Exercise 11.7

Apply an NPCA to the U.S. CRIME data set (Sect. B.8). Interpret the results. Would it be necessary to look at the third PC? Can you see any difference between the four regions? Redo the analysis excluding the variable “area of the state”.

Sol.

U.S. CRIME data set: 50 個州，11 種變數(7 種犯罪方法:$X_3 \sim X_9$)

X_1: land area (land)土地面積

X_2: population 1985 (popu 1985)人口數

X_3: murder (murd)謀殺

X_4: rape 強姦

X_5: robbery (robb)搶劫

X_6: assault (assa)襲擊

X_7: burglary (burg)入室竊盜

X_8: larcery (larc)竊盜

X_9: autothieft (auto)汽車竊盜

X_{10}: U.S. states region number (reg)地區編號

(1: Northeast 東北部; 2: Midwest 中西部; 3: South 南部; 4: West 西部)

X_{11}: U.S. states division number (div)分區編號

由於各變數單位不一致，考慮將資料轉換後標準化再開始做主成分分析。

標準化過後，經由 R 程式計算得特徵值(四捨五入至小數第三位)為：

$\lambda_1 = 4.801 \quad \lambda_2 = 2.435 \quad \lambda_3 = 1.421 \quad \lambda_4 = 0.807 \quad \lambda_5 = 0.563$

$\lambda_6 = 0.297 \quad \lambda_7 = 0.244 \quad \lambda_8 = 0.188 \quad \lambda_9 = 0.135 \quad \lambda_{10} = 0.091 \quad \lambda_{11} = 0.018$

特徵向量為：

$$\begin{bmatrix} -0.17 & 0.31 & -0.03 & 0.85 & 0.25 & -0.04 & 0.09 & 0.23 & -0.007 & 0.16 & -0.10 \\ -0.24 & -0.32 & -0.16 & 0.33 & -0.74 & 0.35 & -0.14 & -0.02 & -0.05 & -0.10 & 0.05 \\ -0.27 & 0.07 & -0.60 & -0.13 & 0.25 & 0.27 & 0.08 & 0.19 & 0.44 & -0.40 & 0.12 \\ -0.40 & 0.03 & -0.06 & 0.06 & 0.01 & -0.60 & -0.54 & -0.35 & 0.11 & -0.20 & 0.02 \\ -0.31 & -0.38 & -0.04 & 0.01 & -0.06 & -0.38 & 0.72 & -0.18 & 0.17 & 0.14 & -0.20 \\ -0.36 & -0.05 & -0.43 & -0.20 & 0.14 & -0.01 & -0.05 & 0.14 & -0.71 & 0.32 & 0.07 \\ -0.38 & -0.14 & 0.28 & -0.20 & 0.02 & 0.099 & -0.27 & 0.42 & 0.41 & 0.51 & -0.13 \\ -0.35 & 0.06 & 0.48 & -0.08 & -0.06 & -0.093 & 0.15 & 0.45 & -0.26 & -0.54 & 0.21 \\ -0.28 & -0.31 & 0.31 & 0.11 & 0.48 & 0.49 & -0.05 & -0.47 & -0.10 & -0.09 & 0.01 \\ -0.24 & 0.51 & 0.06 & -0.19 & -0.16 & 0.146 & 0.17 & -0.21 & -0.07 & -0.06 & -0.69 \\ -0.24 & 0.51 & 0.09 & -0.08 & -0.22 & 0.155 & 0.13 & -0.30 & 0.11 & 0.26 & 0.63 \end{bmatrix}$$

