A study of the log returns of twelve stocks with principal component analysis

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

24th.MAY.2016

Abstract

The purpose of this paper is to study the log returns of twelve stocks which include Yahoo, Dell, HP and other nine stocks. By correlation analysis, we can find that these stocks are correlative. This may means that they have some repeating information. So, we use the method of principal component analysis to convert these possibly correlated variables into a set of values of linearly uncorrelated variables(their principal components). And the result in R shows that the first seven principal components can reach almost 90% of total variation when we use covariance matrix to do it. Particularly, the first principal component's proportion is 47%. However, by using correlation matrix, the first seven principal components decrease to 80% of total variation which is also representative.

Key words:principal component analysis;correlation analysis;R

1 Introduction

Principal component analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables.

Also, PCA is the simplest of the true eigenvector-based multivariate analysis. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualised as a set of coordinates in a high-dimensional data space, PCA can supply the user with a lower-dimensional picture, a projection of this object when viewed from its most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

2 Mathematical Background

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on.

Consider a data matrix, X, with column-wise zero empirical mean, where each of the n rows represents a different repetition of the experiment, and each of the p columns gives a particular kind of feature.

Mathematically, the transformation is defined by a set of p-dimensional vectors of weights or loadings $\mathbf{w_{(k)}} = (w_1, \ldots, w_p)_{(k)}$ that map each row vector $\mathbf{x}_{(i)}$ of X to a new vector of principal component scores $\mathbf{t_{(i)}} = (t_1, \ldots, t_k)_{(i)}$, given by $t_{k(i)} = \mathbf{x_{(i)}} \cdot \mathbf{w_{(k)}}$ in such a way that the individual variables of $\mathbf{t}$ considered over the data set successively inherit the maximum possible variance from $\mathbf{X}$, with each loading vector $\mathbf{w}$ constrained to be a unit vector.

The first loading vector $\mathbf{w}_{(1)}$ thus has to satisfy

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w} \right)^2 \right\}$$

Equivalently, writing this in matrix form gives

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{ \|\mathbf{X}\mathbf{w}\|^2 \} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right\}$$

Since $\mathbf{w}_{(1)}$ has been defined to be a unit vector, it equivalently also satisfies

