다변량자료분석 및 실습 Homework 3

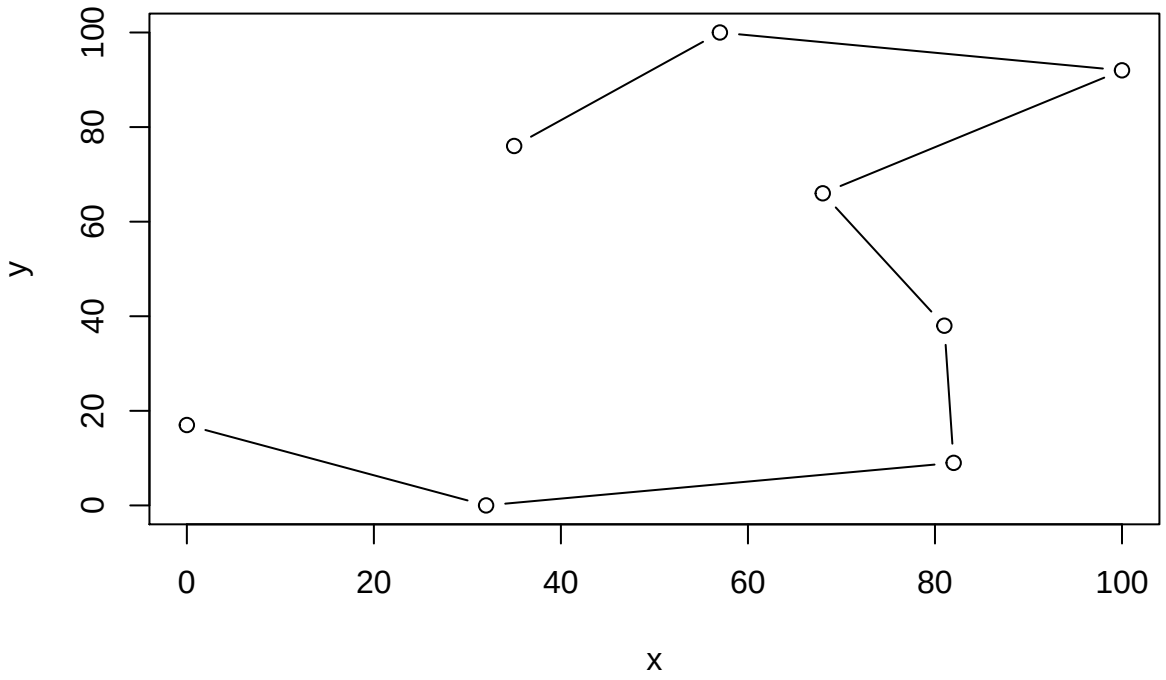
서울대학교 통계학과 2017-11362 박건도

2021년 11월 15일

Problem 3

(a)

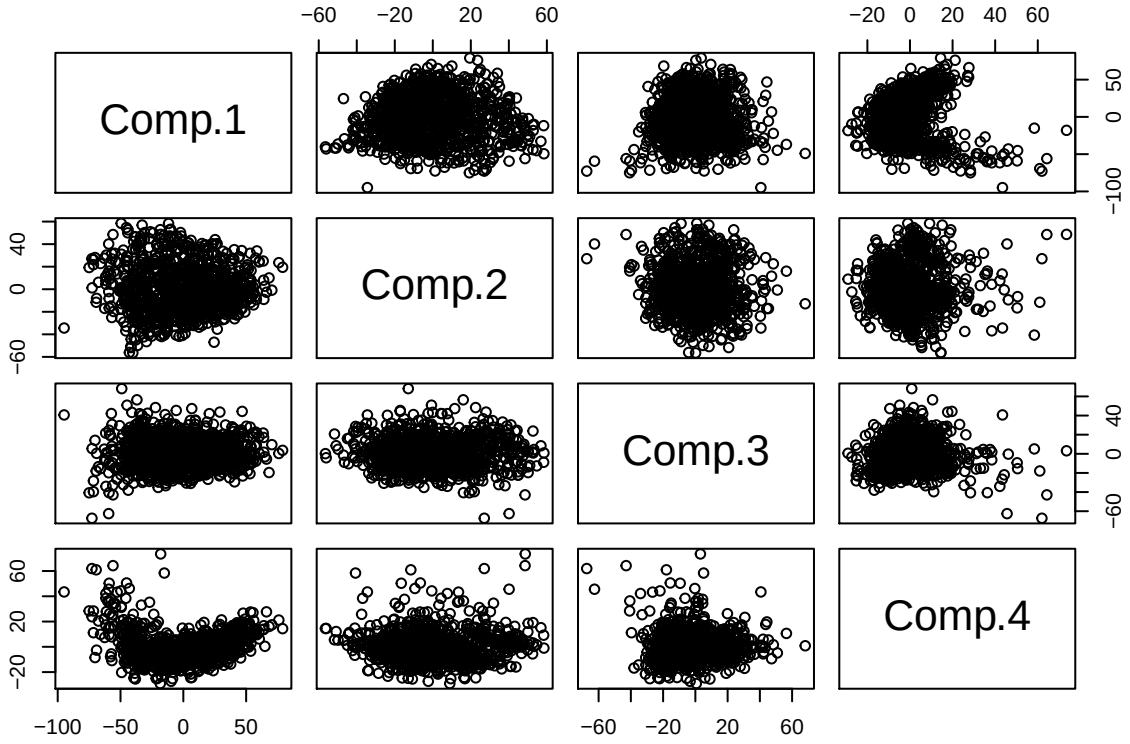
```
draw_digit <- function(pts, main=""){
  x <- pts[seq(1,15,2)]
  y <- pts[seq(2,16,2)]
  plot(unlist(x),unlist(y),'b', xlab="x", ylab="y", main=main)
}

draw_digit(pendigit3[1,])
```