透過特徵值除以特徵值總和可以得到解釋變異的百分比及累積解釋變異百分比

如下表：

表一：特徵值及其解釋變異的百分比

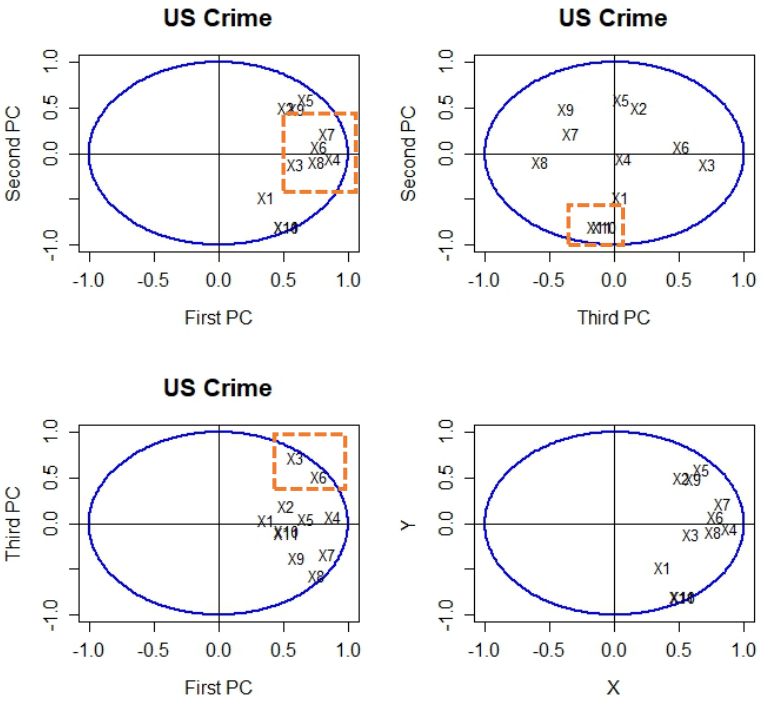
Eigenvalue	Percentages	Cumulated percentages
4.801	0.44	**0.436**
2.435	0.22	**0.658**
1.421	0.13	**0.787**
0.807	0.07	0.860
0.563	0.05	0.911
0.297	0.027	0.939
0.244	0.022	0.961
0.188	0.017	0.978
0.135	0.012	0.990
0.091	0.008	0.998
0.018	0.0016	1

根據表一顯示，第一組主成分可以解釋 43.6%的總變異；前兩組主成分解釋 65.8%總變異；前三組主成分解釋 78.7%總變異。另外，根據經驗法則指出可以選擇特徵值大於 1 的最低那組主成分，因此應取前三組主成分進行分析。

表二：前三組主成分與原始變數關係

	PC1	PC2	PC3
X_1	0.368	-0.483	0.041
X_2	0.519	0.504	0.190
X_3	0.593	-0.113	**0.718**
X_4	**0.884**	-0.054	0.073
X_5	0.674	0.598	0.051
X_6	**0.778**	0.074	0.513
X_7	**0.837**	0.225	-0.337
X_8	**0.757**	-0.093	-0.573
X_9	0.605	0.491	-0.371
X_{10}	0.533	**-0.797**	-0.075
X_{11}	0.528	**-0.797**	-0.109

圖一：前三組主成分與原始變數相關係數圖



圖一為根據表二畫出的關係圖，它們顯示前三組主成分與原始變數之間的關聯，由表二可以看出第一組主成分與強姦、襲擊、入室竊盜、竊盜(X_4, X_6, X_7, X_8)四種變數有明顯正相關；第二組主成分與地區編號及分區編號(X_{10}, X_{11})有較強的負相關；第三組主成分與謀殺(X_3)有較強正相關。圖一左上角的圖可以應證強姦、襲擊、入室竊盜、竊盜在第一組主成分有高度正相關；右上角的圖看出地區及分區編號與第二組主成分有高度負相關；左下圖謀殺為最明顯與群體分離的變數，它與地三組主成分有一定程度正向相關性。因此可以整理出X_4, X_6, X_7, X_8為第一組 PC Score 的主要變異因素；X_{10}, X_{11}為第二組 PC Score 的主要變異因素；X_3為第三組 PC Score 的主要變異因素。