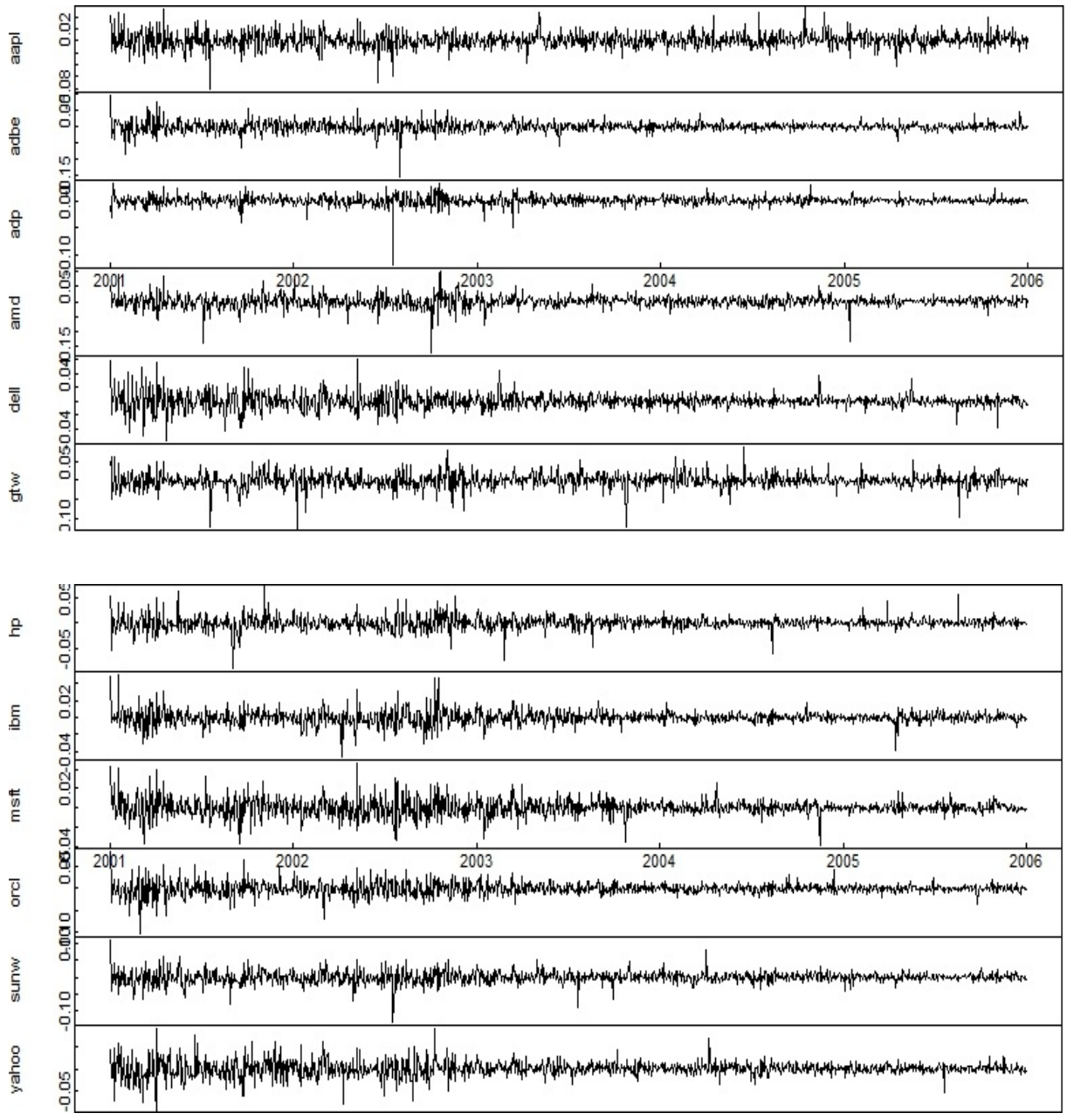
$$\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

The quantity to be maximised can be recognised as a Rayleigh quotient. A standard result for a symmetric matrix such as X^TX is that the quotient's maximum possible value is the largest eigenvalue of the matrix, which occurs when w is the corresponding eigenvector.

With $\mathbf{w}_{(1)}$ found, the first component of a data vector $\mathbf{x}_{(i)}$ can then be given as a score $\mathbf{t}_{1(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}$ in the transformed co-ordinates, or as the corresponding vector in the original variables, $\mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}\mathbf{w}_{(1)}$.

3 principal component analysis

First of all,we use R to get the time series plot of the log returns of the twelve stocks:



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

	aapl	adbe	adp	amd	dell	gtw
mean	7.85e-04	1.59e-04	-1.04e-04	2.61e-04	1.86e-04	-6.67e-04
variance	1.60e-04	2.20e-04	6.51e-05	3.64e-04	1.18e-04	3.50e-04

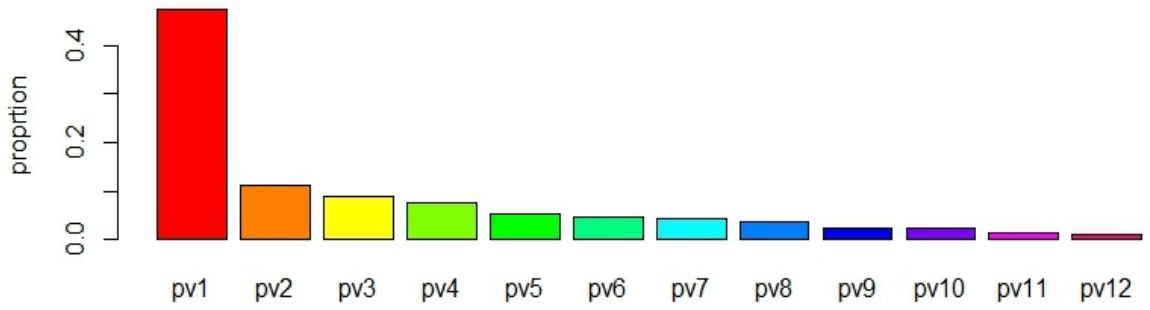
	hp	ibm	msft	orcl	sunw	yhoo
mean	-1.90e-05	-1.08e-05	6.46e-05	-2.67e-04	-6.24e-04	3.54e-04
variance	1.44e-04	6.41e-05	7.68e-05	1.99e-04	3.11e-04	2.87e-04

