Midterm exam

Yu-Xiang Lin (ID: B06305030)

11/9/2020

```
library(tidyverse)
```

```
## ─ Attaching packages ─────────────────────── tidyverse 1.3.0 ─
```

```
## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0
```

```
## Warning: package 'tibble' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
## ─ Conflicts ──────────────────────── tidyverse_conflicts() ─
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(covTestR)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 3.0-1
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(grid)
library(Hotelling)
```

```
## Loading required package: corpcor
```

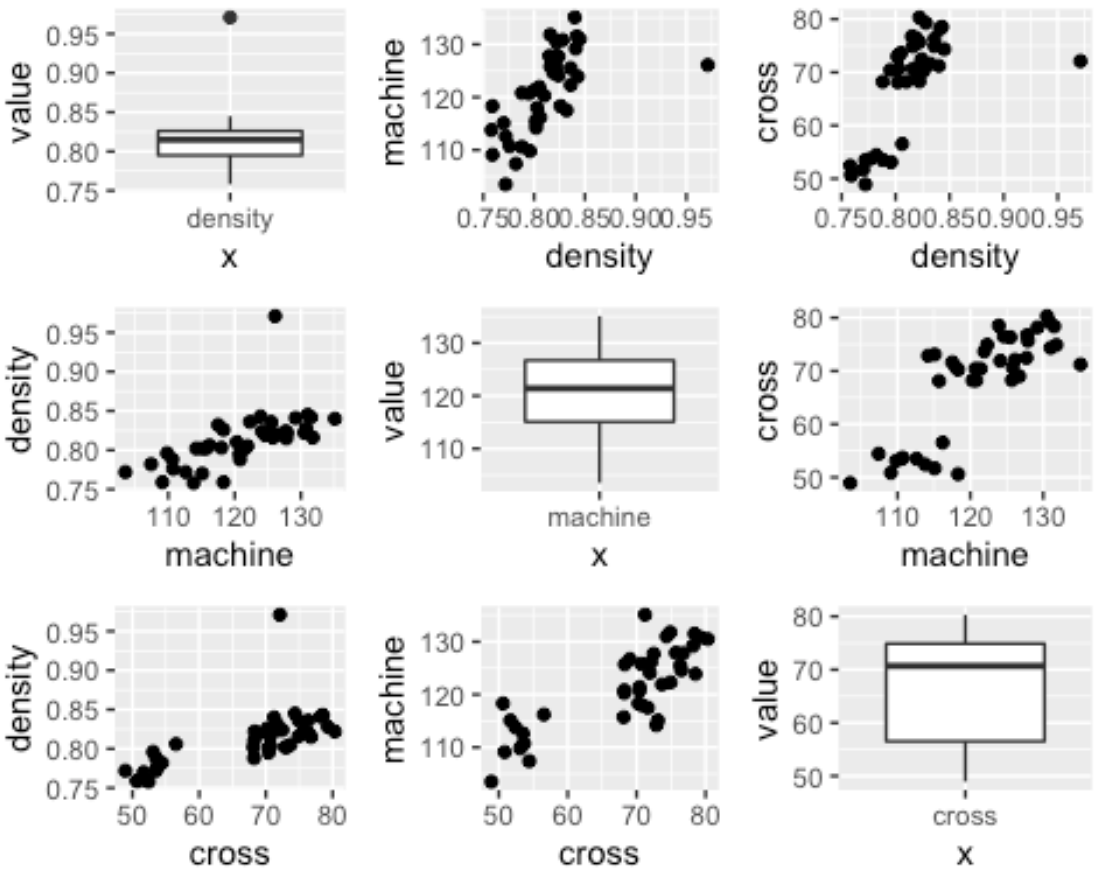
Problem 1

```
p1 <- read.table("/Users/linyuxiang/R/MSA/mid/paper.dat", header = T)
```

Paper is manufactured in continuous sheets several feet wide. Because of the orientation of fibers within the paper, it has a different strength when measured in the direction produced by the machine than when measured across, or at right angles to, the machine direction. The file "paper.dat" shows the measured values of
x1 = density(grams/cubic centinleter)
x2 = strength (pounds) in the machine direction
x3 = strength (pounds) in the cross direction