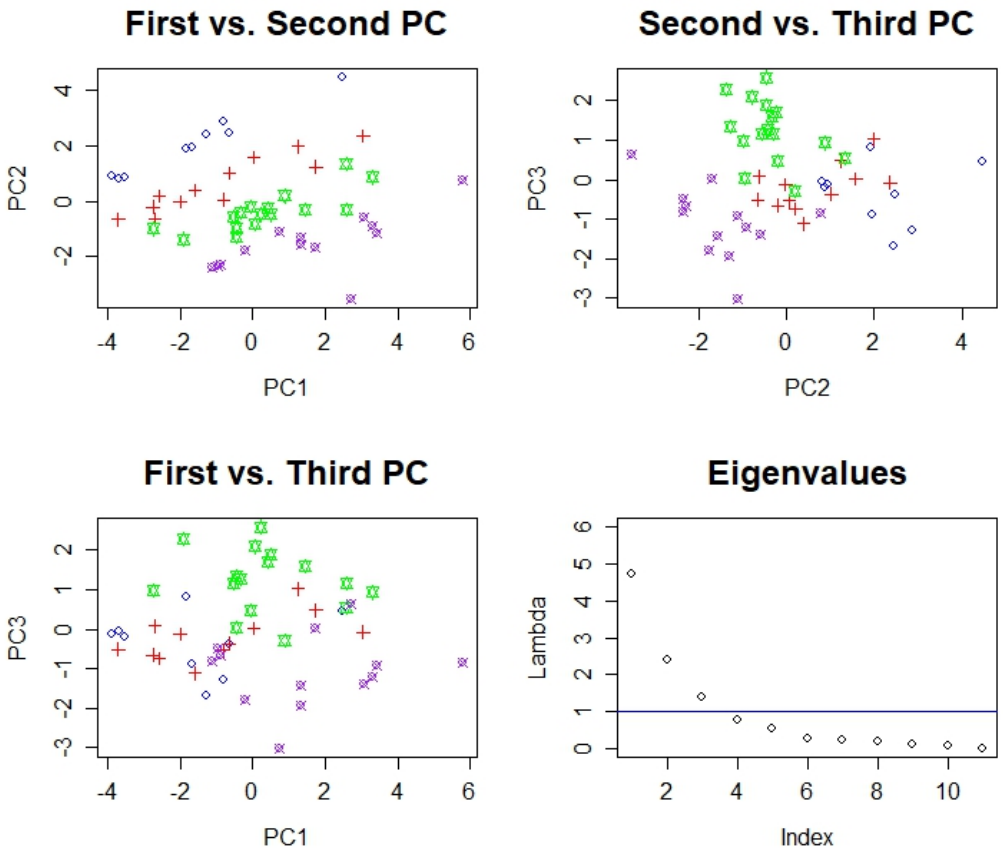
圖二：npca 散佈圖

○ ：東北部

＋ ：中西部

✡ ：南部

⊗ ：西部



圖二左上角的圖可以看出四個地區的變數在 PC1 與 PC2 之間有正向關係，並且四個地區大致被區分開來，其分層在 PC2 較明顯，表示第二組主成分對地區的影響較大。右上角圖可以看出對於 PC3 而言，西部與南部有較明顯的分層，表示在謀殺此一犯罪行為上，西部與南部有較明顯差異。左下角圖終統樣由 PC3 角度觀察可以看出西部與南部較明顯不同，然而東北部與中西部在 PC3 無太大區別度。右下角陡坡圖可在此驗證取到第三組主成分是更為合適的作法。

Excluding the variables X_{10}&X_{11}:

Eigenvalue:

$\lambda_1 = 4.463 \quad \lambda_2 = 1.469 \quad \lambda_3 = 1.154 \quad \lambda_4 = 0.729 \quad \lambda_5 = 0.453$

$\lambda_6 = 0.277 \quad \lambda_7 = 0.225 \quad \lambda_8 = 0.133 \quad \lambda_9 = 0.097$