(b)

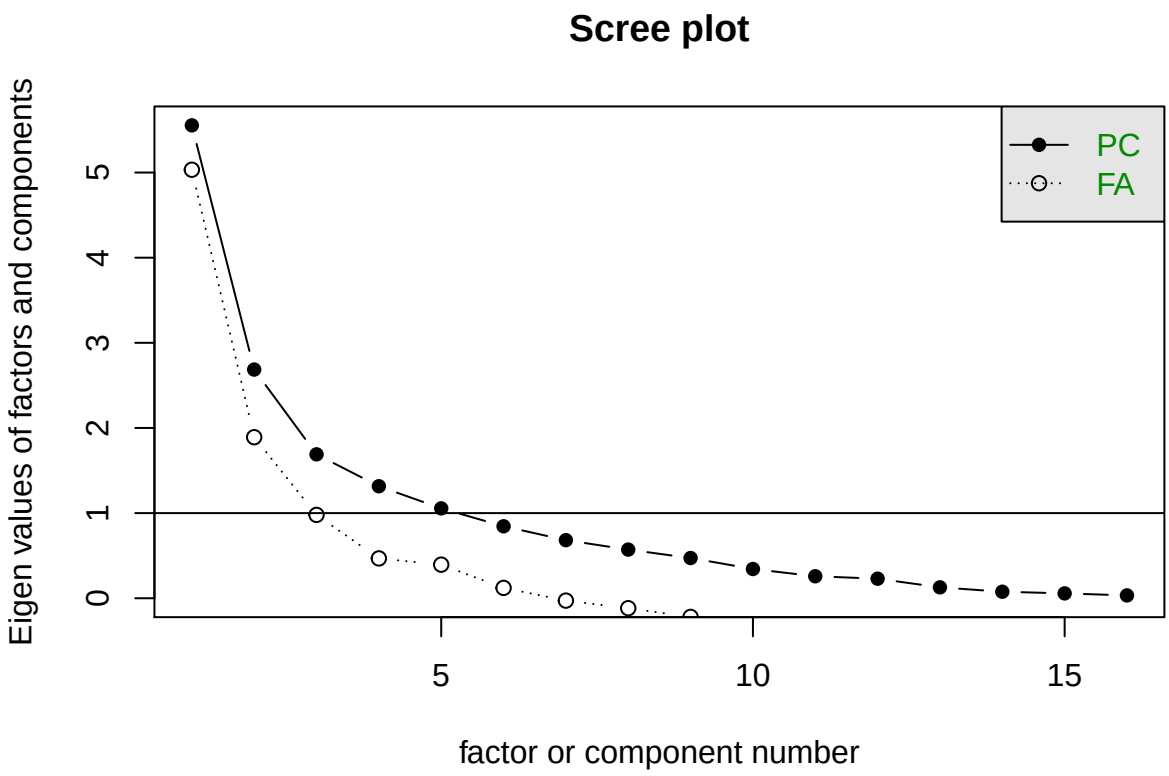
(1) scatterplot matrix

```
x <- pendigit3[,-17]
spr <- princomp(x)
pairs(spr$scores[,1:4])
```



(2) scree plot

```
psych::scree(x)
```



(c) MVN

```
mvnormtest::mshapiro.test(t(x))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.26799, p-value < 2.2e-16
```

p-value 〈 2.2e-16, x is not MVN. Also, scatter plot of principal components seems that x is not MVN.

(d)

In scree plot, 4~5 principal components explains most parts of original data(that point is elbow). I will choose 4 PCs because 4 is more easy to interpret than 5.