(a) Plot the scatter plots as the off-diagonal elements and box plots as the diagonal elements. (10%)

```
myplot <- function(i,ii){
   if (i != ii){
   p <- ggplot(p1, aes(x = p1[,i],y = p1[,ii]))+
   geom_point()+
   xlab(paste0(names(p1)[i]))+ylab(paste0(names(p1)[ii]))
}
return(p)
}
box_p1 <- p1 %>% gather("x", "y")

p1_1 <- subset(box_p1, x =="density") %>%
   ggplot(.,aes(x = x, y = y)) +
   geom_boxplot()+
   ylab("value")
p2_2 <- subset(box_p1, x =="machine") %>%
   ggplot(.,aes(x = x, y = y)) +
   geom_boxplot()+
   ylab("value")
p3_3 <- subset(box_p1, x =="cross") %>%
   ggplot(.,aes(x = x, y = y)) +
   geom_boxplot()+
   ylab("value")

p1_2 <- myplot(1,2);p1_3<- myplot(1,3)
p2_1 <- myplot(2,1);p2_3 <- myplot(2,3)
p3_1 <- myplot(3,1);p3_2 <- myplot(3,2)
grid.arrange(p1_1,p1_2,p1_3,
             p2_1,p2_2,p2_3,
             p3_1,p3_2,p3_3,nrow= 3)
```

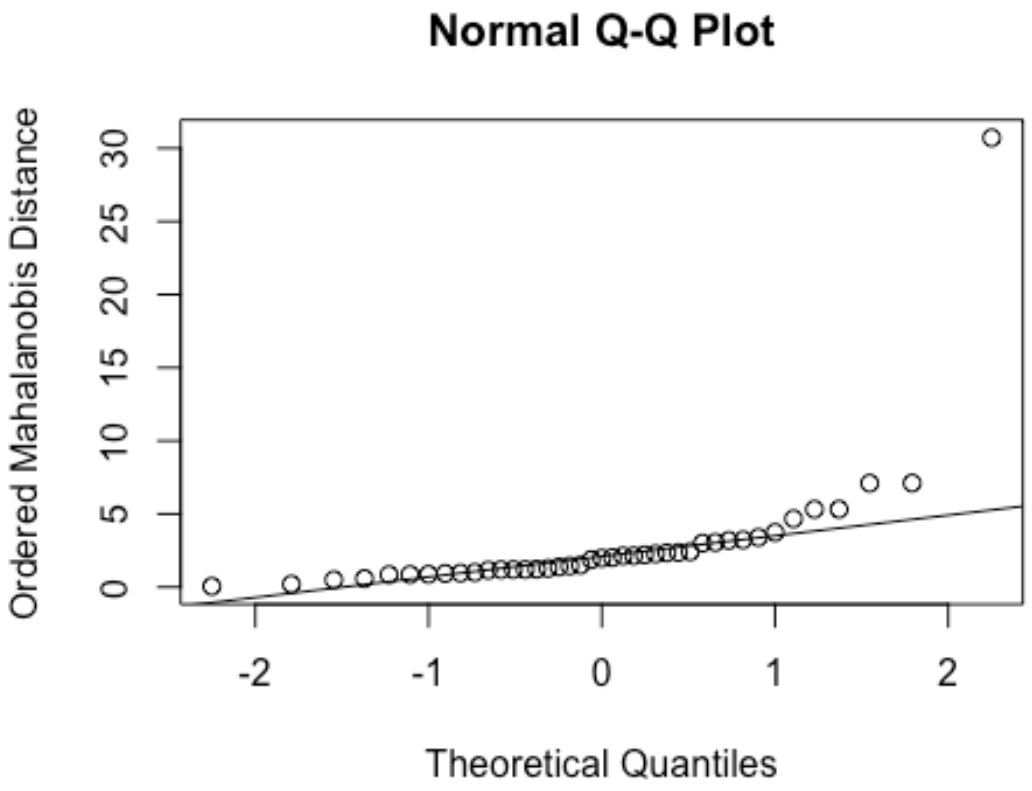


- Ans: 從散佈圖中可以看出變數 machine, density, croww 間幾乎都呈現正相關，且從箱型圖中可以發現似乎有一個觀察值在 density 變項中屬於 outlier