Eigenvector:

$$\begin{bmatrix} -0.12 & -0.28 & 0.72 & 0.51 & 0.22 & -0.05 & -0.16 & 0.01 & -0.23 \\ -0.29 & 0.003 & -0.37 & 0.70 & -0.35 & 0.38 & 0.03 & -0.05 & 0.11 \\ -0.27 & -0.61 & -0.10 & -0.23 & 0.18 & 0.24 & -0.25 & 0.44 & 0.38 \\ -0.40 & -0.14 & 0.21 & -0.07 & -0.28 & -0.32 & 0.73 & 0.12 & 0.18 \\ -0.38 & 0.12 & -0.34 & 0.17 & 0.16 & -0.73 & -0.34 & 0.14 & -0.07 \\ -0.37 & -0.40 & -0.15 & -0.25 & 0.02 & 0.05 & -0.03 & -0.71 & -0.33 \\ -0.40 & 0.28 & 0.06 & -0.25 & -0.15 & 0.30 & -0.05 & 0.43 & -0.62 \\ -0.33 & 0.35 & 0.39 & -0.20 & -0.34 & 0.04 & -0.43 & -0.24 & 0.46 \\ -0.33 & 0.38 & -0.02 & 0.04 & 0.75 & 0.25 & 0.26 & -0.13 & 0.20 \end{bmatrix}$$

表三：特徵值及其解釋變異的百分比(去除X_{10}&X_{11})

Eigenvalue	Percentages	Cumulated percentages
4.463	0.496	**0.496**
1.469	0.163	**0.659**
1.154	0.128	**0.787**
0.729	0.081	0.868
0.453	0.050	0.919
0.277	0.031	0.949
0.225	0.025	0.974
0.133	0.015	0.989
0.097	0.011	1

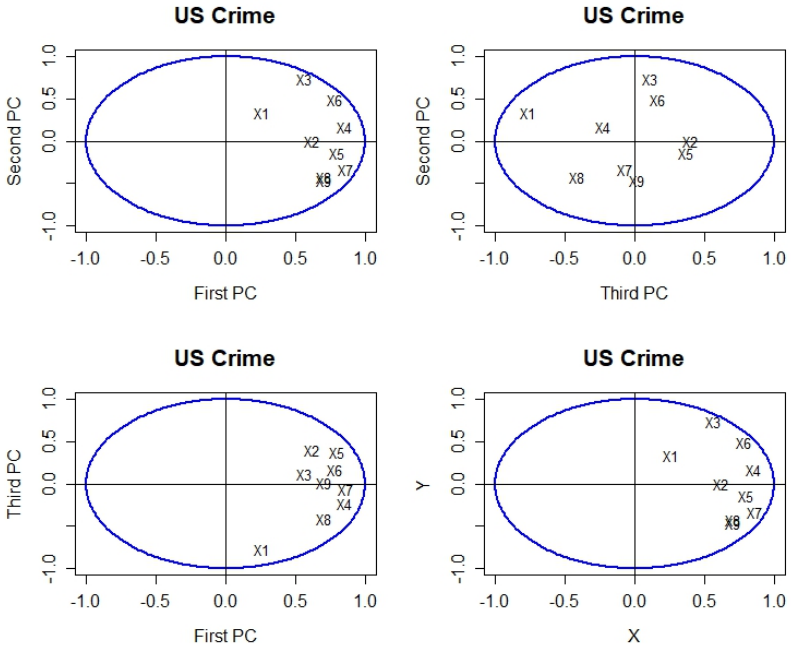
由表三可以得知：第一組主成分可以解釋 49.6%的總變異；前兩組主成分解釋 65.9%總變異；前三組主成分解釋 78.7%總變異。另外，根據經驗法則指出可以選擇特徵值大於 1 的最低那組，因此應取前三組主成分進行分析。下表四取前三組主成分分數與原始變數進行相關性分析。

表四：前三組主成分與原始變數關係(去除X_{10}&X_{11})

	PC1	PC2	PC3
X_1	0.262	0.340	**-0.769**
X_2	0.621	-0.004	0.399
X_3	0.561	**0.742**	0.113
X_4	**0.851**	0.172	-0.227
X_5	**0.796**	-0.146	0.369

