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:)

: land area (land)土地面積

: population 1985 (popu 1985)人口數

: murder (murd)謀殺

: rape強姦

: robbery (robb)搶劫

: assault (assa)襲擊

: burglary (burg)入室竊盜

: larcery (larc)竊盜

: autothieft (auto)汽車竊盜

: 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)

: 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：

特徵向量為：

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

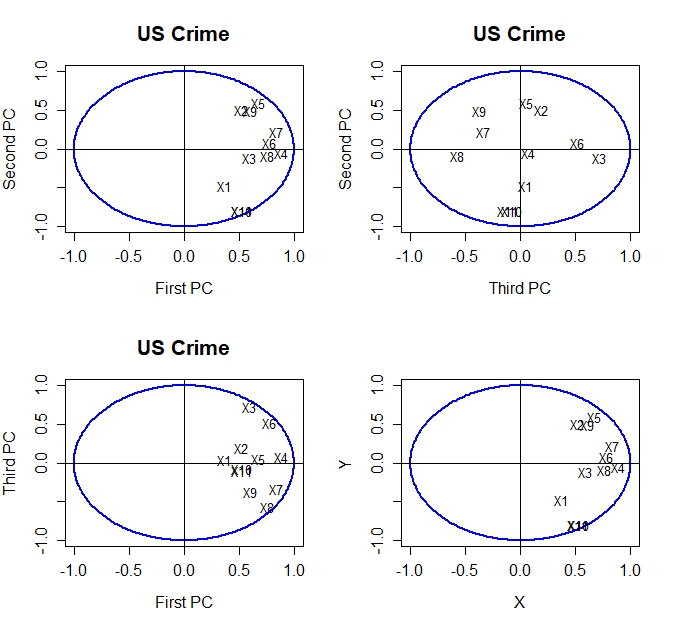
|  |  |  |
| --- | --- | --- |
| Eigenvalue | Percentages | Cumulated percentages |
|  | 0.44 | **0.436** |
|  | 0.22 | **0.658** |
|  | 0.13 | **0.787** |
|  | 0.07 | 0.860 |
|  | 0.05 | 0.911 |
|  | 0.027 | 0.939 |
|  | 0.022 | 0.961 |
|  | 0.017 | 0.978 |
|  | 0.012 | 0.990 |
|  | 0.008 | 0.998 |
|  | 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 |
|  | 0.368 | -0.483 | 0.041 |
|  | 0.519 | 0.504 | 0.190 |
|  | 0.593 | -0.113 | **0.718** |
|  | **0.884** | -0.054 | 0.073 |
|  | 0.674 | 0.598 | 0.051 |
|  | **0.778** | 0.074 | 0.513 |
|  | **0.837** | 0.225 | -0.337 |
|  | **0.757** | -0.093 | -0.573 |
|  | 0.605 | 0.491 | -0.371 |
|  | 0.533 | **-0.797** | -0.075 |
|  | 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有較強的負相關；第三組主成分與謀殺有較強正相關。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。因此可以整理出為第一組PC Score的主要變異因素；為第二組PC Score的主要變異因素；為第三組PC Score的主要變異因素。

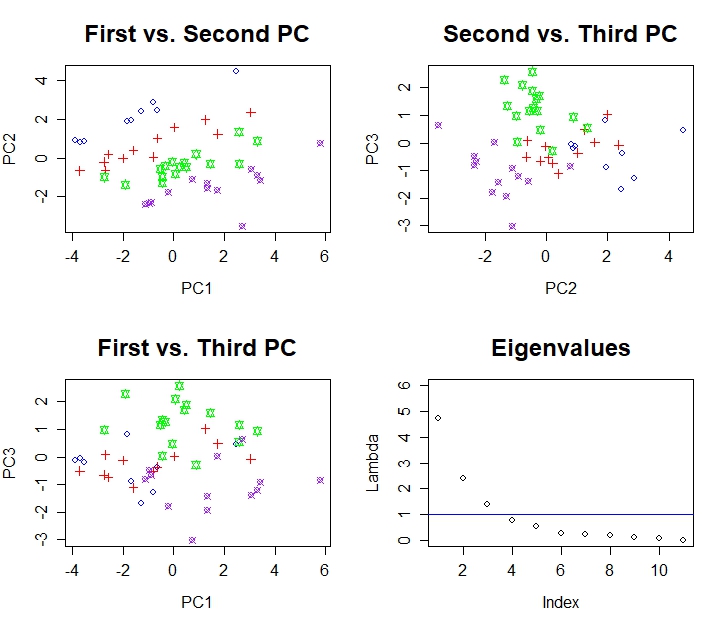
圖二：npca散佈圖

* ：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 :**

Eigenvalue:

