Principle Component Analysis on Stock Returns

ISYE/MATH 6783 - Assignment2

Quan Zhou
Quantitative and Computational Finance
Georgia Institute of Technology
Email: qzhou81@gatech.edu

***Abstract*—Factor model is a widely used method to characterize the relationship between several variables. As the dimensions of factors increases, the cost of computation grows at a tremendous speed. However, these factors are often highly collineared and there are only a few important data information. Principle component analysis(PCA) is able to identify those important variables and reduce dimensionality, saving a great amount of computing time. In this report, PCA is performed on the log return of 12 stocks. Results shows 9 of them are enough to explain the variance and the rest can be eliminated.**

I. Introduction

In the finance industry, factor model is widely used as an efficient method to express the relationship between different variables. Classical factor model includes the FamaFrench three-factor model[1], the Carhart four-factor model[2], the Five-factor model[3] and Prof. Deng's Seven-factor portfolio trading strategy[4].

The financial market is an extremely complex and unpredictable system. In most cases, there are dozens of factors that could have an influence. The problem is that with the number of factors increasing, the cost of computation also grows at a tremendous speed. Knowing the fact that some of these factors are highly correlated and there are only a few important data information among them, it would be much more efficient if some of the unimportant factors can be emilinated without hurting the result. For example, the computational complexity of linear regression is quadratic. If half of the factors are eliminated, the whole process will be four times faster. Principle component ananlysis(PCA) is able to identify the most important factors in a model with the least collinearity, reducing dimensionality. In this report, the log return of 12 stocks are provided. By applying principle component analysis, 3 of them can be eliminated.

The rest of this report is organized as follows. Section 2 will introduce the basic process of principle component analysis. Detailed data manipulation and result interpretation will be introduced in Section 3, followed by a brief conclusion in Section 4.

II. Principle Component Analysis

A. Preliminaries

Before performing principle component analysis, the data must satisfy some conditions. First of all, the input data has to be stationary, which means its joint probability distribution remains the same all the time. The data used in this report is the log return of stock prices. It is stationary.

Since the purpose of principle component analysis is finding the componenets that explain most of the total variance, the input data also has to be normalized. Otherwise, different data may have different weights from the beginning. The first principle component is more likely to be the component with the largest volatility.

B. Solving Process

Given a dataset with *k* factors and *n* observations, the first step is to compute the variance-covariance matrix of the *k* factors. Then find the vector that inherit the maximum possible variance from the matrix in descending order. The detailed process is decribes as below.

First Principle Component:

$$a_1 = argmaxVar(a'Va)$$

$$s.t.||a_1|| = 1$$

Second Principle Component:

$$a_2 = argmaxVar(a'Va)$$

$$s.t.||a_2|| = 1 \ and \ a_1'a_2 = 0$$

Third Principle Component:

$$...$$

In the end, principle component analysis gives the portion of total variance explained by each factor.

III. Data and Result Interpretation

The data provided consists of the log return of 12 stocks on over 1200 days. The variance-covariance matrix of these 12 stocks is shown below.

The covariance matrix indicates that the data more or less is correlated with each other. Principle component analysis should be able to eliminate some of the stocks in the original dataset.

Before running PCA, the data has to satisfy the two conditions mentioned in Section 2. Since the data is log return, it can be considered as stationary. Also it needs to be normalized. This is done by setting the parameter *scale* to *TRUE* when performing PCA in R.

TABLE I
VARIANCE-COVARIANCE MATRIX

	aapl	adbe	adp	amd	dell	gtw	hp	ibm	msft	orcl	sunw	yhoo
aapl	1.00	0.39	0.28	0.45	0.51	0.35	0.43	0.43	0.48	0.41	0.42	0.42
adbe	0.39	1.00	0.31	0.38	0.49	0.27	0.42	0.45	0.53	0.45	0.41	0.5
adp	0.28	0.31	1.00	0.31	0.32	0.2	0.34	0.39	0.37	0.32	0.27	0.32
amd	0.45	0.38	0.31	1.00	0.47	0.35	0.43	0.45	0.46	0.38	0.44	0.42
dell	0.51	0.49	0.32	0.47	1.00	0.37	0.53	0.53	0.62	0.51	0.51	0.5
gtw	0.35	0.27	0.2	0.35	0.37	1.00	0.38	0.28	0.38	0.28	0.37	0.28
hp	0.43	0.42	0.34	0.43	0.53	0.38	1.00	0.48	0.49	0.45	0.47	0.42
ibm	0.43	0.45	0.39	0.45	0.53	0.28	0.48	1.00	0.6	0.54	0.47	0.44
msft	0.48	0.53	0.37	0.46	0.62	0.38	0.49	0.6	1.00	0.59	0.45	0.49
orcl	0.41	0.45	0.32	0.38	0.51	0.28	0.45	0.54	0.59	1.00	0.52	0.47
sunw	0.42	0.41	0.27	0.44	0.51	0.37	0.47	0.47	0.45	0.52	1.00	0.41
yhoo	0.42	0.5	0.32	0.42	0.5	0.28	0.42	0.44	0.49	0.47	0.41	1.00

The summary of PCA below shows the first principle component explains 47.6% of the total variance. In order to explain over 90% of the variance, 9 out of 12 factors are needed.