As we can see in the table, half of the stocks have negative mean of their log returns, which means they might be bad object to invest.And, the variances of three stocks(adp, ibm, msft) are smaller than others. This can be thought that these three stocks are stable and less risky.

Secondly, we compute the covariance matrix and correlation matrix of the twelve stocks.From the correlation matrix, we can know that the correlation between every two stocks range from 30% to 60%. This means they have repeating information, so we think that maybe we can use the method of principal component analysis to decrease the dimension of data. There are two ways to do this, using covariance matrix or correlation matrix. Following is the result from R.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	0.03343	0.01622	0.01442	0.01335	0.01127	0.01052
Proportion of Variance	0.47365	0.11150	0.08812	0.07550	0.05383	0.04693
Cumulative Proportion	0.47365	0.58515	0.67326	0.74876	0.80259	0.84952

	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.01001	0.00915	0.00754	0.00727	0.00594	0.00513
Proportion of Variance	0.04250	0.03546	0.02407	0.02237	0.01494	0.01114
Cumulative Proportion	0.89202	0.92749	0.95156	0.97392	0.98886	1.00000



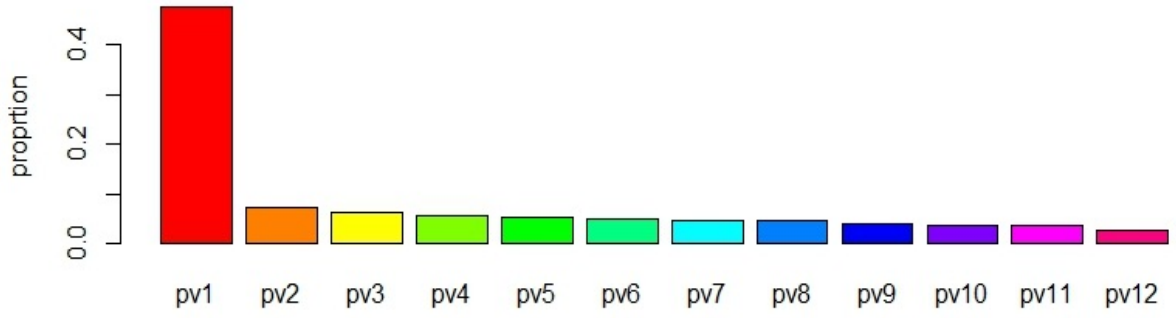
The result in R shows that the first five principal components can reach almost 80% of total variation when we use covariance matrix to do it. Furthermore, the first seven principal components' proportion is up to 90%. Particularly, the first principal component's proportion is 47% which is so near to a half. This can help us reduce the dimension of data to five or seven, according to our requirement.

Things are different when we use correlation matrix to do principal components analysis. Each principal component's proportion of variance is showed below.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.39000	0.94031	0.87840	0.81283	0.47600	0.07368
Proportion of Variance	0.47600	0.07368	0.06430	0.05506	0.05248	0.04814
Cumulative Proportion	0.47600	0.54972	0.61400	0.66908	0.72156	0.76970

	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.74010	0.73416	0.70243	0.65937	0.65073	0.57030
Proportion of Variance	0.04565	0.04492	0.04112	0.03623	0.03529	0.02710
Cumulative Proportion	0.81535	0.86026	0.90138	0.93761	0.97290	1.00000

Following is a bar plot of these twelve principal components which can give us a more intuitive impression.



