Homework 4

陈旭鹏

2018/4/29

1

```
mp <- matrix(c(5,2,2,2),2,2,byrow=T)
eigen(mp)
```

```
## eigen() decomposition
## $values
## [1] 6 1
##
## $vectors
##            [,1]       [,2]
## [1,] -0.8944272  0.4472136
## [2,] -0.4472136 -0.8944272
```

The eigenvalue-eigenvector pairs are $\lambda_1 = 6, e_1 = [\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}]; \lambda_2 = 1, e_2 = [-\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}]$.

Therefore, the principle componenets become:

$$Y_1 = e_1^T X = \frac{2}{\sqrt{5}} X_1 + \frac{1}{\sqrt{5}} X_2$$

$$Y_1 = e_2^T X = -\frac{1}{\sqrt{5}} X_1 + \frac{2}{\sqrt{5}} X_2$$

The total population variance explained by first principal component is:

$$\frac{var(Y_1)}{var(Y_1) + var(Y_2)} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{6}{1+6} \approx 85.71\%$$

2

a

```
cov2cor(mp)
```

```
##           [,1]      [,2]
## [1,] 1.0000000 0.6324555
## [2,] 0.6324555 1.0000000
```

$$\text{The correlation matrix } \rho = \begin{bmatrix} 1 & \sqrt{\frac{2}{5}} \\ \sqrt{\frac{2}{5}} & 1 \end{bmatrix}$$

The eigenvalue-eigenvector pairs are

$$\lambda_1 = \frac{5+\sqrt{10}}{5}, \; e_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$\lambda_1 = \frac{5-\sqrt{10}}{5}, \; e_1 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$Let\ \mathbf{Z_i} = \frac{\mathbf{X_i} - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, \ldots, p. \text{ The principal components become:}$$

$$Y_1 = e_1^T Z = \frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2$$

$$Y_1 = e_2^T Z = -\frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2$$

The total population variance explained by first principal component is:

$$\frac{var(Y_1)}{var(Y_1) + var(Y_2)} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{5+\sqrt{10}}{10} \approx 81.6\%$$

b

The principal components of Z obtained from the eigenvectors of the correlation matrix ρ of X is different from those calculated from covariance matrix Σ. Because the eigen pairs derived from Σ, in general not the same as the ones derived from ρ

c

THe correlations between Y_j and Z_i are:

$$\rho_{Y_1,Z_1} = e_{11}\sqrt{\lambda_1} = \frac{1}{\sqrt{2}}\sqrt{\frac{5+\sqrt{10}}{5}} \approx 0.903$$

$$\rho_{Y_1,Z_2} = e_{12}\sqrt{\lambda_1} = \frac{1}{\sqrt{2}}\sqrt{\frac{5+\sqrt{10}}{5}} \approx 0.903$$

$$\rho_{Y_2,Z_1} = e_{21}\sqrt{\lambda_2} = -\frac{1}{\sqrt{2}}\sqrt{\frac{5-\sqrt{10}}{5}} \approx -0.429$$