(b) Examine the data on paper-quality measurements for single and multivariate outliers as well as the multivariate normality.(10%)

```
## Single Outlier
density <- scale(p1$density); which(density > 2.5)

## [1] 25

machine <- scale(p1$machine); which(machine > 2.5)

## integer(0)

cross <- scale(p1$cross); which(cross > 2.5)

## integer(0)

## multivariate outliers
Sx <- cov(p1)
D2 <- mahalanobis(p1,colMeans(p1),Sx)
# simple
p <- ncol(p1)
which((D2 / p) >2.5)
```

```
## [1] 25

# Wilks' statisc
n <- nrow(p1)
w <- 1-((n*(D2)) / (n-1)^2)
Fi <- ((n-p-1)/p) * ((1/w)-1)
which(Fi > qf(0.995,p,n-p-1))

## [1] 25

## Multivariate normality
index <-  ((1:n)-0.5)/n
quant <- qchisq(index,p)
qq_p1 <- bind_cols(p1,data_frame(D2 = D2,quant = quant))

## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was gene
rated.

  qqnorm(qq_p1$D2,ylab = "Ordered Mahalanobis Distance");qqline(qq_p1$D
2)
```



- Ans: 針對個觀察樣本的每個變項進行 outlier 的檢定，結果發現只有第 25 個樣本的 density 數值屬於 outlier，其他樣本以及其他變數均沒有 outlier 存在。使用D_2/p > 2.5 作為 Multivariate outlier 檢定的標準，發現第 25 個樣本為 outlier。若進一步使用使用 Wilks' statisc 進行 Multivariate outlier 檢定可以得到該樣本為 outlier, 和起初使用 D2 / p > 2.5 的標準得到相同結果，outlier 為第 25 個觀察值。針對 Multivariate normality，從 Q-Q 圖中可以發現該資料可能是常態分配，大部分樣本的 Mahalanobis Distance 的數值均座落於 Q-Q line，只有少部分 Mahalannobis Distance 的數值略高於 Q-Q line，且只有一個樣本遠高於 Q-Q line

Problem 2

```
p2 <- read.table("/Users/linyuxiang/R/MSA/mid/love.dat")
```

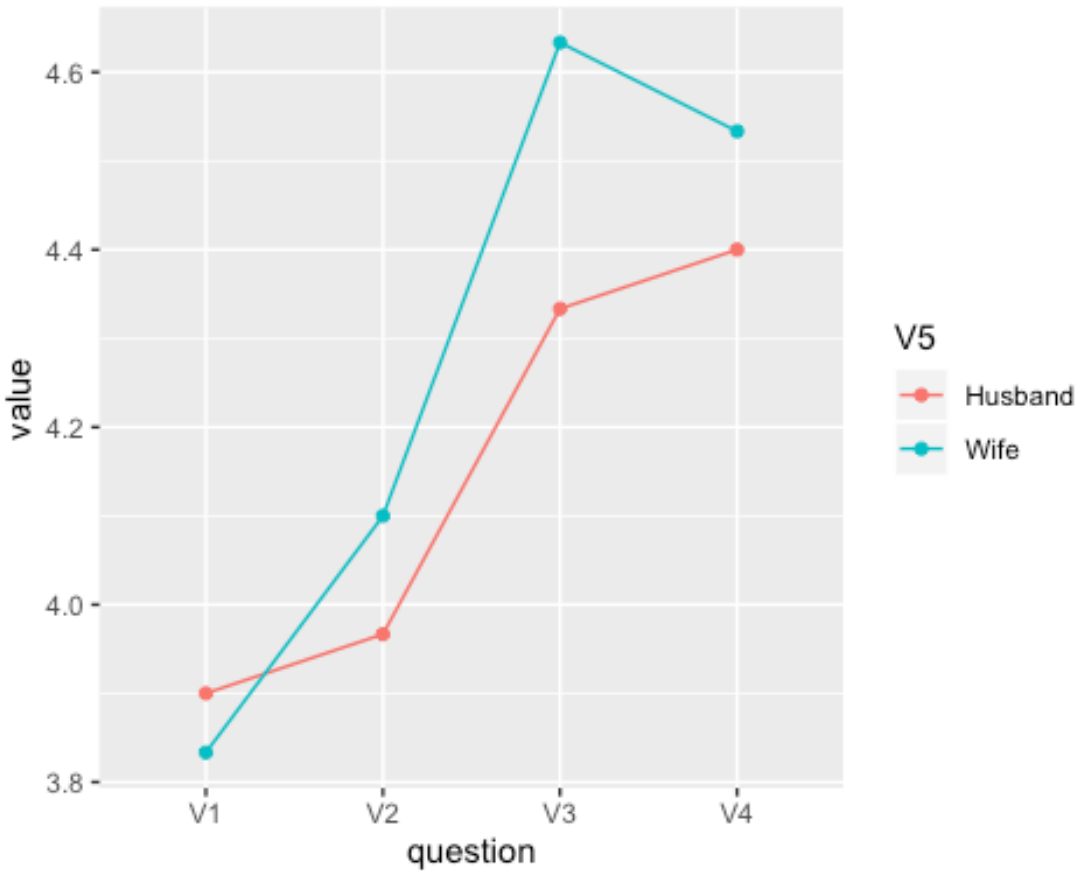
As part of the study of love and marriage, a sample of husbands and wives were asked to respond to these questions:
1. What is the level of passionate love you feel for your partner?
2. What is the level of passionate love that your partner feels for you?
3. What is the level of companionate love that you feel for your partner?
4. What is the level of companionate love that your partner feels for you?

