Factor Analysis on Stock Returns

ISYE/MATH 6783 - Assignment3

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. Factor analysis is an extension of Principle Component Analysis. It reduces dimensionality by checking if the hypothesis holds for factor loadings. In this report, factor analysis is performed on the excess log return of ten stocks. According to the result, 3 factors are enough to get a p value greater than the threshold. Then, a modified CAPM model is tested and the optimal value of parameters in the model is estimated via the least squares criterion.**

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. Last time, Principle Component Analysis was used for dimensionality deduction. Principle Component Analysis determines the number of factors based on eigenvalues and variance. PCA and factor analysis are similar but they differs in the following ways[5]. Firstly, PCA results in principal components that account for a maximal amount of variance for observed variables while factor analysis account for common variance in the data. Secondly, PCA inserts ones on the diagonals of the correlation matrix while factor analysis adjusts the diagonals of the correlation matrix with the unique factors. Also, PCA minimizes the sum of squared perpendicular distance to the component axis. factor analysis estimates factors which influence responses on observed variables. Another difference is the component scores in PCA represent a linear combination of the observed variables weighted by eigenvectors while the observed variables in factor analysis are linear combinations of the underlying and unique factors. The last differenct mentioned in Suhr's paper was the components in PCA yielded are uninterpretable while in factor analysis, the underlying constructs can be labeled and readily interpreted, given an accurate model specification.

The rest of this report is organized as follows. Section 2 will introduce the process and result of performing factor analysis on the given dataset. A model validation will be done in Section 4 followed by conclusions in Section 5.

II. Process and Result of Factor Analysis

A. Data Manipulation

Two datasets are given. One contains monthly returns of the S&P500 and the rates of the 3-month U. S. Treasury bill from January 1994 to December 2006. The other contains the monthly log returns of ten stocks. And the excess return is defined as below.

$$excess\_return = stock\_return - risk\_free\_rate$$

where the risk free rate is approximated by the rates of 3-month treasury bills. Also, since the units and time of these rates differs, all data is transformed in to monthly log returns before factor analysis is performed.

B. Factor Analysis

Given a dataset with k factors and n observations, the hypothesis in factor analysis is that t factors are sufficient to explain the variation in the dataset. If the p value is greater than 0.05 means there is a high probability that the hypothesis would hold. So the goal is finding the smallest t that has p value greater than 0.05.

The p values of each number of factors without rotation is shown in Table I.

TABLE I
Factor Loadings without Rotation

Number of Factors	1	2	3	4	5
p value	7.28e-6	0.00219	0.0548	0.357	0.684

The result shows the minimum number of factors for explaining the variation is three. The the result when number of factors equals to three is shown below.

With rotation maximizing the varimax criterion, the result with 3 factors is shown as below.

However, the scree plot indicates a different number of factors.

According to the scree plot, the optimal number of factors should be 2.

```
Call:
factanal(x = stock_ex, factors = 3, rotation = "none")

Uniquenesses:
 AAPL  ADBE   ADP   AMD  DELL   GTW    HP   IBM  MSFT  ORCL
0.566 0.720 0.838 0.535 0.383 0.526 0.835 0.005 0.264 0.789

Loadings:
     Factor1 Factor2 Factor3
AAPL  0.377   0.441   0.312
ADBE  0.180   0.400   0.295
ADP   0.382   0.115
AMD   0.493   0.333   0.332
DELL  0.480   0.622
GTW   0.338   0.568   0.193
HP    0.261           0.305
IBM   0.997
MSFT  0.528   0.539  -0.407
ORCL  0.352   0.286

               Factor1 Factor2 Factor3
SS loadings      2.375   1.563   0.600
Proportion Var   0.237   0.156   0.060
Cumulative Var   0.237   0.394   0.454

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 28.5 on 18 degrees of freedom.
The p-value is 0.0548
```

Fig. 1. Factor Analysis with 3 Factors

```
Call:
factanal(x = stock_ex, factors = 3, rotation = "varimax")

Uniquenesses:
 AAPL  ADBE   ADP   AMD  DELL   GTW    HP   IBM  MSFT  ORCL
0.566 0.720 0.838 0.535 0.383 0.526 0.835 0.005 0.264 0.789

Loadings:
     Factor1 Factor2 Factor3
AAPL  0.575   0.229   0.224
ADBE  0.506   0.142
ADP   0.175   0.177   0.317
AMD   0.539   0.185   0.374
DELL  0.451   0.609   0.209
GTW   0.561   0.378   0.127
HP    0.308           0.257
IBM   0.143   0.346   0.925
MSFT  0.122   0.815   0.237
ORCL  0.291   0.272   0.229

               Factor1 Factor2 Factor3
SS loadings      1.641   1.514   1.383
Proportion Var   0.164   0.151   0.138
Cumulative Var   0.164   0.315   0.454

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 28.5 on 18 degrees of freedom.
The p-value is 0.0548
```

