Homework 8:

Perform a factor analysis on the variables X3–X9 in the U.S. crime data set (Sec. B.8). Would it make sense to use all of the variables for the analysis?

變數資料: (p=7, n=50)

: 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

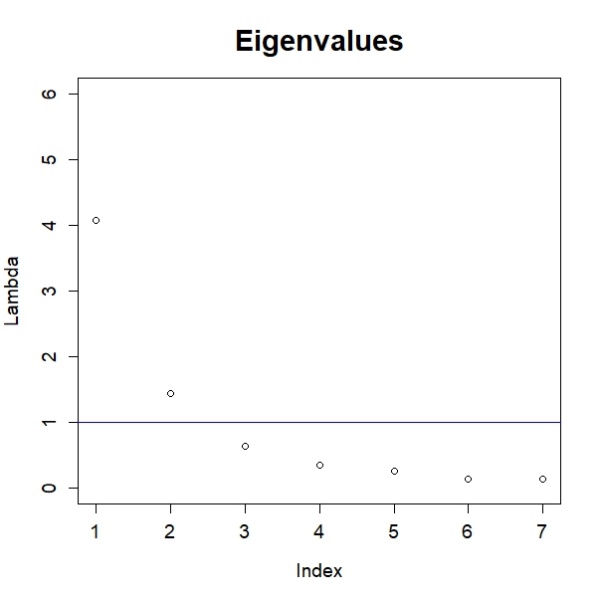
: autothieft (auto)汽車竊盜

Sol.

相關係數矩陣：

　　從相關係數矩陣可以看出襲擊跟謀殺具有高度相關，搶劫跟竊盜及搶劫跟汽車竊盜也有一定程度上的相關，三種相關性皆為邏輯上可以合理解釋。

Eigenvalue:



圖一：陡坡圖

由上式可以發現取前兩組主成分可以解釋78.7%的變異，從陡坡圖也可以驗證取m=2足夠解釋大部分資訊。

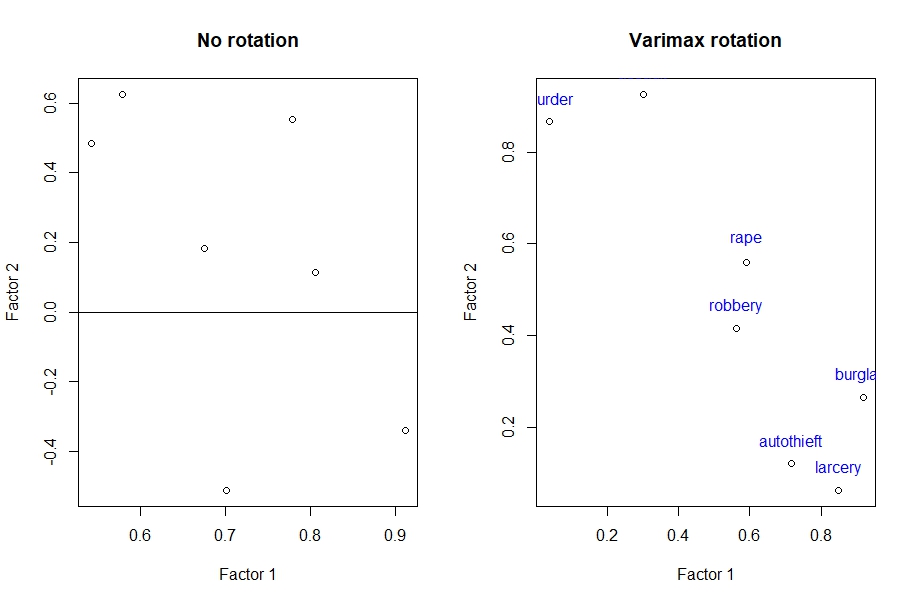
表一：因子旋轉前後表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Estimated factor loadings | | Specific variances | Estimated rotated factor loadings | | Communalities |
|  |  |  |  |  |  |
| Murder() | 0.7014 | -0.5109 | 0.247 | 0.0375 | 0.8670 | 0.753 |
| Rape() | 0.8057 | 0.1128 | 0.338 | 0.5903 | 0.5599 | 0.661 |
| Robbery() | 0.6761 | 0.1820 | 0.510 | 0.5637 | 0.4154 | 0.490 |
| Assault() | 0.9116 | -0.3404 | 0.053 | 0.3018 | 0.9251 | 0.947 |
| Burglary() | 0.7792 | 0.5526 | 0.087 | 0.9177 | 0.2651 | 0.913 |
| Larcery() | 0.5788 | 0.6245 | 0.275 | 0.8490 | 0.0636 | 0.725 |
| Autothieft() | 0.5424 | 0.4851 | 0.471 | 0.7172 | 0.1220 | 0.529 |
| Cumulative proportion of total (standardize) sample variance explained | 0.524 | 0.717 |  | 0.405 | 0.717 |  |