The responses were recorded on the following 5-point scale. 1. None at all. 2. Very little. 3. Some. 4. A great deal. 5. Tremendous amount.

30 husbands and 30 wives gave the responses in the Thble, where X_1 = a 5-point-scale response to Question 1, X_2 = a 5-point-scale response to Question 2, X_3 = a 5-point-scale response to Question 3, and X_4 = a 5-point-scale response to Question 4. (love.dat)

(a) Plot the mean vectors for husbands and wives as sample profiles. (10%)

```
p2 %>% group_by(V5) %>%
   summarise_at(c("V1","V2","V3","V4"),mean) %>%
   gather("question","value", -V5) %>%
   ggplot(aes(x = factor(question), y = value, group = V5, color = V5))
+
   geom_line()+
   geom_point()+
   xlab("question")
```



- Ans: 從途中大致上可以看出 Husband 和 Wife 間在不同的問題上可能有平行的關係，但從 V2 和 V3 在 Husband 和 Wife 的差距不同即可看出可能沒有 same level

(b) Is the husband rating wife profile parallel to the wife rating husband profile? Test for parallel profiles with $\alpha = 0.05$. If the profiles appear to be parallel, test for coin- cident profiles at the same level of significance. Finally, if the profiles are coinci- dent, test for level profiles with $\alpha = 0.05$. What conclusion(s) can be drawn from this analysis? (20%)

```
g1 <- subset(p2, V5=="Husband")[,-5]
g2 <- subset(p2, V5 =="Wife")[,-5]
S_g1 <- cov(g1); S_g2 <- cov(g2); pooled <- (S_g1+S_g2)/2
n1 <- nrow(g1); n2 <- nrow(g2);n <- n1+n2
x1_bar <- apply(g1, 2, mean); x2_bar <- apply(g2, 2, mean)
x_bar <- (x1_bar+x2_bar)/2
p <- ncol(p2)-1
A <- matrix(c(1,-1,0,0,0,1,-1,0,0,0,1,-1),nrow = 3, byrow= T)
# parallelism
para_T <- ((n1*n2) / (n1+n2)) * t(x1_bar - x2_bar) %*% t(A) %*%
   solve(A%*%pooled%*%t(A)) %*% A %*%(x1_bar - x2_bar)
```

```
df1 <- p-1; df2 <- n1+n2-p
(n1+n2-p) / ((n1+n2-2)*(p-1)) * para_T;

##          [,1]
## [1,] 2.579917

(n1+n2-p) / ((n1+n2-2)*(p-1)) * para_T > qf(0.95, df1,df2); ## Paralle
relation 2.579917

##       [,1]
## [1,] FALSE

# Equal mean response
equal_resp_T2 <- (n1+n2)* (t(x_bar)%*%t(A))%*%solve(A%*%pooled%*%t(A))%
*%(A%*%x_bar)
df1 <- p-1; df2 <- n1+n2-p
(n1+n2-p)/((n1+n2-2)*(p-1)) * equal_resp_T2;

##         [,1]
## [1,] 8.18807

(n1+n2-p)/((n1+n2-2)*(p-1)) * equal_resp_T2 > qf(0.95, df1, df2); ## No
 equal mean response 8.18807

##      [,1]
## [1,] TRUE

# same level
vec_1 <- matrix(rep(1,length(x_bar)),nrow = 1)
abs(vec_1%*%(x1_bar-x2_bar) /
   sqrt((vec_1%*%pooled%*%t(vec_1)) * (1/n1 + 1/n2)));

##          [,1]
## [1,] 1.238051

abs(vec_1%*%(x1_bar-x2_bar) /
   sqrt((vec_1%*%pooled%*%t(vec_1)) * (1/n1 + 1/n2))) > qt(0.95,n1+n2-
2); ## Same Level T 1.238051

##       [,1]
## [1,] FALSE
```