Fig. 2. Rotated Factor Analysis with 3 Factors

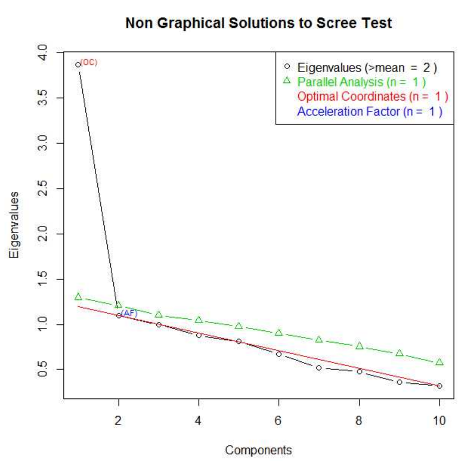

Fig. 3. Scree Test

III. Model validation

The CAPM model describes the relationship between excess stock return and excess market return. However, the β of one stock may change over time. In order to model the change of β, the following modified CAPM model is proposed.

$$r_t^* = \beta_1 1_{\{t<t_0\}} r_M^* + \beta_2 1_{\{t>t_0\}} r_M^* + \epsilon_t$$

where $r_t^* = r_t - r_f$ and $r_M^* = r_M - r_f$ are the excess returns of the stock and the S&P500 index.

A. Test of t_0

In order to validate this model, t_0 is set to February 2001 for each stock and the null hypothesis is $\beta_1 = \beta_2$.

Linear regression is used to test the null hypothesis. The regression model is

$$r_t^* = \alpha_1 1_{\{t<t_0\}} r_M^* + \alpha_2 * r_M^*$$

If α_1 is 0, then it means $\beta_1 = \beta_2$. Otherwise it suggests $\beta_1 \neq \beta_2$.

The p-value of α_1 of each stock is listed in the following table. For ADP and AMD $\beta_1 \neq \beta_2$ and for the others, $\beta_1 = \beta_2$.

TABLE II
P VALUES

Stock	AAPL	ADBE	ADP	AMD	DELL
p value	0.87616	0.148	0.0704	0.0174	0.05763
Stock	GTW	HP	IBM	MSFT	ORCL
p value	0.111	0.85387	0.174	0.11	0.937

B. Optimal t_0

To find the optimal t_0 that minimizes the square error, every possiple value of t_0 is tested and the optimal t_0 for each stock is shown in TableIII.

TABLE III
OPTIMAL t_0 FOR EACH STOCK

Stock	AAPL	ADBE	ADP	AMD	DELL
t_0	60	148	93	81	87
Stock	GTW	HP	IBM	MSFT	ORCL
t_0	77	113	74	110	51

IV. Conclusion

Given the log return of ten stocks, market return and risk free rate, factor analysis is performed to reduce dimensionality. Result shows the at least three factor are needed to explain the viriation of the data. Then, a modified model decribing the relationship between stock return and market return is tested and it is valid for eight out of ten stocks.

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.

[5] Suhr, Diane (2009). "Principal component analysis vs. exploratory factor analysis". SUGI 30 Proceedings. Retrieved 5 April 2012.

Appendix A
R code

```
library(nFactors)
# compute market excess log return
mkt ¡- read.csv(”m_sp500ret_3mtcm.txt”, header = T, sep = ’t’, skip = 1)
mkt$log_rf < − log(mkt$X3mTCM/100 + 1)/12
mkt$ex_logret < − log(mkt$sp500+1) - mkt$log_rf
mkt < − subset(mkt, select=c(”Date”, ”ex_logret”, ”log_rf”))
# compute stock excess log return
stock < − read.csv(”m_logret_10stocks.txt”, header = T, sep = ’t’)
stock$Date < − NULL
stock = na.omit(stock)
stock_ex = stock - mkt$log_rf
# factor analysis without rotation
for (i in 1:10){
fa < − factanal(stock_ex, i, rotation=”none”)
print(fa)
}
# factor analysis with rotation
for (i in 1:10){
fa.rotate < − factanal(stock_ex, i, rotation=”varimax”)
print(fa.rotate)
}
fa < − factanal(stock_ex, 3, rotation=”none”)
print(fa)
fa.rotate < − factanal(stock_ex, 3, rotation=”varimax”)
print(fa.rotate)
# scree test
ev < − eigen(cor(stock_ex))
# get eigenvalues
ap < − parallel(subject=nrow(stock_ex),var=ncol(stock_ex), rep=100,cent=.05)
nS < − nScree(x=ev$values, aparallel=ap$eigen$qevpea)
plotnScree(nS)
# model validation
idx < − seq(1,156)
idx.t0 < − 86
data.before < − mkt$ex_logret*I(idx ¡ idx.t0)
for (i in 1:10){
data.test < − cbind(stock_ex[i], data.before, mkt$ex_logret)
linreg < − lm(data.test[,1]    data.test[,2] + data.test[,3])
print(summary(linreg))
}
# find optimal t0
for (i in 1:10){
minError < − 999999
for (j in 1:156){
data.before < − mkt$ex_logret*I(idx ¡ j)
data.after < − mkt$ex_logret*I(idx ¿= j)
data.test < − cbind(stock_ex[i], data.before, data.after)
linreg < − lm(data.test[,1]    data.test[,2] + data.test[,3])
error < − sum((linreg$residuals)2̂)
if (error ¡ minError){
minError < − error
t0 < − j
}
}
```

```
print(t0)
}
```