From the table and bar plot above, we can find that the first five principal components' cumulative proportion is about 67% of total variation by using correlation matrix. And, the first seven principal components' cumulative proportion is 82% which is less than it using covariance matrix. That means we can use the data of less dimension to simplify our analysis, but it still has most of the original information.

On the other hand, we find that there are some difference with different matrix of PCA. Compared with the correlation covariance matrix of principal component analysis, principal component analysis of covariance matrix is more efficient when a single index variance are crucial for research purposes. However, some data involving different measurement, and the variables are not comparable with variance. So, for this kind of data, principal component analysis with covariance matrix is improper. As we know, the correlation coefficient matrix is a standard matrix. Through the standardization of random variables, it only retains the correlation between variables which don't need yo worry about different measurement of the data.

4 conclusion

We use the method of principal component analysis to convert these twelve possibly correlated variables into their principal components. The result shows that the first seven principal components can reach almost 90% of total variation when we use covariance matrix to do it. Particularly, the first principal component's proportion is 47%. However, by using correlation matrix, the first seven principal components decrease to 80% of total variation which is also representative. From the study of the analysis with two different matrix, we have learned that we should analyze the data first before we do PCA, which can help us know if we need to think about the data's measurement.

Reference:

1. Comparative Research Main Components Analysis with Covariance Matrix and Correlation Matrix, ZHU Xiao-feng, CHINA SPORT SCIENCE AND TECHNOLOGY, Vol.41,No.3,1342136,2005.