- Ans: 在$\alpha = 0.95$ 下Profile analysis 的 Parallelism 檢驗中，F 值為 2.579917，小於 F(0.95, 3,56) = 2.76，因此無法拒絕兩者關係平行的虛無假設。接著在假設平行的前提下檢測 Equal mean response 和 Same level，在 Equal mean response 的檢驗中，F 值為 8.18807，大於 F(0.95, 3,56) = 2.76，拒絕 Equal mean response 的虛無假設，即 Husband 和 Wife 在不同變數間的差距可能是不同的；在 Same level 的檢驗中，T 值為 1.238051，小於 T(0.95, 58) = 1.67，無法拒絕 Same level 的虛無假設，也就是說在 Husband 和 Wife 組間問題的平均分數可能沒有明顯差異。

Problem 3

```
p3 <- read.csv("/Users/linyuxiang/R/MSA/mid/blood.csv")
```

The following is the investigation of hemophilia by the National Taiwan University Medical School. Among the 68 women under investigation, 31 were normal people and the other 37 were definite carriers of hemophilia. (blood.csv) The test items were:
X1 = factor VIII coagulant activity %
X2 = factor VIII related antigen %

(a) Test the homogeneity of covariance matrices. (10%)

```
homogeneityCovariances(p3, group = carrier)

##
##  Boxes' M Homogeneity of Covariance Matrices Test
##
## data:  0 and 1
## Chi-Squared = 3.7455, df = 496, p-value = 1
## alternative hypothesis: true difference in covariance matrices is no
t equal to 0
```

- Ans: p-value = 1, 拒絕$H_0: \Sigma_1 = \Sigma_2$，也就是說 Carrier 兩個組別 Covariance matrices 間可能存在差異。

(b) Test the equality between the mean vectors of two groups. (15%)

```
y <- cbind(p3$X1, p3$X2)
mod <- manova(y~as.factor(carrier), data = p3)
summary(mod, test = "Wilks") # F: 58.27

##                     Df   Wilks approx F num Df den Df    Pr(>F)
## as.factor(carrier)  1 0.35805   58.271      2     65 3.184e-15 ***
## Residuals          66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

fit <- hotelling.test(data = p3,.~carrier); fit

## Test stat:  58.271
## Numerator df:  2
## Denominator df:  65
## P-value:  3.109e-15
```

- Ans: 在$\alpha = 0.001$下達統計顯著，拒絕兩個組別間變數間平均數沒有不同的虛無假設，即在 carrier = 1 和 carrier = 0 的組別間 factor VIII coagulant activity, factor VIII related antigen 存在差異。

Problem 4

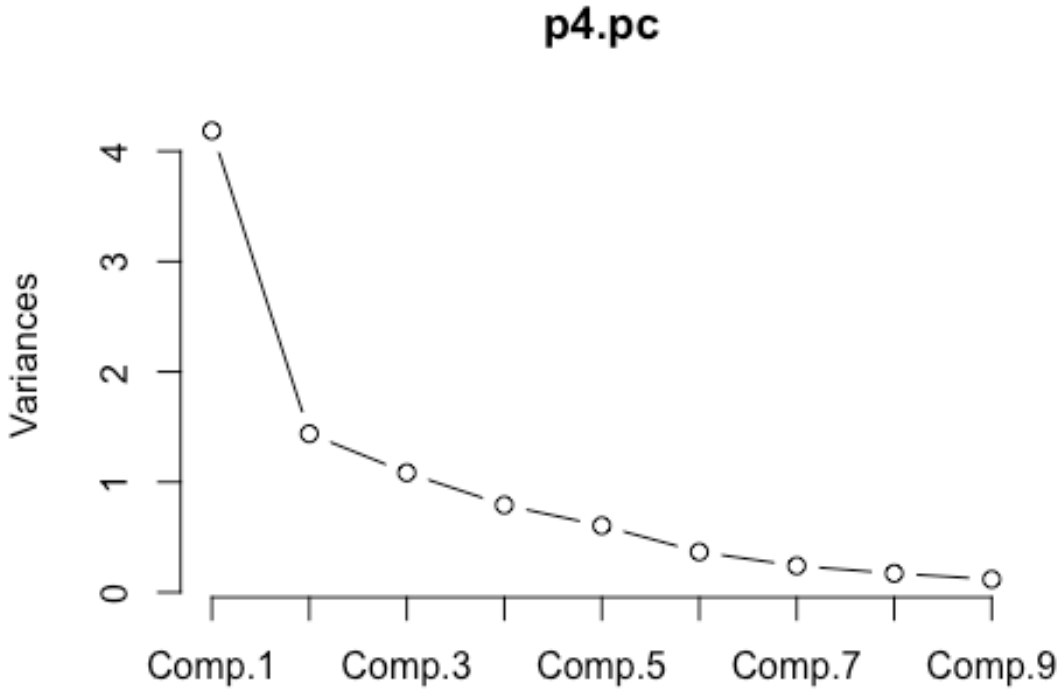
```
p4 <- read.table("/Users/linyuxiang/R/MSA/mid/food.dat", header = T)
```