3

a

```
# read the data
setwd('~/Desktop/三春/3多元统计分析/作业/作业4/')
dat<-read.csv("table8.4.csv")
X1<-dat$x1
X2<-dat$x2
X3<-dat$x3
X4<-dat$x4
X5<-dat$x5
covar <- cov(dat)
covar
```

```
##              x1           x2           x3           x4           x5
## x1 4.332695e-04 0.0002756679 1.590265e-04 6.411929e-05 8.896616e-05
## x2 2.756679e-04 0.0004387172 1.799737e-04 1.814512e-04 1.232623e-04
## x3 1.590265e-04 0.0001799737 2.239722e-04 7.341348e-05 6.054612e-05
## x4 6.411929e-05 0.0001814512 7.341348e-05 7.224964e-04 5.082772e-04
## x5 8.896616e-05 0.0001232623 6.054612e-05 5.082772e-04 7.656742e-04
```

```
eigen(cov(dat))
```

```
## eigen() decomposition
## $values
## [1] 0.0013676780 0.0007011596 0.0002538024 0.0001426026 0.0001188868
##
## $vectors
##           [,1]       [,2]        [,3]       [,4]        [,5]
## [1,] 0.2228228  0.6252260  0.32611218  0.6627590  0.11765952
## [2,] 0.3072900  0.5703900 -0.24959014 -0.4140935 -0.58860803
## [3,] 0.1548103  0.3445049 -0.03763929 -0.4970499  0.78030428
## [4,] 0.6389680 -0.2479475 -0.64249741  0.3088689  0.14845546
## [5,] 0.6509044 -0.3218478  0.64586064 -0.2163758 -0.09371777
```

```
prcomp(dat)
```

```
## Standard deviations (1, .., p=5):
## [1] 0.03698213 0.02647942 0.01593118 0.01194163 0.01090352
##
## Rotation (n x k) = (5 x 5):
##           PC1        PC2         PC3        PC4         PC5
## x1 -0.2228228  0.6252260 -0.32611218  0.6627590 -0.11765952
## x2 -0.3072900  0.5703900  0.24959014 -0.4140935  0.58860803
## x3 -0.1548103  0.3445049  0.03763929 -0.4970499 -0.78030428
## x4 -0.6389680 -0.2479475  0.64249741  0.3088689 -0.14845546
## x5 -0.6509044 -0.3218478 -0.64586064 -0.2163758  0.09371777
```

```
summary(prcomp(dat))
```

```
## Importance of components%s:
##                            PC1     PC2     PC3     PC4     PC5
## Standard deviation     0.03698 0.02648 0.01593 0.01194 0.01090
## Proportion of Variance 0.52926 0.27133 0.09822 0.05518 0.04601
## Cumulative Proportion  0.52926 0.80059 0.89881 0.95399 1.00000
```

$$
\begin{aligned}
Y_1 &= e_1^T X = -0.2228228X_1 - 0.3072900X_2 - 0.1548103X_3 - 0.6389680X_4 - 0.6509044X_5 \\
Y_2 &= e_2^T X = 0.6252260X_1 + 0.5703900X_2 + 0.3445049X_3 - 0.2479475X_4 - 0.3218478X_5 \\
Y_3 &= e_3^T X = -0.32611218X_1 + 0.24959014X_2 + 0.03763929X_3 + 0.64249741X_4 - 0.64586064X_5 \\
Y_4 &= e_4^T X = 0.6627590X_1 - 0.4140935X_2 - 0.4970499X_3 + 0.3088689X_4 - 0.2163758X_5 \\
Y_5 &= e_5^T X = -0.11765952X_1 + 0.58860803X_2 - 0.78030428X_3 - 0.14845546X_4 + 0.09371777X_5
\end{aligned}
$$

b From the summary above, the proportion of the total sample variance explained by the rst three principal components is: 89.881%. It means that the first three explain almost all variance.

c

From 8-33, we have the CI of m λ_i:

$$
[\frac{\hat{\lambda}_i}{1 + z(\alpha/2m)\sqrt{2/n}}, \frac{\hat{\lambda}_i}{1 - z(\alpha/2m)\sqrt{2/n}}]
$$

```
z <-qnorm(1-1/6)
cical <-function(lambda){
  c(lambda/(1+z*(1/103)**0.5),lambda/(1-z*(1/103)**0.5))
}
cical(0.0013676780)
```

```
## [1] 0.001248653 0.001511786
```

```
cical(0.0007011596)
```

```
## [1] 0.0006401396 0.0007750385
```