　　首先觀察表一整體狀況，總樣本方差累積比例顯示兩個因子解釋了整個數據集71.7%的方差。由程式可以驗證經過因子轉換並不會影響共同因子及特殊因子的值。其中，的特殊因子值很小，表示對襲擊與入室竊盜造成影響的多來自共同因子；相反地，的特殊因子值相對高於其他變數，表示搶劫跟汽車竊盜造成影響可能來自於共同因子之外的其他因素。

　　另外，比較左半邊未旋轉過與右半邊旋轉過後的估計因子載荷量，可以看出旋轉後有較明顯的分群，因子變得更好解釋，大致可歸類為兩區，在第二個因子上載荷較大；在第一個因子上載荷較大。

圖二：因子旋轉前後圖



　　由圖二可更直覺看出表一因子旋轉後的結果，原本每種變數四散各處，使用正交旋轉後，形成較明顯的兩塊區域，其中一塊較靠近第一個因子，另一塊靠近第二個因子，中間兩種變數較靠近於中間位置但仍稍微偏向第二個因子，因此我將它們歸類在第二個因子。

　　綜合表一與圖二可以將各變數分別歸類於兩個因子，第一個因子有強姦、搶劫、入室竊盜、竊盜、汽車竊盜，我認為這類相較第二類屬於較不致命的犯罪行為，而第二個因子包含謀殺、襲擊，則屬於較致命的犯罪行為。

Factor 1=較不致命的因子(強姦、搶劫、入室竊盜、竊盜、汽車竊盜)

Factor 2=較致命的因子(謀殺、襲擊)

Residual matrix殘差矩陣：

透過殘差矩陣可檢測當殘差值越接近0，表示因子模型足夠解釋該資料。因此根據上述算式得到的殘差矩陣，可以觀察出多數殘差值都非常接近0，表示m取2是足夠代表該資料的基本概況。另外，我們也可以透過概似比檢定再次檢驗m取2是否足夠解釋資料集：

Likelihood ratio statistic:

Where

根據Bartlett correction，當，可以拒絕。

　　以上兩種方法都可以證明m=2有足夠證據解釋該變數資料。

Factor score因子分數：

　　用迴歸方法估計因子分數，將變數中有高(大於.40)載荷量組成一組，其因子分數則是根據載荷的組合對組中變量的觀察值加總。第二個因子分數則是加總第二個因子中載荷量較高的變數。藉由簡化的因子分數我們可以達成數據降維。

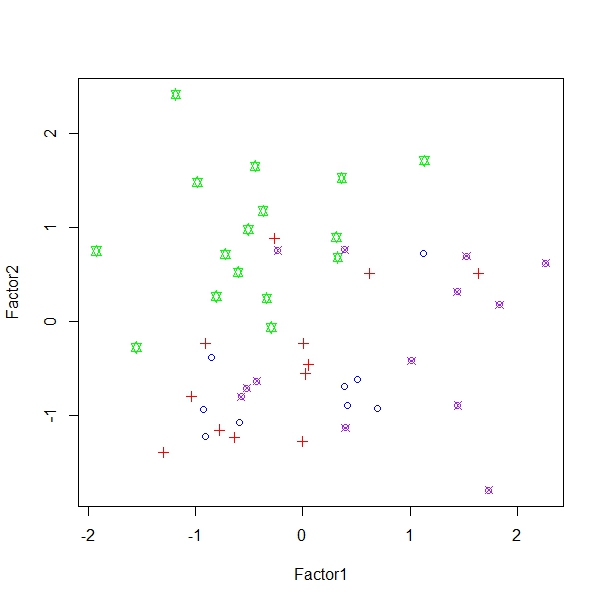
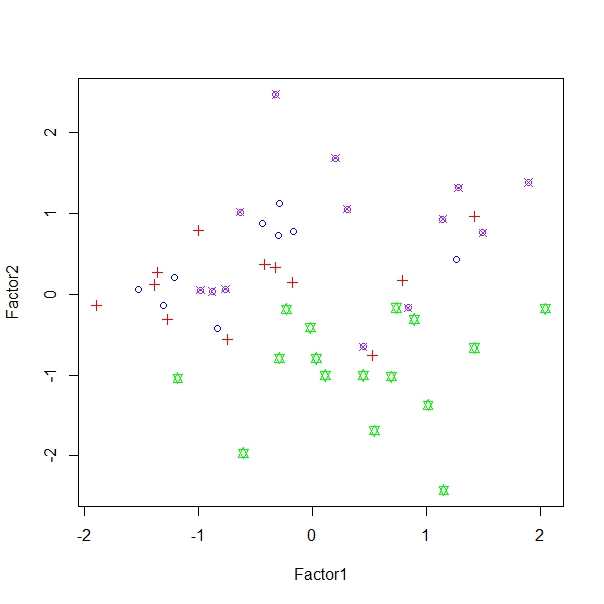
圖二：旋轉因子前後因子分數分佈圖