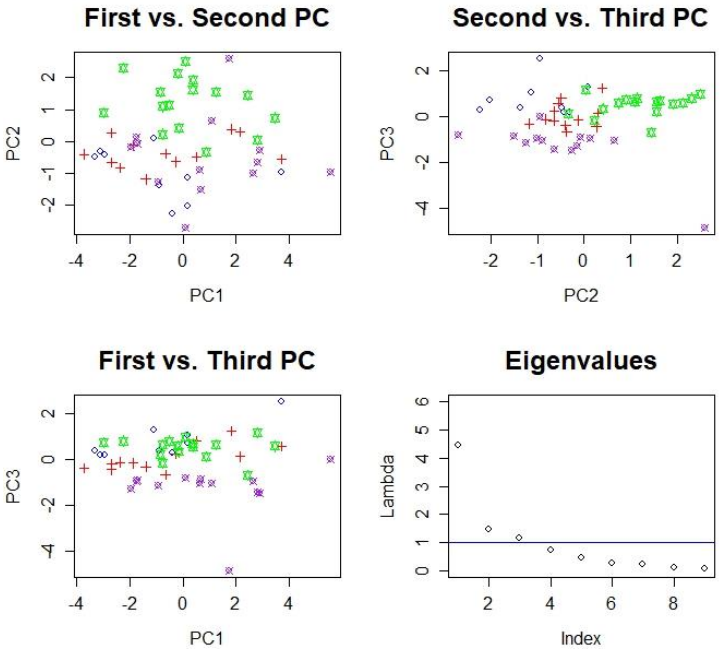
X_6	**0.781**	0.490	0.164
X_7	**0.859**	-0.338	-0.068
X_8	0.702	-0.427	-0.415
X_9	0.702	-0.464	0.017

圖三：前三組主成分與原始變數相關係數圖(去除X_{10}&X_{11})



透過表四及圖三可看出除了x_1外的其他變數在 PC1 皆呈現一定的相關性，代表各犯罪行為對 PC1 影響最甚。另外，也可以觀察到 PC2 與 PC3 在各變數的表現皆沒有太明顯的分群，表示各犯罪行為無明顯的關聯性。

圖四：npca 散佈圖(去除X_{10}&X_{11})



將圖四與圖二做比較，可以明顯看出若將地區編號及分區編號去除在外，死個地區得分群就沒那麼明顯。但由左上與右上圖中仍可以發現美國南部與其他地區的不同，且為 PC2 產生的效果。

NPCA of the variables $X_3 \sim X_9$:

Eigenvalue:

$\lambda_3 = 4.077 \quad \lambda_4 = 1.432 \quad \lambda_5 = 0.631 \quad \lambda_6 = 0.340$

$\lambda_7 = 0.248 \quad \lambda_8 = 0.140 \quad \lambda_9 = 0.132$

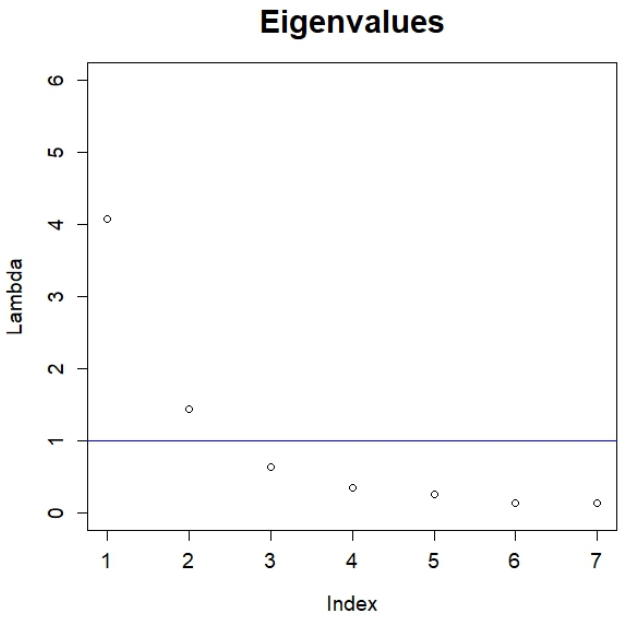
Eigenvector:

$$\begin{bmatrix} -0.276 & 0.644 & -0.010 & -0.329 & 0.203 & 0.100 & 0.591 \\ -0.421 & 0.116 & -0.360 & 0.296 & -0.759 & -0.065 & 0.107 \\ -0.387 & -0.046 & 0.604 & 0.645 & 0.190 & 0.069 & 0.161 \\ -0.388 & 0.456 & 0.011 & -0.067 & 0.136 & 0.100 & -0.780 \\ -0.436 & -0.257 & -0.155 & -0.144 & 0.292 & -0.783 & 0.027 \\ -0.360 & -0.401 & -0.508 & 0.048 & 0.360 & 0.561 & 0.069 \\ -0.354 & -0.366 & 0.472 & -0.601 & -0.337 & 0.208 & -0.012 \end{bmatrix}$$

表五：特徵值及其解釋變異的百分比($X_3 \sim X_9$)

Eigenvalue	Percentages	Cumulated percentages
4.077	0.5824	**0.5824**
1.432	0.2045	**0.7869**
0.631	0.0901	0.8771
0.340	0.0486	0.9257
0.248	0.0355	0.9612
0.140	0.0200	0.9811
0.132	0.0189	1

圖五：陡坡圖($X_3 \sim X_9$)

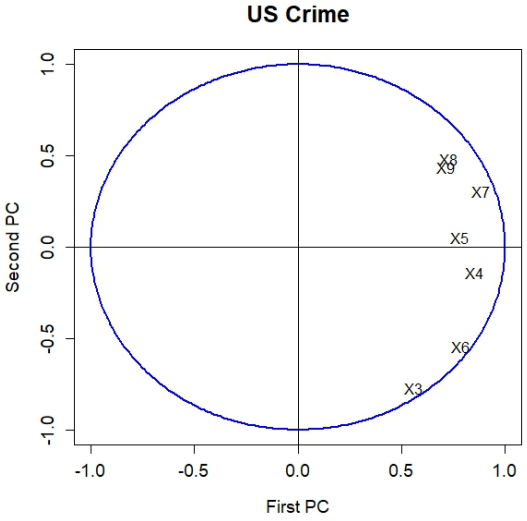


表五顯示第一組主成分可以解釋 58.24%的變異；前兩組主成分可以解釋 78.69%的變異。由圖五可看出前兩組特徵值位於 1 之上，因此該組資料可以取前兩組主成分進行分析即可。下表六為擷取前兩組主成分與原始變數的相關係數表。

表六：前二組主成分與原始變數關係(X_3~X_9)

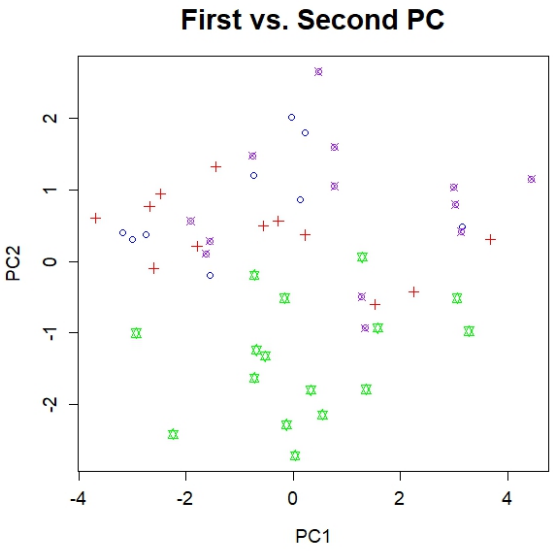
	PC1	PC2
X_3	0.557	**-0.771**
X_4	**0.851**	-0.139
X_5	**0.782**	0.055
X_6	**0.784**	-0.546
X_7	**0.881**	0.308
X_8	0.728	0.480
X_9	0.714	0.438

圖六：前兩組主成分與原始變數相關係數圖(X_3~X_9)



表六可看出在 PC1 中，強姦、搶劫、襲擊、入室竊盜可以解釋多數變異，在圖六也可應證變數X_4~X_7最為貼近圓圈邊界，表示其相關係數較高。

圖七：npca 散佈圖(X_3~X_9)