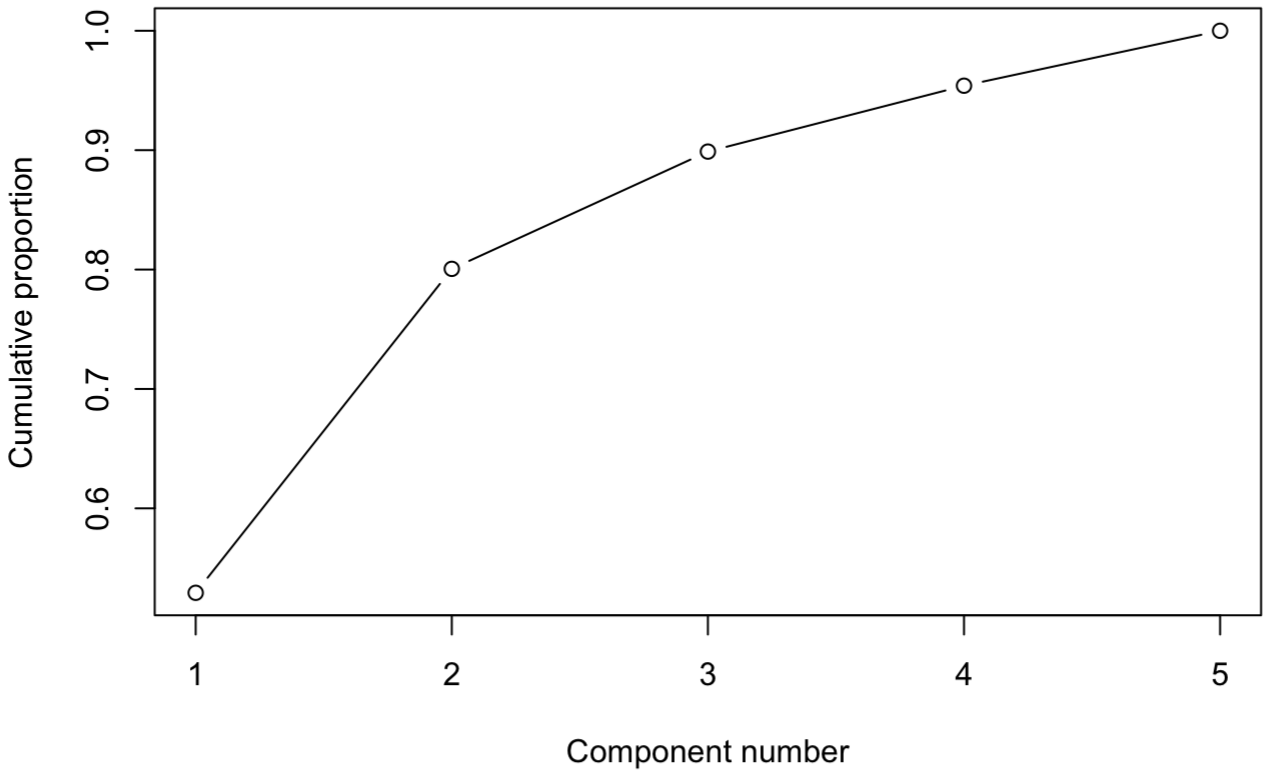
```
cical(0.0002538024)
```

```
## [1] 0.0002317147 0.0002805447
```

CIs are: [0.001248653 0.001511786], [0.0006401396 0.0007750385], [0.0002317147 0.0002805447]

d

```
plot(c(0.52926, 0.80059, 0.89881, 0.95399, 1.00000),ylab="Cumulative proportion",xlab
="Component number",type='b')
```