* ：東北部

＋ ：中西部

* : 南部

: 西部



　　圖二顯示所有因子分數的散布情況，左邊是因子旋轉前的散佈圖，右邊是因子旋轉後的散佈圖。兩軸的旋轉並不影響因子分數彼此的距離與相對位置，只影響他們的實際位置。由右圖來看，南部地區多靠近第二個因子，也就是謀殺、襲擊等較致命的因子，而東北部、中西部、西部則平均散佈在較靠近第一個因子，表示這些地區的犯罪類型多為較不致命的強姦、搶劫、竊盜等等。

R Code:

# clear variables and close windows

rm(list = ls(all = TRUE))

graphics.off()

# load data

data <- read.table("C:/Users/user/Desktop/多變量11101/uscrime.dat")

x <- data[,(3:9)]

# define variable names

colnames(x) = c("murder", "rape", "robbery", "assault", "burglary", "larcery", "autothieft")

# correlation matrix

r = cor(x)

# determine the nb of factors

n1 <- nrow(x)

n2 <- ncol(x)

xm <- (x - matrix(mean(as.matrix(x)), n1, n2, byrow = T))/matrix(sqrt((n1 - 1) \*

apply(x, 2, var)/n1), n1, n2, byrow = T)

eig <- eigen((n1 - 1) \* cov(xm)/n1)

e <- eig$values

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")

# factor analysis

# without rotate

x.fac <- factanal(x, factors = 2, rotation = "none", scores = "regression")

x.fac$loadings[,1]

x.fac$loadings[,2]

x.fac$scores

# rotated

x.fac.r <- factanal(x, factors = 2, rotation="varimax", scores = "regression")

x.fac.r$loadings[,1]

x.fac.r$loadings[,2]

x.fac.r$scores

com <- 1 - x.fac.r$uniquenesses

# residual matrix

Lambda <- x.fac$loadings

Psi <- diag(x.fac$uniquenesses)

S <- x.fac$correlation

Sigma <- Lambda %\*% t(Lambda) + Psi

round(S - Sigma, 4) # round the result to 4 digits

det(Sigma)/det(S)

# scatter plot

plot(x.fac$scores, 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)))

plot(x.fac.r$scores, 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)))

par(mfrow = c(1,2))

plot(x.fac$loadings[,1],

x.fac$loadings[,2],

xlab = "Factor 1",

ylab = "Factor 2",

main = "No rotation")

abline(h = 0, v = 0)

plot(x.fac.r$loadings[,1],

x.fac.r$loadings[,2],

xlab = "Factor 1",

ylab = "Factor 2",

main = "Varimax rotation")

text(x.fac.r$loadings[,1],

x.fac.r$loadings[,2]+0.05,

colnames(x),

col="blue")

abline(h = 0, v = 0)

# 法二

library(psych)

fa <- fa(r, nfactors = 2, rotate = "none", fm = "ml", scores = "regression") # ml:最大似然法;pa:主軸迭代法;wls:加權最小二乘法

fa.varimax <- fa(r, nfactors = 2, rotate = "varimax", fm = "ml", scores = "regression")

factor.plot(fa.varimax, labels = rownames(fa.varimax$loadings), pch = fa.varimax$loadings)

fa.diagram(fa.varimax, digits = 3)

# digits = 3表示保留為小數