Survey data were collected as part of a study to assess options for enhancing food security through the sustainable use of natural resources in the West Africa. A total of n = 72 farmers were surveyed and observations on the nine variables. (food.dat)

x1 = Family (total number of individuals in household)
x2 = DistRd (distance in kilometers to nearest passable road)
x3 = Cotton (hectares of cotton planted in year 2000)
x4 = Maize (hectares of maize planted in year 2000)
x5 = Sorg (hectares of sorghum planted in year 2000)
x6 = Millet (hectares of millet planted in year 2000)
x7 = Bull (total number of bullocks or draft animals)
x8 = Cattle (total) x9 = Goats (total)

(b) Perform a principal component analysis using the correlation matrix R. Determine the number of components to effectively summarize the variability. Use the proportion of variation explained and a scree plot to aid in your determination. (15%)

```
p4.pc <- princomp(p4,cor=TRUE)
plot(p4.pc, type = "l")
```



- Ans: 從圖中可以看出在第六個主成分後的 Variance 幾乎趨近於零，因此我會選擇 6 個主成分作為模型的變項。

(c) Interpret the first five principal components. Can you identify, for example, a "farm-size" component? A, perhaps, "goats and distance to road" component? (10%)

```
princomp(p4,cor=TRUE) %>% summary(loadings = T)

## Importance of components:
##                           Comp.1    Comp.2   Comp.3     Comp.4    Co
mp.5
## Standard deviation     2.0457593 1.1992026 1.041393 0.88984136 0.777
3833
## Proportion of Variance 0.4650146 0.1597874 0.120500 0.08797974 0.067
1472
## Cumulative Proportion  0.4650146 0.6248020 0.745302 0.83328173 0.900
4289
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.60509161 0.48992202 0.41451805 0.34373679
## Proportion of Variance 0.04068176 0.02666929 0.01909169 0.01312833
## Cumulative Proportion  0.94111069 0.96777998 0.98687167 1.00000000
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.
9
## Family  0.434                0.171                0.797  0.263  0.24
9
## DistRD        -0.497 -0.569  0.496 -0.378 -0.187

## Cotton  0.446         0.132        -0.219  0.200 -0.361 -0.329  0.67
5
## Maize   0.352 -0.353  0.388  0.240         0.273        -0.363 -0.57
4
## Sorg    0.204  0.604 -0.111        -0.645 -0.246        -0.126 -0.29
3
## Millet  0.240  0.415 -0.116  0.616  0.527 -0.181 -0.241

## Bull    0.445               -0.146         0.134 -0.396  0.751 -0.19
0
## Cattle  0.355 -0.284        -0.373  0.218 -0.759        -0.169

## Goats   0.255        -0.687 -0.351  0.249  0.402  0.131 -0.274 -0.14
9
```

- Ans: 第一個因子由變數 Family, Cotton, Bull 貢獻做多，是家戶人數、棉花種植面積、牛隻圈養數量的相關因子； 第二個因子主要由 Sorg, Millet, DistRD 貢獻最多，其中 DistRd 是反向的關係，是 Sorg 種植面積、Millet 種植面積以及與最近道路距離相關的因子； 第三個因子是 DistRd, Goats 為主要貢獻，兩者均

為負向關係，是與最近道路距離以及總 Goat 圈養數量相關的因子； 第四個因子主要由 DistRd, Millet 貢獻，兩者均為正向關係，是與最近道路距離以及 Millet 的種植面積相關的因子； 第五個因子主要由 Sorg, Millet 構成，其中 Sorg 為負向關係，該因子應該是和 Sorg 種植面積負相關，和 Millet 種植面積正相關的因子。