A study of the log returns of ten stocks with factor analysis

Junfeng Li[1,2]
(1.School of Mathematics and Statistics；2.Student ID:1120132819)

6th.June.2016

Abstract

The purpose of this paper is to study the log returns of ten stocks which include MSFT, DELL, IBM and other seven stocks. By correlation analysis, we can find that these stocks are correlative even though their correlation isn't strong enough. This may means that they have some repeating information or some factors in common. So, we use the method of factor analysis to search and choose common information from these possibly correlated variables(their common factors). And the result in SPSS shows that we extract two or three principal and common factors using different methods of extraction. In model test, we use the data of DELL as an example, and we reject that they have the same coefficient by t-test.

Key words:factor analysis;correlation analysis;SPSS

1 Introduction

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved (underlying) variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Factor analysis originated in psychometrics and is used in behavioral sciences, social sciences, marketing, product management, operations research, and other fields that deal with data sets where there are large numbers of observed variables that are thought to reflect a smaller number of underlying/latent variables.

Factor analysis is related to principal component analysis (PCA), but the two are not identical. There has been significant controversy in the field over differences between the two techniques. Clearly though, PCA is a more basic version of exploratory factor analysis that was developed in the early days prior to the advent of high-speed computers. From the point of view of exploratory analysis, the eigenvalues of PCA are inflated component loadings, i.e., contaminated with error variance.

2 Mathematical Background

For the mathematical definition of factor analysis, suppose we have a set of p observable random variables, $x_1, \ldots, x_p$ with means $\mu_1, \ldots, \mu_p$.

Suppose for some unknown constants l_{ij} and k unobserved random variables F_j , where $i \in 1, \ldots, p$ and $j \in 1, \ldots, k$, where $k < p$, we have

$$x_i - \mu_i = l_{i1}F_1 + \cdots + l_{ik}F_k + \varepsilon_i.$$

Here, the ε_i are independently distributed error terms with zero mean and finite variance, which may not be the same for all i. Let $\mathrm{Var}(\varepsilon_i) = \psi_i$, so that we have

$$\mathrm{Cov}(\varepsilon) = \mathrm{Diag}(\psi_1, \ldots, \psi_p) = \Psi \text{ and } \mathrm{E}(\varepsilon) = 0.$$

In matrix terms, we have

$$x - \mu = LF + \varepsilon.$$

If we have n observations, then we will have the dimensions $x_{p\times n}$, $L_{p\times k}$, and $F_{k\times n}$. Each column of x and F denote values for one particular observation, and matrix L does not vary across observations.

Also we will impose the following assumptions on F:

1. F and ε are independent.
2. $\mathrm{E}(F) = 0$.
3. $\mathrm{Cov}(F) = I$(to make sure that the factors are uncorrelated).

Any solution of the above set of equations following the constraints for F is defined as the factors, and L as the loading matrix.

Suppose $\mathrm{Cov}(x-\mu)=\Sigma$. Then note that from the conditions just imposed on F , we have