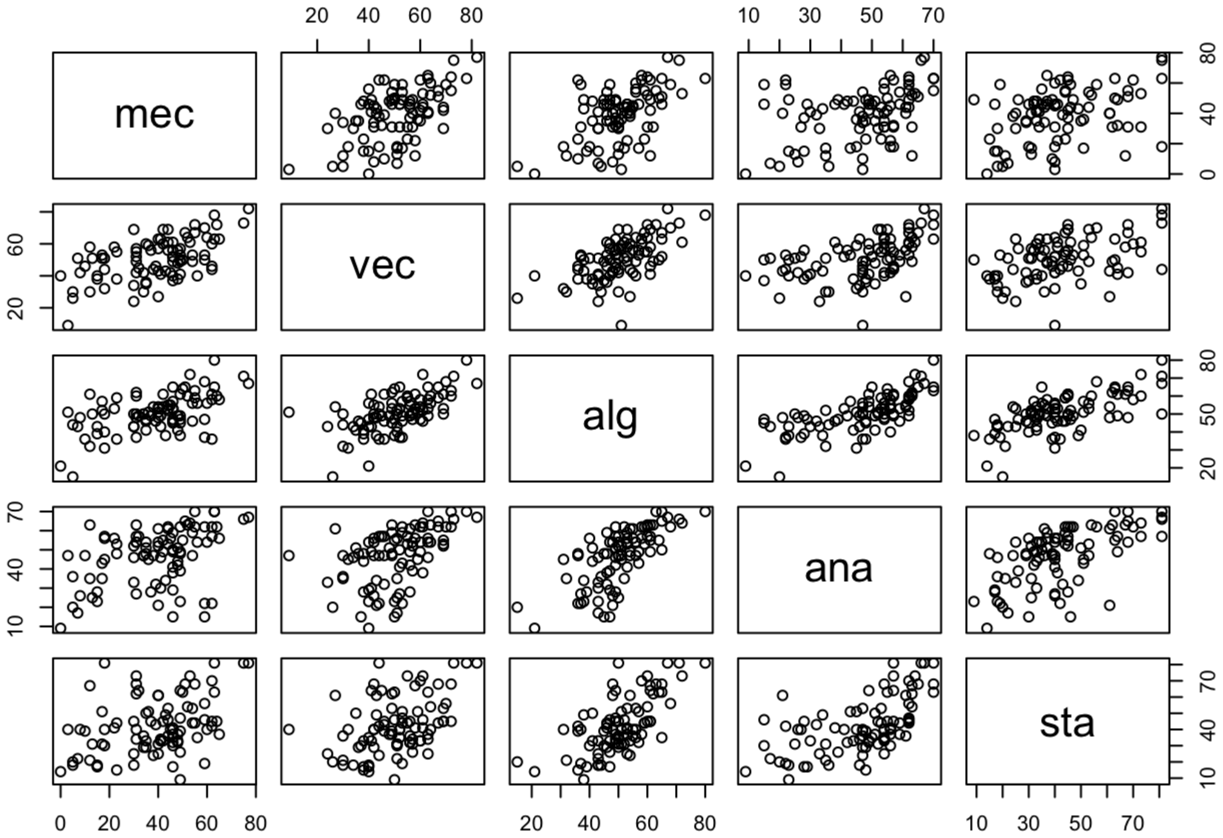


From the cumulative proportion plot, it seems that three dimensions' PC are enough.

4

a

```
library(bootstrap)
data(scor)
plot(scor)
```



b

```
cor(scor)
```

```
##           mec       vec       alg       ana       sta
## mec 1.0000000 0.5534052 0.5467511 0.4093920 0.3890993
## vec 0.5534052 1.0000000 0.6096447 0.4850813 0.4364487
## alg 0.5467511 0.6096447 1.0000000 0.7108059 0.6647357
## ana 0.4093920 0.4850813 0.7108059 1.0000000 0.6071743
## sta 0.3890993 0.4364487 0.6647357 0.6071743 1.0000000
```

c

```
prcomp(scor)
```

```
## Standard deviations (1, .., p=5):
## [1] 26.210490 14.216577 10.185642  9.199481  5.670387
##
## Rotation (n x k) = (5 x 5):
##            PC1         PC2        PC3          PC4         PC5
## mec -0.5054457 -0.74874751  0.2997888 -0.296184264 -0.07939388
## vec -0.3683486 -0.20740314 -0.4155900  0.782888173 -0.18887639
## alg -0.3456612  0.07590813 -0.1453182  0.003236339  0.92392015
## ana -0.4511226  0.30088849 -0.5966265 -0.518139724 -0.28552169
## sta -0.5346501  0.54778205  0.6002758  0.175732020 -0.15123239
```