Eigenvector:

表三：特徵值及其解釋變異的百分比(去除)

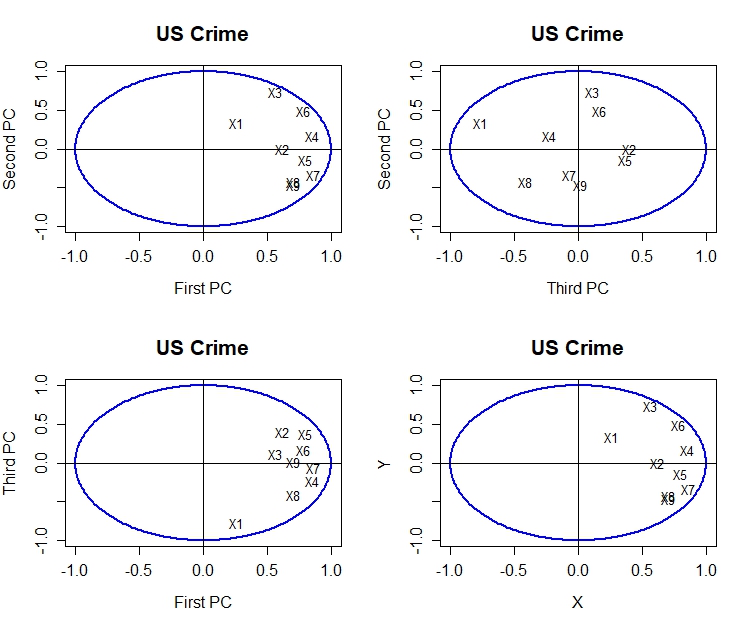
|  |  |  |
| --- | --- | --- |
| Eigenvalue | Percentages | Cumulated percentages |
|  | 0.496 | **0.496** |
|  | 0.163 | **0.659** |
|  | 0.128 | **0.787** |
|  | 0.081 | 0.868 |
|  | 0.050 | 0.919 |
|  | 0.031 | 0.949 |
|  | 0.025 | 0.974 |
|  | 0.015 | 0.989 |
|  | 0.011 | 1 |

　　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)

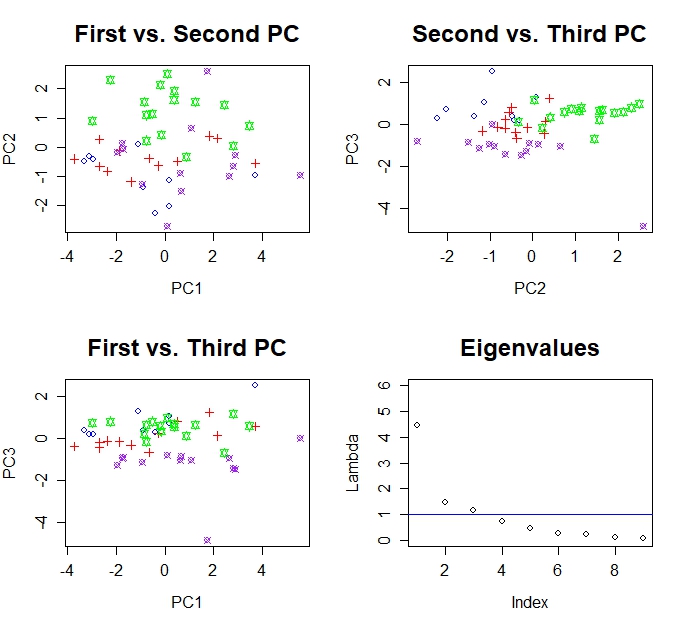
|  |  |  |  |
| --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 |
|  | 0.262 | 0.340 | **-0.769** |
|  | 0.621 | -0.004 | 0.399 |
|  | 0.561 | **0.742** | 0.113 |
|  | **0.851** | 0.172 | -0.227 |
|  | **0.796** | -0.146 | 0.369 |
|  | **0.781** | 0.490 | 0.164 |
|  | **0.859** | -0.338 | -0.068 |
|  | 0.702 | -0.427 | -0.415 |
|  | 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)



透過表四及圖三可看出除了外的其他變數在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)