$$\mathrm{Cov}(x-\mu)=\mathrm{Cov}(LF+\varepsilon)$$

or

$$\Sigma = L\mathrm{Cov}(F)L^T+\mathrm{Cov}(\varepsilon)$$

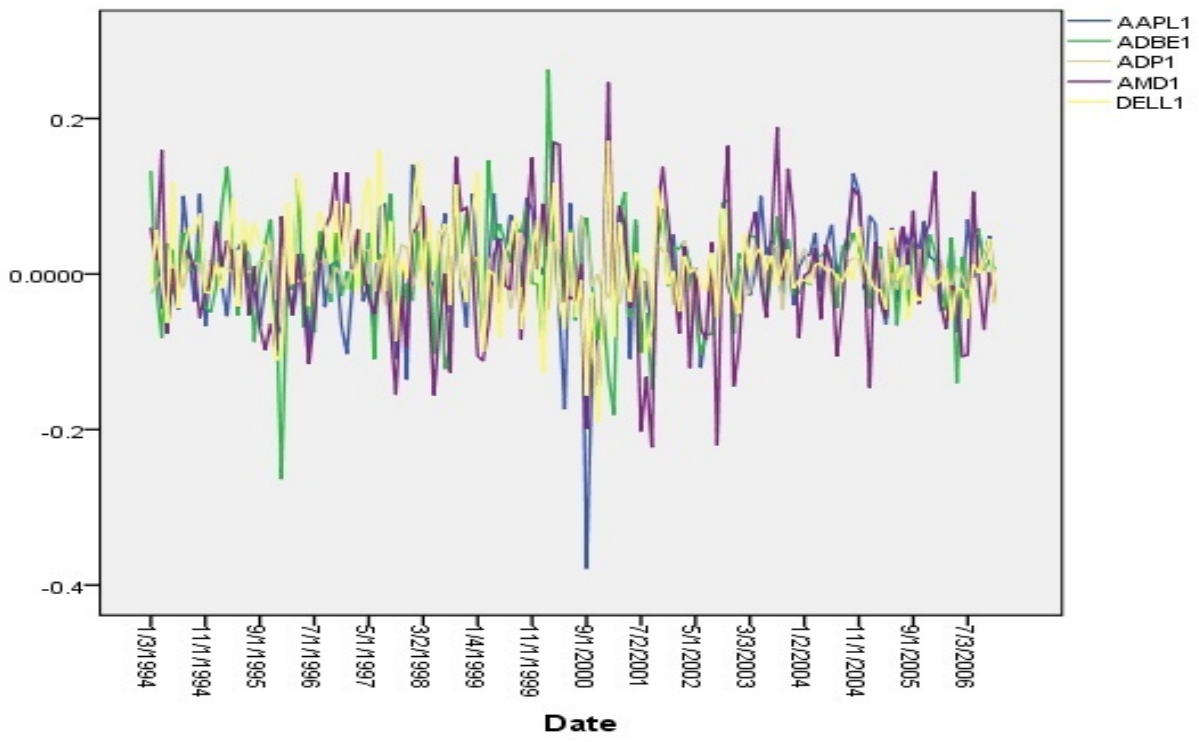
or

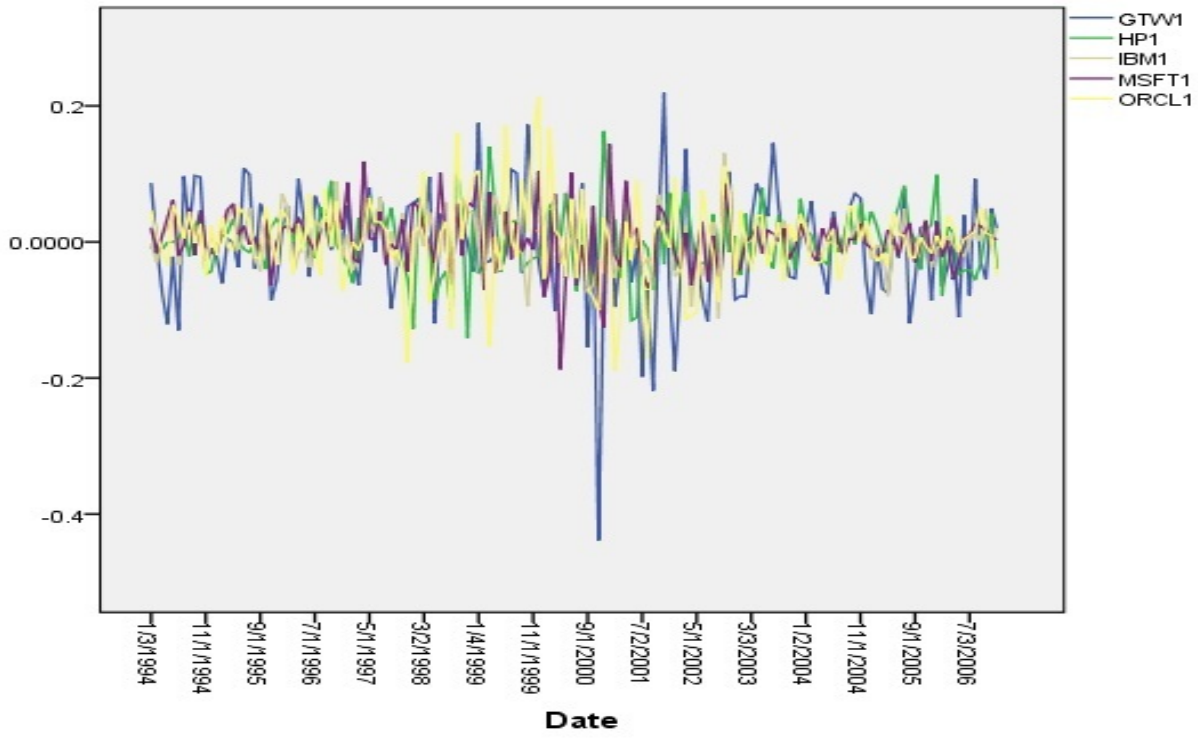
$$\Sigma = LL^T+\Psi.$$

Note that for any orthogonal matrix Q, if we set $L=LQ$ and $F=Q^TF$, the criteria for being factors and factor loadings still hold. Hence a set of factors and factor loadings is unique only up to orthogonal transformation.

3 factor analysis

First of all,we use spss to get the time series plot of the log returns of the ten stocks to see if we can find some information from it:





From the time series plot above, we can find that the ten stocks may have different mean and variance. But, there must be some repeating information among them because they looks similar and their mean are so close. Then, we compare their mean and variance using SPSS.

	AAPL1	ADBE1	ADP1	AMD1	DELL1
mean	0.0037	0.0045	0.0007	-0.0009	0.0087
variance	0.005	0.004	0.001	0.007	0.004

	GTW1	HP1	IBM1	MSFT1	ORCL1
mean	-0.0057	0.0018	0.0025	0.0041	0.0037
variance	0.007	0.002	0.001	0.002	0.004

As we can see in the table, some stocks have negative mean of their log returns, which means they might be bad object to invest. And, the variances of two stocks(AMD,GTW) are larger than others. This can be thought that these two stocks are unstable and more risky.

Secondly, we compute the correlation matrix of the ten stocks.From the correlation matrix, we can know that the correlation between every two stocks range from 5% to 60%. This means they have repeating information, so we think that maybe we can use the method of

factor analysis to decrease the dimension of data. Then, we do the Bartlett's test of sphericity to test the hypothesis that the variables are uncorrelated in the population. And the result from SPSS show that P value is almost 0 which means the ten stocks are correlated. Also, the Kaiser-Meyer-Olkin(KMO) measure of sampling adequacy test that used to examine the appropriateness of factor analysis shows it's value is up to 0.8. This indicates appropriateness to use factor analysis here.