2. Principal Component Analysis , Wikipedia

Appendix:

```
> D<-read.table("E:/课件/统计计算/统计计算作业/2/d_logret_12stocks.txt",header=T)
> time<-as.Date(as.character(D[,1]),format="%Y%m%d", origin="2001-01-03")
> par(mfrow=c(6,1),mai=c(0,0.5,0,0.5),tck=0.05,mgp=c(3,0.1,0))
> plot(time,D[,2],type="l",ylab="aapl",xaxt="n")
> plot(time,D[,3],type="l",ylab="adbe",xaxt="n")
> plot(time,D[,4],type="l",ylab="adp")
> plot(time,D[,5],type="l",ylab="amd",xaxt="n")
> plot(time,D[,6],type="l",ylab="dell",xaxt="n")
> plot(time,D[,7],type="l",ylab="gtw",xaxt="n")
> par(mfrow=c(6,1),mai=c(0,0.5,0,0.5),tck=0.05,mgp=c(3,0.1,0))
> plot(time,D[,8],type="l",ylab="hp",xaxt="n")
> plot(time,D[,9],type="l",ylab="ibm",xaxt="n")
> plot(time,D[,10],type="l",ylab="msft")
> plot(time,D[,11],type="l",ylab="orcl",xaxt="n")
> plot(time,D[,12],type="l",ylab="sunw",xaxt="n")
> plot(time,D[,13],type="l",ylab="yahoo",xaxt="n")
> apply(D[,-1],2,mean)
              aapl              adbe               adp               amd
 7.850530e-04  1.586685e-04 -1.043985e-04  2.614451e-04
              dell               gtw                hp               ibm
 1.859433e-04 -6.670242e-04 -1.904704e-05 -1.082917e-05
              msft              orcl              sunw              yhoo
 6.475075e-05 -2.665154e-04 -6.241187e-04  3.538089e-04
> cov(D[,-1])
                aapl              adbe               adp               amd
aapl 1.602972e-04 7.261495e-05 2.911049e-05 1.083805e-04
adbe 7.261495e-05 2.196976e-04 3.650095e-05 1.081695e-04
adp  2.911049e-05 3.650095e-05 6.514283e-05 4.782619e-05
amd  1.083805e-04 1.081695e-04 4.782619e-05 3.644885e-04
dell 7.088990e-05 7.842859e-05 2.775145e-05 9.768945e-05
gtw  8.390727e-05 7.381208e-05 2.944381e-05 1.239572e-04
hp   6.558035e-05 7.566500e-05 3.320855e-05 9.837326e-05
ibm  4.355035e-05 5.349149e-05 2.503108e-05 6.807450e-05
msft 5.314971e-05 6.829328e-05 2.585040e-05 7.646203e-05
orcl 7.281651e-05 9.488668e-05 3.635144e-05 1.037446e-04
sunw 9.477767e-05 1.059337e-04 3.884307e-05 1.479069e-04
yhoo 9.039319e-05 1.248751e-04 4.403850e-05 1.363715e-04
                dell               gtw                hp               ibm
aapl 7.088990e-05 8.390727e-05 6.558035e-05 4.355035e-05
adbe 7.842859e-05 7.381208e-05 7.566500e-05 5.349149e-05
adp  2.775145e-05 2.944381e-05 3.320855e-05 2.503108e-05
```

```
amd    9.768945e-05 1.239572e-04 9.837326e-05 6.807450e-05
dell 1.184557e-04 7.489606e-05 6.904001e-05 4.636074e-05
gtw    7.489606e-05 3.495261e-04 8.534350e-05 4.227032e-05
hp     6.904001e-05 8.534350e-05 1.444505e-04 4.591866e-05
ibm    4.636074e-05 4.227032e-05 4.591866e-05 6.411143e-05
msft 5.907876e-05 6.148801e-05 5.197987e-05 4.229886e-05
orcl 7.837947e-05 7.486131e-05 7.565041e-05 6.095342e-05
sunw 9.808163e-05 1.214318e-04 9.999179e-05 6.653126e-05
yhoo 9.193902e-05 8.838912e-05 8.594824e-05 5.975147e-05
                msft          orcl          sunw          yhoo
aapl 5.314971e-05 7.281651e-05 9.477767e-05 9.039319e-05
adbe 6.829328e-05 9.488668e-05 1.059337e-04 1.248751e-04
adp    2.585040e-05 3.635144e-05 3.884307e-05 4.403850e-05
amd    7.646203e-05 1.037446e-04 1.479069e-04 1.363715e-04
dell 5.907876e-05 7.837947e-05 9.808163e-05 9.193902e-05
gtw    6.148801e-05 7.486131e-05 1.214318e-04 8.838912e-05
hp     5.197987e-05 7.565041e-05 9.999179e-05 8.594824e-05
ibm    4.229886e-05 6.095342e-05 6.653126e-05 5.975147e-05
msft 7.677275e-05 7.261635e-05 7.024557e-05 7.322809e-05
orcl 7.261635e-05 1.993962e-04 1.286252e-04 1.121691e-04
sunw 7.024557e-05 1.286252e-04 3.105267e-04 1.234453e-04
yhoo 7.322809e-05 1.121691e-04 1.234453e-04 2.868855e-04
> cor(D[,-1])
                aapl          adbe          adp          amd          dell
aapl 1.0000000 0.3869461 0.2848743 0.4483803 0.5144503