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 :**

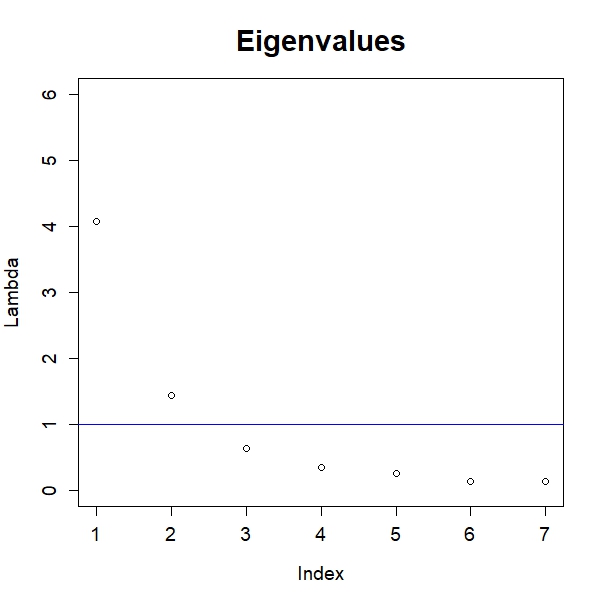
Eigenvalue:

Eigenvector:

表五：特徵值及其解釋變異的百分比()

|  |  |  |
| --- | --- | --- |
| Eigenvalue | Percentages | Cumulated percentages |
|  | 0.5824 | **0.5824** |
|  | 0.2045 | **0.7869** |
|  | 0.0901 | 0.8771 |
|  | 0.0486 | 0.9257 |
|  | 0.0355 | 0.9612 |
|  | 0.0200 | 0.9811 |
|  | 0.0189 | 1 |

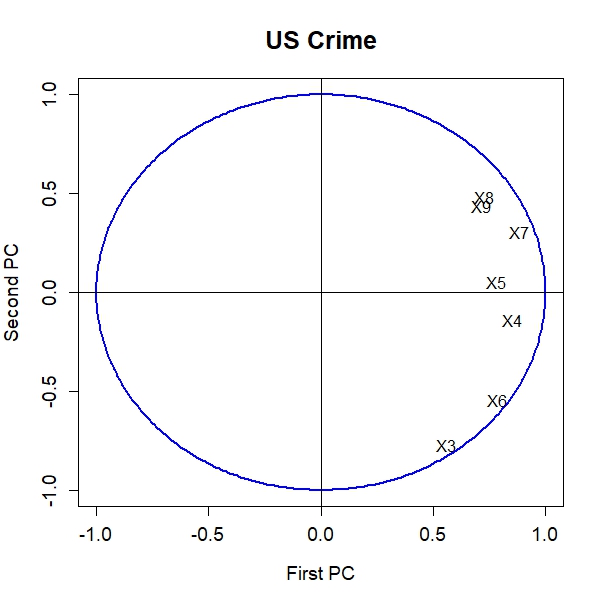
圖五：陡坡圖()



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。

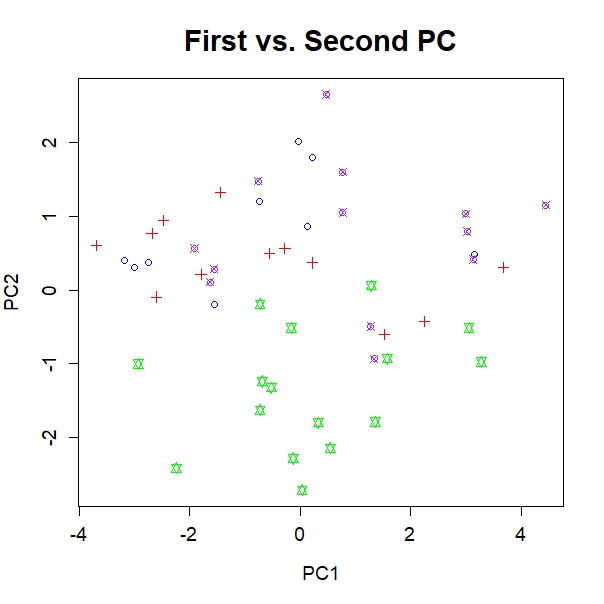
表六：前二組主成分與原始變數關係()

|  |  |  |
| --- | --- | --- |
|  | PC1 | PC2 |
|  | 0.557 | **-0.771** |
|  | **0.851** | -0.139 |
|  | **0.782** | 0.055 |
|  | **0.784** | -0.546 |
|  | **0.881** | 0.308 |
|  | 0.728 | 0.480 |
|  | 0.714 | 0.438 |

圖六：前兩組主成分與原始變數相關係數圖()

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最為貼近圓圈邊界，表示其相關係數較高。

圖七：npca散佈圖()



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")