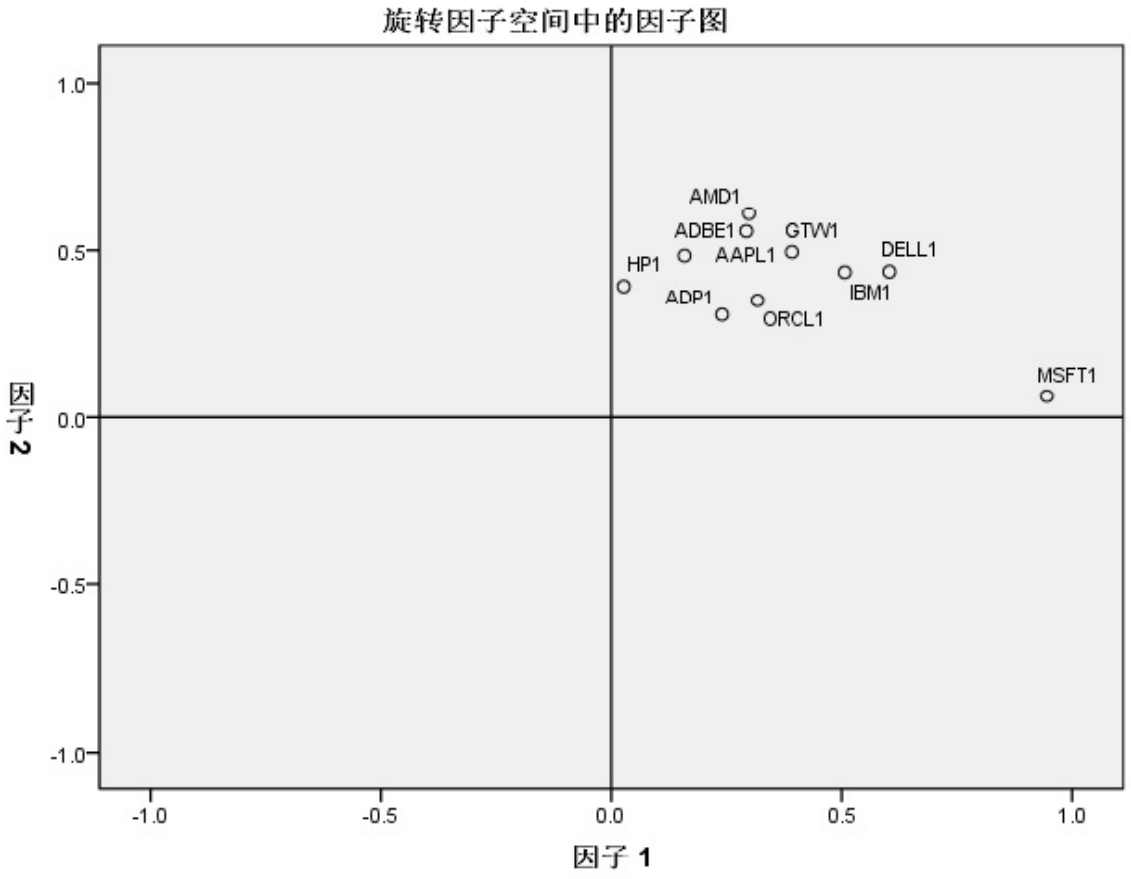
Now, we use the method of principal axis factoring to get the common factors.

	Factor1	Factor2	Initial Communalities	Extractive Communalities
AAPL1	0.597	0.199	0.378	0.396
ADBE1	0.450	0.239	0.247	0.259
ADP1	0.386	0.055	0.236	0.152
AMD1	0.638	0.233	0.393	0.461
DELL1	0.738	-0.103	0.527	0.555
GTW1	0.626	0.086	0.402	0.399
HP1	0.292	0.265	0.156	0.155
IBM1	0.667	-0.037	0.477	0.446
MSFT1	0.726	-0.608	0.465	0.898
ORCL1	0.472	0.033	0.222	0.222

From the table above, we know that there are two common factors. The factor loadings are shown. We will try to improve the loadings later by allowing rotations. And we can see the extractive communalities which is the Amount of variance a variable shares with all the other variables in the table. This is the proportion of variance explained by the common factors.

There are five main method to do rotation in SPSS: Varimax, Direct Oblimin, Quartimax, Equamax, Promax. And we just choose the method of Varimax. The new common factors after rotation are shown below.

	Factor1	score1	Factor2	score2
AAPL1	0.293	-0.001	0.557	0.210
ADBE1	0.159	-0.049	0.484	0.199
ADP1	0.240	0.005	0.307	0.092
AMD1	0.299	-0.008	0.610	0.277
DELL1	0.603	0.062	0.437	0.194
GTW1	0.392	0.016	0.495	0.173
HP1	0.027	-0.016	0.393	0.124
IBM1	0.507	0.016	0.435	0.179
MSFT1	0.945	0.092	0.064	-0.465
ORCL1	0.317	0.010	0.350	0.102



In factor analysis, a factor can be interpreted in terms of the variables that load high on it.Another useful aid in interpretation is to plot the variables, using the factor loadings as coordinates. Variables at the end of an axis are those that have high loadings on only that factor, and hence describe the factor.