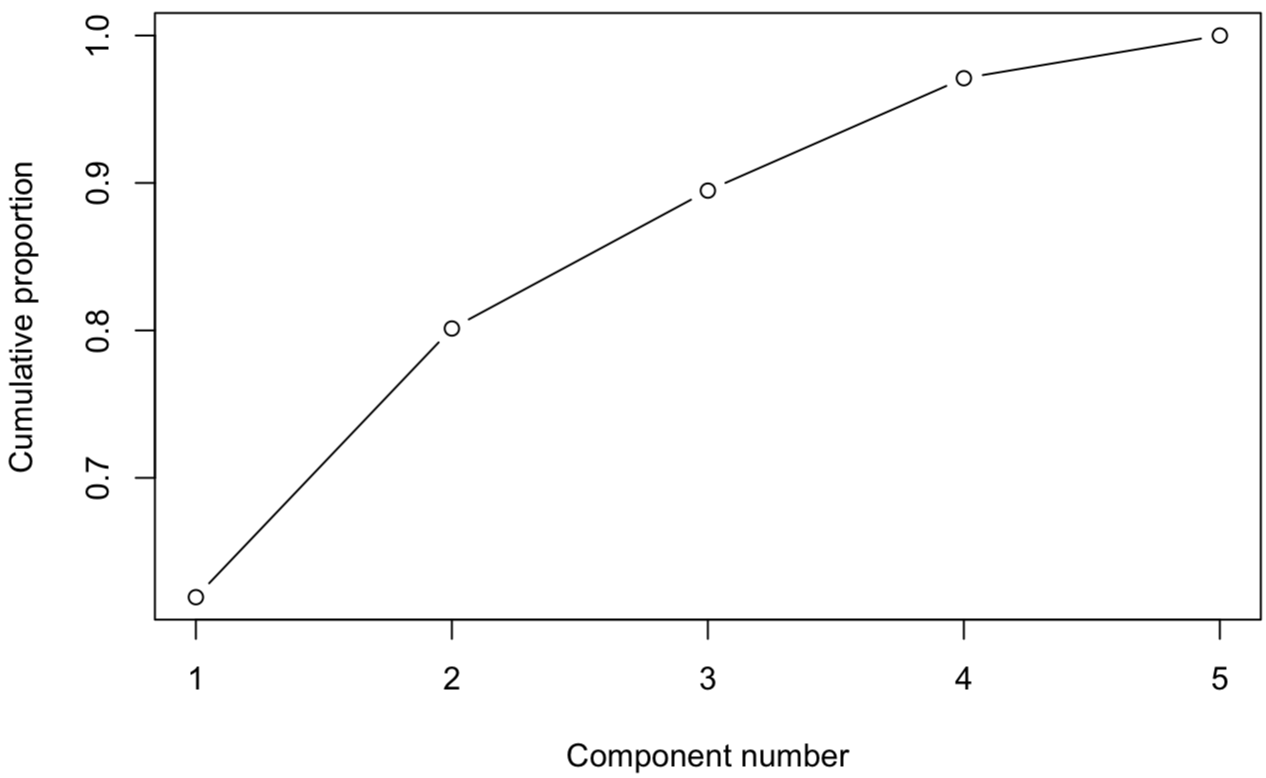
```
summary(prcomp(scor))
```

```
## Importance of components%s:
##                           PC1     PC2     PC3     PC4     PC5
## Standard deviation     26.2105 14.2166 10.1856 9.19948 5.67039
## Proportion of Variance  0.6191  0.1821  0.0935 0.07627 0.02898
## Cumulative Proportion   0.6191  0.8013  0.8948 0.97102 1.00000
```

$$Y_1 = e_1^T X = -0.5054457X_1 - 0.3683486X_2 - 0.3456612X_3 - 0.4511226X_4 - 0.5346501X_5$$
$$Y_2 = e_2^T X = -0.74874751X_1 - 0.20740314X_2 + 0.07590813X_3 + 0.30088849X_4 + 0.54778205X_5$$
$$Y_3 = e_3^T X = 0.2997888X_1 - 0.4155900X_2 - 0.1453182X_3 - 0.5966265X_4 + 0.6002758X_5$$
$$Y_4 = e_4^T X = -0.296184264X_1 + 0.78288817X_2 + 0.003236339X_3 - 0.518139724X_4 + 0.175732020X_5$$
$$Y_5 = e_5^T X = -0.07939388X_1 - 0.18887639X_2 + 0.92392015X_3 - 0.28552169X_4 - 0.15123239X_5$$

d

```
plot(c( 0.6191,0.8013 ,0.8948 ,0.97102, 1.00000),ylab="Cumulative proportion",xlab="C
omponent number",type='b')
```



will choose the first too for these three PCs take almost 80% of total variance.

e

PC1 may stand for the indicator of scores on all subjects. PC2 has more straightforward mearning: it is related to closed or open rules.

