdle, Zdenk Hlvka. *Multivariate Statistics: Exercises and Solutions*. Springer Finance Textbooks, Springer Science, 2007.

[14] Libin Yang. *An Application of Principal Component Analysis to Stock Portfolio Management*. Department of Economics and Finance, University of Canterbury, January 2015.