藉由圖七 PC1 與 PC2 的散佈圖可以看到四個地區之間沒有明顯分層，而美國南部較為獨立於其他地區。推測南部犯罪率高於美國其他州，根據統計大部分平均家庭收入在全國排名靠後州的犯罪率會高於全國平均水平。

R Code:

```
rm(list = ls(all = TRUE))
graphics.off()
# load data
x <-   read.table("C:/Users/user/Desktop/多變量
11101/HW7_1123/uscrime.dat")
n1 <- nrow(x)
n2 <- ncol(x)

# standardize the data
x <- (x - matrix(mean(as.matrix(x)), n1, n2, byrow = T))/matrix(sqrt((n1 - 1)
*
            apply(x, 2, var)/n1), n1, n2, byrow = T)

# spectral decomposition
eig <- eigen((n1 - 1) * cov(x)/n1)
e <- eig$values
v <- eig$vectors
# eigenvalues and percentage
perc = e/sum(e)
cum   = cumsum(e)/sum(e)
xv    = as.matrix(x) %*% v # principal components
xv    = xv * (-1)

# correlation of the first 3 PC
corr = cor(x, xv)[, 1:3]
r12   = corr[1:11, 1:2]
r13   = cbind(corr[1:11, 1], corr[1:11, 3])
r32   = cbind(corr[1:11, 3], corr[1:11, 2])
r123 = corr[1:11, 1:3]

# plot of cor of PC1&2
par(mfrow = c(2, 2))
```

```
ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab =
"Second PC",
     main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9", "X10", "X11")
text(r12, label)
# plot of cor of PC3&2
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "Third PC", ylab =
"Second PC",
     main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9", "X10", "X11")
text(r32, label)
# plot of cor of PC1&3
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab =
"Third PC",
     main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9", "X10", "X11")
text(r13, label)
# plot of cor of PC1&2&3
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "X", ylab = "Y",
cex.lab = 1.2,
     cex.axis = 1.2, lwd = 2)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9", "X10", "X11")
text(r123, label)

# plot of PC1&2
```

```
par(mfrow = c(2, 2))
plot(xv[, 1], xv[, 2], pch = c(rep(1, 9), rep(3, 12), rep(11, 16), rep(13, 13)), col
= c(rep("blue", 9),
      rep("red", 12), rep("green1", 16), rep("purple", 13)), xlab = "PC1", ylab
= "PC2", main = "First vs. Second PC", cex.lab = 1.2,
      cex.axis = 1.2, cex.main = 1.8)
# plot of PC2&3
plot(xv[, 2], xv[, 3], pch = c(rep(1, 9), rep(3, 12), rep(11, 16), rep(13, 13)), col
= c(rep("blue", 9),
      rep("red", 12), rep("green1", 16), rep("purple", 13)), xlab = "PC2", ylab
= "PC3", main = "Second vs. Third PC", cex.lab = 1.2,
      cex.axis = 1.2, cex.main = 1.8)
# plot of PC1&3
plot(xv[, 1], xv[, 3], pch = c(rep(1, 9), rep(3, 12), rep(11, 16), rep(13, 13)), col
= c(rep("blue", 9),
      rep("red", 12), rep("green1", 16), rep("purple", 13)), xlab = "PC1", ylab
= "PC3", main = "First vs. Third PC", cex.lab = 1.2,
      cex.axis = 1.2, cex.main = 1.8)
# plot of the eigenvalues
plot(e, ylim = c(0, 6), xlab = "Index", ylab = "Lambda", main =
"Eigenvalues",
      cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.8)
abline(h=1, col="blue")

#  去除 x10,x11
x1 <- x[,-(10:11)]
n3 <- ncol(x1)
x11 <- (x1 - matrix(mean(as.matrix(x1)), n1, n3, byrow =
T))/matrix(sqrt((n1 - 1) *

apply(x1, 2, var)/n1), n1, n3, byrow = T)
eig1 <- eigen((n1 - 1) * cov(x11)/n1)
e1 <- eig1$values
```

```
v1 <- eig1$vectors
perc1 = e1/sum(e1)
cum1   = cumsum(e1)/sum(e1)
xv1    = as.matrix(x11) %*% v1
xv1    = xv1 * (-1)