4 model test

Consider the model:

$$r_t^* = \beta_1 l_{t<t_0} r_M^* + \beta_2 l_{t\geq t_0} r_M^* + \epsilon_t$$

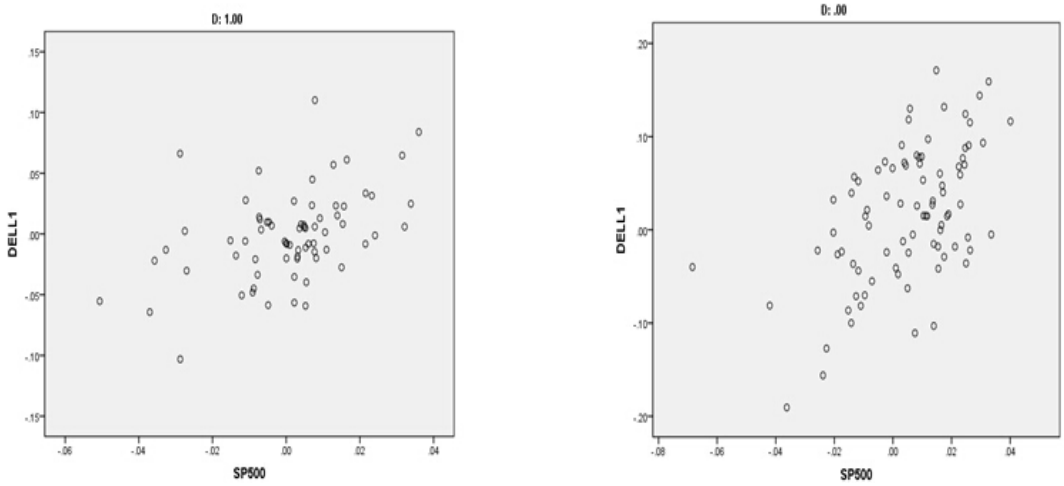
where $r_t^* = r_t - r_f$ and $r_M = r_M - r_f$ are the excess returns of the stock and the S&P500 index. The model suggests that the β may vary over the time. Taking February 2001 as the month t_0 , then we test for each stock the null hypothesis that $\beta_1 = \beta_2$.

Firstly, we transfer the model to:

$$r_t^* = \beta_1 r_M^* + \beta_3 D + \epsilon_t$$

where $D = 0$ if $t < t_0$ and $D = 1$ if $t \geq t_0$. Then, we test wether $\beta_3 = 0$. It's equal to the initial null hypothesis that $\beta_1 = \beta_2$.

Now, we use the data of DELL as an example. bellow are two scatter plot between DELL and S&P500 in different time. We can see that they may have different slope.



However, the T-test of coefficient in SPSS indicates that we need to reject the initial hypothesis $\beta_3 = 0$ which means $\beta_1 = \beta_2$.(more information in appendix)

5 conclusion

We use the method of factor analysis to convert these ten possibly correlated variables into two common factors. This paper shows the log returns of ten stocks which include MSFT, DELL,

IBM and other seven stocks. By correlation analysis, we found that these stocks are correlative even though their correlation isn't strong enough. This means that they have some repeating information or some factors in common. So, we use the method of factor analysis to search and choose common information from these possibly correlated variables(their common factors). And the result shows that we extract two common factors using the method of Varimax. In model test, we use the data of DELL as an example, and we reject that they have the same coefficient by t-test.

Reference:

1. Principal Component Analysis , Wikipedia

appendix:

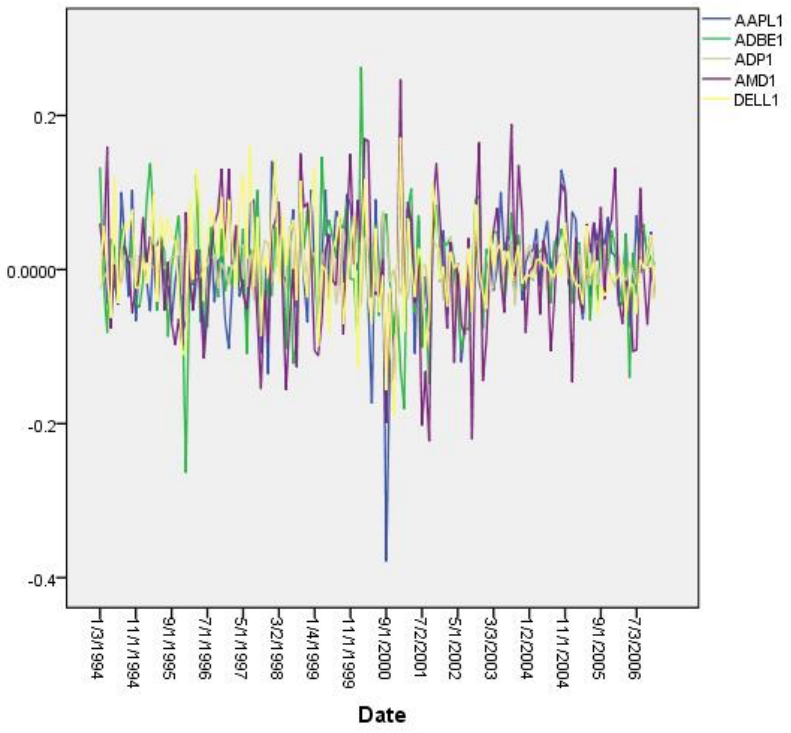
序列图

模型描述

模型名称		MOD_1
序列或顺序	1	AAPL1
	2	ADBE1
	3	ADP1
	4	AMD1
	5	DELL1
转换		无
非季节性差分		0
季节性差分		0
季节性期间的长度		无周期性
水平轴标签		Date
干预开始		无
对于每个观测值		未连接的值

个案处理摘要

		AAPL1	ADBE1	ADP1	AMD1	DELL1
序列或顺序长度		159	159	159	159	159
图中的缺失值数	用户缺失	0	0	0	0	0
	系统缺失	3	3	3	3	3



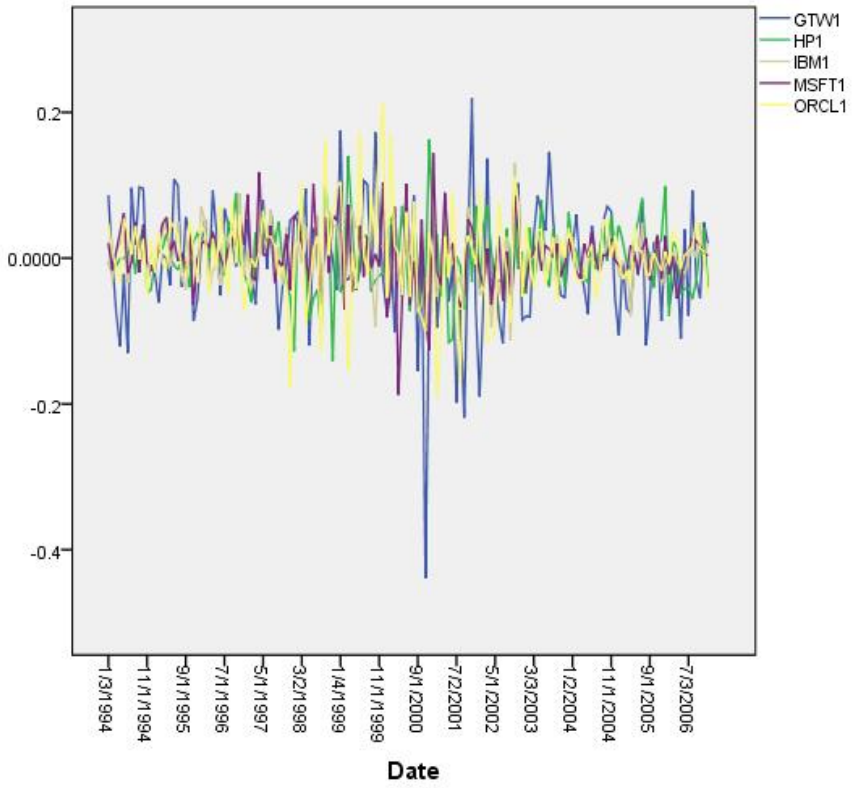
序列图

模型描述

模型名称		MOD_2
序列或顺序	1	GTW1
	2	HP1
	3	IBM1
	4	MSFT1
	5	ORCL1
转换		无
非季节性差分		0
季节性差分		0
季节性期间的长度		无周期性
水平轴标签		Date
干预开始		无
对于每个观测值		未连接的值