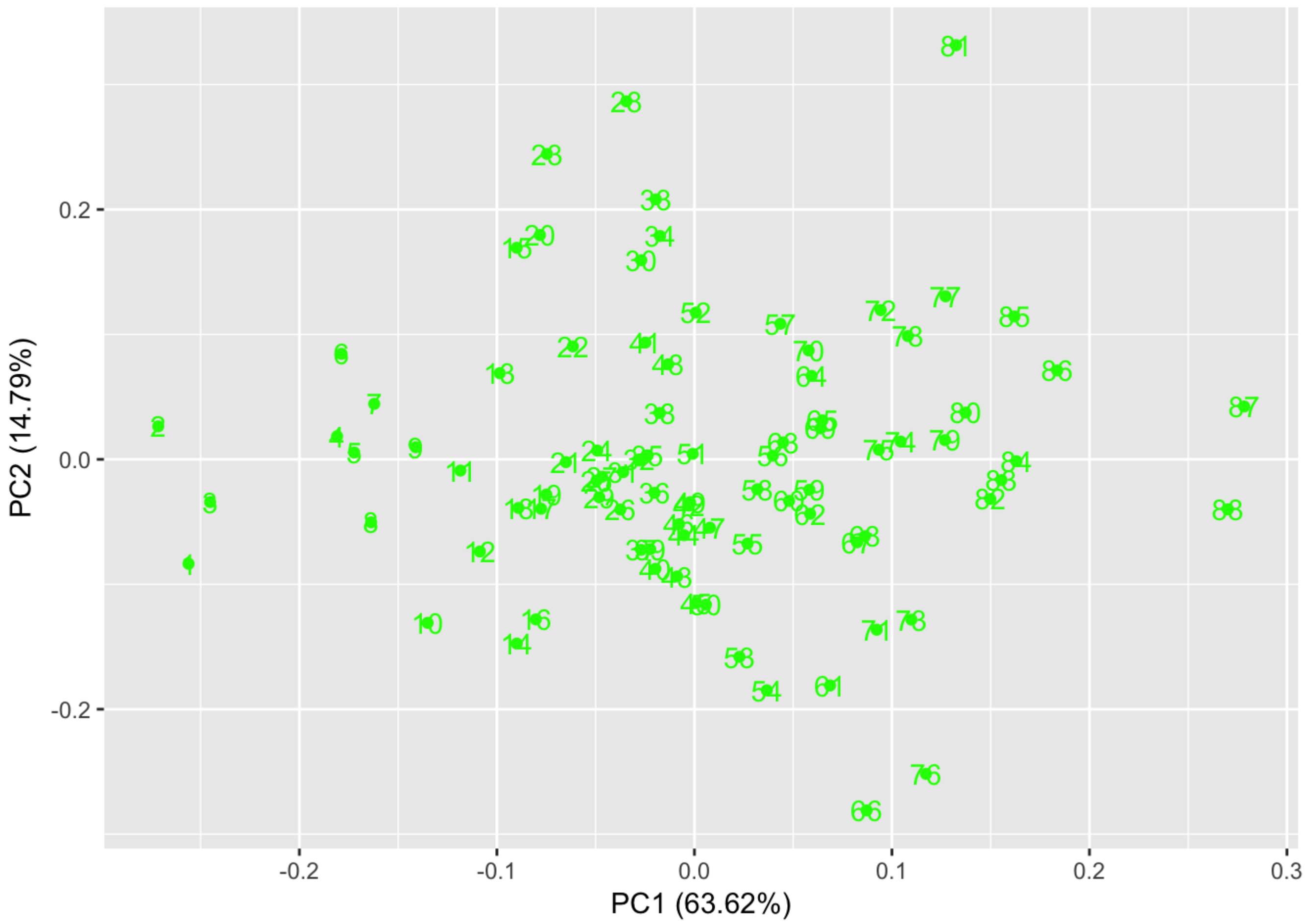
f

```r
library('ggfortify')
```

```
## Warning: package 'ggfortify' was built under R version 3.4.4
```

```
## Loading required package: ggplot2
```

```r
autoplot(prcomp(scor,scale=TRUE),colour='green',label=TRUE)
```



g

$\chi_2^2(0.05) = 5.99$ I use python to check the outlier:

```
import numpy as np
from sklearn.decomposition import PCA
def convert(strr):
    return np.array(strr.split(' ')).astype('float').reshape(-1,1)
pca = PCA(n_components=2, svd_solver='full')
dat = pca.fit_transform(data)
def ellipse(i):
    x,y = dat[i,0],dat[i,1]
    a =  (x/26.2105)**2 + (y/14.2166)**2
    if a >=5.99:
        print (i,a)
for i in range(data.shape[0]):
    ellipse(i)
```

And we can find eight outliers: 1,2,23,28,66,76,81,87