corr1 = cor(x11, xv1)[, 1:3]

r12_1   = corr1[1:9, 1:2]
r13_1   = cbind(corr1[1:9, 1], corr1[1:9, 3])
r32_1   = cbind(corr1[1:9, 3], corr1[1:9, 2])
r123_1 = corr1[1:9, 1:3]

par(mfrow = c(2, 2))
ucircle = cbind(cos((0:360)/180 * pi), sin((0:360)/180 * pi))
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab =
"Second PC",
      main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9")
text(r12_1, label)
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "Third PC", ylab =
"Second PC",
      main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9")
text(r32_1, label)
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab =
"Third PC",
      main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
```

```
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9")
text(r13_1, label)
plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "X", ylab = "Y",
     main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
label = c("X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8", "X9")
text(r123_1, label)

par(mfrow = c(2, 2))
plot(xv1[, 1], xv1[, 2], pch = c(rep(1, 9), rep(3, 12), rep(11, 16), rep(13, 13)),
col = c(rep("blue", 9),

rep("red", 12), rep("green1", 16), rep("purple", 13)), xlab = "PC1", ylab =
"PC2", main = "First vs. Second PC", cex.lab = 1.2,
     cex.axis = 1.2, cex.main = 1.8)
plot(xv1[, 2], xv1[, 3], pch = c(rep(1, 9), rep(3, 12), rep(11, 16), rep(13, 13)),
col = c(rep("blue", 9),

rep("red", 12), rep("green1", 16), rep("purple", 13)), xlab = "PC2", ylab =
"PC3", main = "Second vs. Third PC", cex.lab = 1.2,
     cex.axis = 1.2, cex.main = 1.8)
plot(xv1[, 1], xv1[, 3], pch = c(rep(1, 9), rep(3, 12), rep(11, 16), rep(13, 13)),
col = c(rep("blue", 9),

rep("red", 12), rep("green1", 16), rep("purple", 13)), xlab = "PC1", ylab =
"PC3", main = "First vs. Third PC", cex.lab = 1.2,
     cex.axis = 1.2, cex.main = 1.8)
plot(e1, ylim = c(0, 6), xlab = "Index", ylab = "Lambda", main =
"Eigenvalues",
     cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.8)
abline(h=1, col="blue")
```

```
# BONUS:只留 X3~X9
x2 <- x[,(3:9)]
n4 <- ncol(x2)
x12 <- (x2 - matrix(mean(as.matrix(x2)), n1, n4, byrow =
T))/matrix(sqrt((n1 - 1) *

apply(x2, 2, var)/n1), n1, n4, byrow = T)
eig2 <- eigen((n1 - 1) * cov(x12)/n1)
e2 <- eig2$values
v2 <- eig2$vectors
perc2 = e2/sum(e2)
cum2   = cumsum(e2)/sum(e2)
xv2    = as.matrix(x12) %*% v2
xv2    = xv2 * (-1)

corr2 = cor(x12, xv2)[, 1:2]
r12_2   = corr2[1:7, 1:2]

plot(ucircle, type = "l", lty = "solid", col = "blue", xlab = "First PC", ylab =
"Second PC",
       main = "US Crime", cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.6, lwd =
2)
abline(h = 0, v = 0)
label = c("X3", "X4", "X5", "X6", "X7", "X8", "X9")
text(r12_2, label)

plot(xv2[, 1], xv2[, 2], pch = c(rep(1, 9), rep(3, 12), rep(11, 16), rep(13, 13)),
col = c(rep("blue", 9),

rep("red", 12), rep("green1", 16), rep("purple", 13)), xlab = "PC1", ylab =
"PC2", main = "First vs. Second PC", cex.lab = 1.2,
       cex.axis = 1.2, cex.main = 1.8)
```

```
plot(e2, ylim = c(0, 6), xlab = "Index", ylab = "Lambda", main =
"Eigenvalues",
     cex.lab = 1.2, cex.axis = 1.2, cex.main = 1.8)
abline(h=1, col="blue")
```