个案处理摘要

		GTW1	HP1	IBM1	MSFT1	ORCL1
序列或顺序长度		159	159	159	159	159
图中的缺失值数	用户缺失	0	0	0	0	0
	系统缺失	3	3	3	3	3



因子分析

KMO 和 Bartlett 的检验

取样足够度的 Kaiser-Meyer-Olkin 度量。		.823
Bartlett 的球形度检验	近似卡方	428.843
	df	45
	Sig.	.000

相关矩阵

		AAPL1	ADBE1	ADP1	AMD1	DELL1	GTW1	HP1	IBM1	MSFT1	ORCL1
Sig.（单侧）	AAPL1		.000	.185	.000	.000	.000	.002	.000	.000	.000
	ADBE1	.000		.000	.000	.000	.000	.003	.015	.005	.000
	ADP1	.185	.000		.004	.003	.001	.003	.000	.001	.004
	AMD1	.000	.000	.004		.000	.000	.003	.000	.000	.000
	DELL1	.000	.000	.003	.000		.000	.009	.000	.000	.000
	GTW1	.000	.000	.001	.000	.000		.067	.000	.000	.003
	HP1	.002	.003	.003	.003	.009	.067		.001	.311	.102
	IBM1	.000	.015	.000	.000	.000	.000	.001		.000	.000
	MSFT1	.000	.005	.001	.000	.000	.000	.311	.000		.000
	ORCL1	.000	.000	.004	.000	.000	.003	.102	.000	.000	

公因子方差

	初始	提取
AAPL1	.378	.396
ADBE1	.247	.259
ADP1	.236	.152
AMD1	.393	.461
DELL1	.527	.555
GTW1	.402	.399
HP1	.156	.155
IBM1	.477	.446
MSFT1	.465	.898
ORCL1	.222	.223

解释的总方差

因子	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	3.865	38.654	38.654	3.329	33.293	33.293	2.027	20.271	20.271
2	1.098	10.985	49.639	.614	6.143	39.437	1.917	19.165	39.437
3	.997	9.974	59.613						
4	.879	8.786	68.399						
5	.811	8.110	76.509						
6	.670	6.698	83.207						
7	.518	5.180	88.387						
8	.478	4.784	93.171						
9	.359	3.590	96.761						
10	.324	3.239	100.000						

因子矩阵

	因子	
	1	2
AAPL1	.597	.199
ADBE1	.450	.239
ADP1	.386	.055
AMD1	.638	.233
DELL1	.738	-.103
GTW1	.626	.086
HP1	.292	.265
IBM1	.667	-.037
MSFT1	.726	-.608
ORCL1	.472	.033

旋转因子矩阵

	因子	
	1	2
AAPL1	.293	.557
ADBE1	.159	.484
ADP1	.240	.307
AMD1	.299	.610
DELL1	.603	.437
GTW1	.392	.495
HP1	.027	.393
IBM1	.507	.435
MSFT1	.945	.064

ORCL1	.317	.350

因子转换矩阵

因子	1	2
1	.721	.693
2	-.693	.721

回归

输入／移去的变量

模型	输入的变量	移去的变量	方法
1	D, SP500	.	输入

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差	Durbin-Watson
1	.516	.267	.257	.05109	1.938

Anova

模型		平方和	df	均方	F	Sig.
1	回归	.145	2	.073	27.817	.000
	残差	.399	153	.003		
	总计	.545	155			

系数

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	.009	.006		1.532	.128
	SP500	1.639	.230	.497	7.134	.000
	D	-.011	.008	-.094	-1.352	.178

残差统计量

	极小值	极大值	均值	标准 偏差	N
预测值	-.1036	.0743	.0087	.03061	156
残差	-.13973	.13828	.00000	.05076	156
标准 预测值	-3.667	2.146	.000	1.000	156
标准 残差	-2.735	2.706	.000	.994	156