TABLE II
SUMMARY OF PCA

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.39	0.94031	0.8784	0.81283	0.79357	0.76007
Proportion of Variance	0.476	0.07368	0.0643	0.05506	0.05248	0.04814
Cumulative Proportion	0.476	0.54972	0.614	0.66908	0.72156	0.7697
	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.7401	0.73416	0.70243	0.65937	0.65073	0.5703
Proportion of Variance	0.04565	0.04492	0.04112	0.03623	0.03529	0.0271
Cumulative Proportion	0.81535	0.86026	0.90138	0.93761	0.9729	1

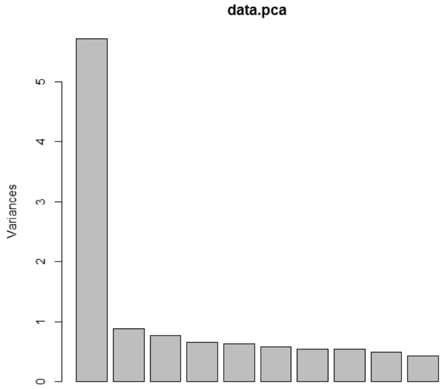

Fig. 1. Variance Explained by Principle Components

The plot shows that from the second principle component to the last principle component, the variance explained by each of them is quiet similar. This is not a desirable phenomenon. It means there is no factor that dominates the variance except for the first principle component. And the first principle component only explained less than half of the variance.

The eigenvalues and eigenvectors of the first 5 principle components are shown as follows.

The eigenvector of the first principle component shows general movement of the market and the eigenvector of the second principle component shows the differences across sectors.

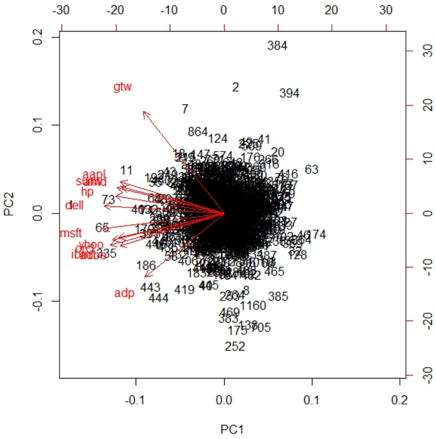

Fig. 2. Bitplot of the first two principle component

TABLE III
EIGENVALUES AND EIGENVECTORS

eigenvalues:	1.1E-03	2.6E-04	2.1E-04	1.8E-04	1.3E-04
eigenvectors:	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	-0.250	0.004	-0.025	-0.032	0.093
[2,]	-0.296	0.234	0.198	-0.254	0.602
[3,]	-0.108	0.049	0.000	-0.035	0.058
[4,]	-0.420	0.027	-0.887	0.017	-0.012
[5,]	-0.240	0.054	0.058	-0.001	0.101
[6,]	-0.335	-0.892	0.155	-0.212	-0.025
[7,]	-0.245	-0.004	0.060	0.029	0.145
[8,]	-0.160	0.072	0.027	0.027	0.095
[9,]	-0.190	0.049	0.060	-0.044	0.148
[10,]	-0.295	0.180	0.228	0.139	0.194
[11,]	-0.392	0.024	0.222	0.762	-0.297
[12,]	-0.360	0.322	0.194	-0.533	-0.661

IV. CONCLUSION

Given the log return of 12 stocks, principle component analysis is performed to reduce dimensionality. Result shows the first 9 components can explain over 90% of the total variance. However, there is no significant difference between the variance explained by each principle component except for the first one, indicating that the collearity between the stock prices might not be large enough to perform principle component analysis.

REFERENCES

[1] Fama, E. F.; French, K. R. (1993). ”Common risk factors in the returns on stocks and bonds”. Journal of Financial Economics 33: 3. doi:10.1016/0304-405X(93)90023-5. CiteSeerX: 10.1.1.139.5892.

[2] Carhart, M. M. (1997). ”On Persistence in Mutual Fund Performance”. The Journal of Finance 52: 5782. doi:10.1111/j.1540-6261.1997.tb03808.x. JSTOR 2329556.

[3] Fama, E. F.; French, K. R. (2015). ”A Five-Factor Asset Pricing Model”. Journal of Financial Economics 116: 122.

[4] Shijie Deng, Design and Implementation of Systems to Support Computational Finance. Final Project Topic 2. 2015Fall.

APPENDIX A
R CODE

```
   # pca
data.full < − read.csv(”d_logret_12stocks.txt”, header = T, sep = ’\t ’)
names(data.full)[names(data.full)==”X.”] < − ”DATE”
data < − data.full
data[1] < − NULL
head(data)
round(cor(data), 2)

data.pca < − prcomp(data, center=TRUE, scale.=TRUE)
summary(data.pca)
biplot(data.pca)

print(eigen(cov(data)))
```