adbe 0.3869461 1.0000000 0.3051113 0.3822526 0.4861651
adp    0.2848743 0.3051113 1.0000000 0.3103777 0.3159181
amd    0.4483803 0.3822526 0.3103777 1.0000000 0.4701407
dell 0.5144503 0.4861651 0.3159181 0.4701407 1.0000000
gtw    0.3544842 0.2663637 0.1951286 0.3472881 0.3680791
hp     0.4309742 0.4247397 0.3423396 0.4287219 0.5277927
ibm    0.4295968 0.4507174 0.3873276 0.4453227 0.5319915
msft 0.4791091 0.5258499 0.3655362 0.4570886 0.6195122
orcl 0.4072944 0.4533506 0.3189553 0.3848265 0.5099950
sunw 0.4248089 0.4055758 0.2731057 0.4396397 0.5113994
yhoo 0.4215203 0.4974036 0.3221402 0.4217233 0.4987325
                gtw          hp          ibm          msft          orcl
aapl 0.3544842 0.4309742 0.4295968 0.4791091 0.4072944
adbe 0.2663637 0.4247397 0.4507174 0.5258499 0.4533506
adp    0.1951286 0.3423396 0.3873276 0.3655362 0.3189553
amd    0.3472881 0.4287219 0.4453227 0.4570886 0.3848265
dell 0.3680791 0.5277927 0.5319915 0.6195122 0.5099950
gtw    1.0000000 0.3798141 0.2823761 0.3753592 0.2835692
hp     0.3798141 1.0000000 0.4771577 0.4935965 0.4457520
```

```
ibm    0.2823761 0.4771577 1.0000000 0.6029168 0.5391032
msft 0.3753592 0.4935965 0.6029168 1.0000000 0.5869112
orcl 0.2835692 0.4457520 0.5391032 0.5869112 1.0000000
sunw 0.3685897 0.4721232 0.4715291 0.4549523 0.5169139
yhoo 0.2791289 0.4222048 0.4405819 0.4934236 0.4689866
                sunw           yhoo
aapl 0.4248089 0.4215203
adbe 0.4055758 0.4974036
adp    0.2731057 0.3221402
amd    0.4396397 0.4217233
dell 0.5113994 0.4987325
gtw    0.3685897 0.2791289
hp      0.4721232 0.4222048
ibm    0.4715291 0.4405819
msft 0.4549523 0.4934236
orcl 0.5169139 0.4689866
sunw 1.0000000 0.4135907
yhoo 0.4135907 1.0000000
> a=prcomp(D[,-1],scale=F)
> summary(a)
Importance of components:
                                        PC1         PC2         PC3         PC4
Standard deviation          0.03343 0.01622 0.01442 0.01335
Proportion of Variance 0.47365 0.11150 0.08812 0.07550
Cumulative Proportion    0.47365 0.58515 0.67326 0.74876
                                        PC5         PC6         PC7          PC8
Standard deviation          0.01127 0.01052 0.01001 0.009148
Proportion of Variance 0.05383 0.04693 0.04250 0.035460
Cumulative Proportion    0.80259 0.84952 0.89202 0.927490
                                         PC9         PC10         PC11         PC12
Standard deviation          0.007536 0.007265 0.005937 0.005127
Proportion of Variance 0.024070 0.022370 0.014940 0.011140
Cumulative Proportion    0.951560 0.973920 0.988860 1.000000
> b=prcomp(D[,-1],scale=T)
> summary(b)
Importance of components:
                                       PC1         PC2        PC3         PC4
Standard deviation          2.390 0.94031 0.8784 0.81283
Proportion of Variance 0.476 0.07368 0.0643 0.05506
Cumulative Proportion    0.476 0.54972 0.6140 0.66908
                                        PC5         PC6         PC7         PC8
Standard deviation          0.79357 0.76007 0.74010 0.73416
Proportion of Variance 0.05248 0.04814 0.04565 0.04492
Cumulative Proportion    0.72156 0.76970 0.81535 0.86026
```

```
                                       PC9       PC10       PC11      PC12
Standard deviation          0.70243 0.65937 0.65073 0.5703
Proportion of Variance 0.04112 0.03623 0.03529 0.0271
Cumulative Proportion    0.90138 0.93761 0.97290 1.0000
>
pov1<-c(0.47365,0.11150,0.08812,0.07550,0.05383,0.04693,0.04250,0.035460,0.024070,0.0223
70,0.014940,0.011140)
>
pov2<-c(0.476,0.07368,0.0643,0.05506,0.05248,0.04814,0.04565,0.04492,0.04112,0.03623,0.03
529,0.0271)
> barplot(pov1,ylab="proprtion",col=rainbow(12),names.arg = d)
> barplot(pov2,ylab="proprtion",col=rainbow(12),names.arg = d)
```