(e)

```
spr$loadings[,1:4]
```

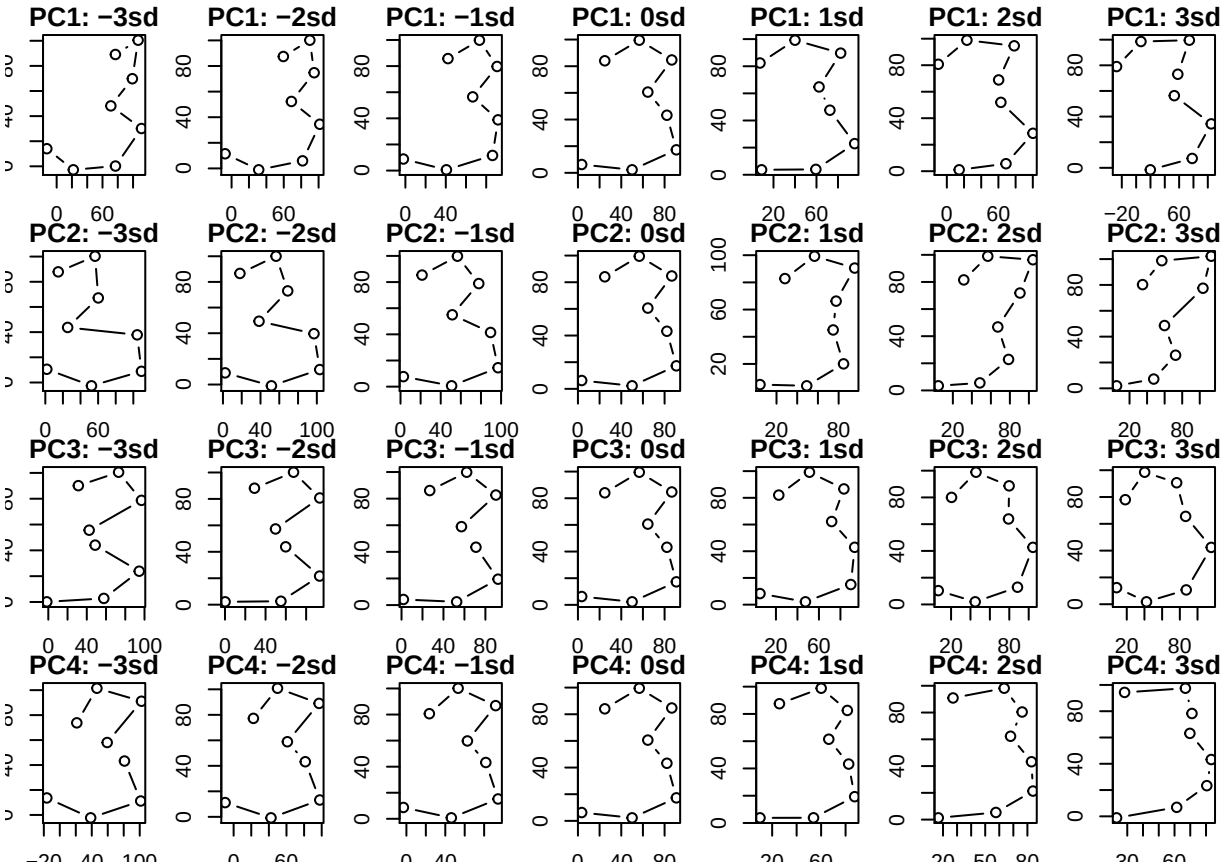
```
##           Comp.1      Comp.2      Comp.3       Comp.4
## V1   0.56675198  0.16095487  0.13441149  0.0980077234
```

```
## V2   0.05476928 -0.06085538  0.12634327  0.2862525847
## V3   0.54355798  0.00209875  0.34258930  0.2858999375
## V4   0.01043010 -0.01250342  0.02111939 -0.0562747673
## V5   0.13508717  0.42222362  0.21337798 -0.4343041699
## V6  -0.16231465  0.27985185 -0.12431921 -0.1789793698
## V7   0.06944982  0.62464361 -0.46038723  0.1460671818
## V8  -0.13536517  0.26917960 -0.10366929  0.0711567058
## V9   0.31431909 -0.35543330 -0.69736611  0.0333726022
## V10 -0.14256123  0.08571592  0.01705454  0.0005000155
## V11 -0.15166906 -0.29370657  0.07577042 -0.2896284850
## V12 -0.18538719  0.13394666  0.13740251  0.1698718679
## V13 -0.30683501 -0.03882578  0.16063624  0.3172710736
## V14 -0.05497909  0.07839996  0.01022358  0.1279395364
## V15 -0.17789608  0.02840049 -0.10493047  0.5506046169
## V16  0.08471901 -0.06823769 -0.12774363 -0.2060235434
```

- PC1 has highly related with V1, V3. These values are x-coordinates of first, second points.
- PC2 has highly related with V5, V7. These values are x-coordinates of third, fourth points.
- PC3 has highly related with V7, V9. These values are x-coordinates of fourth, fifth points.
- PC4 has highly related with V5, V15. These values are x-coordinates of third, last points.

For example, let's show the difference from the values of PCs.

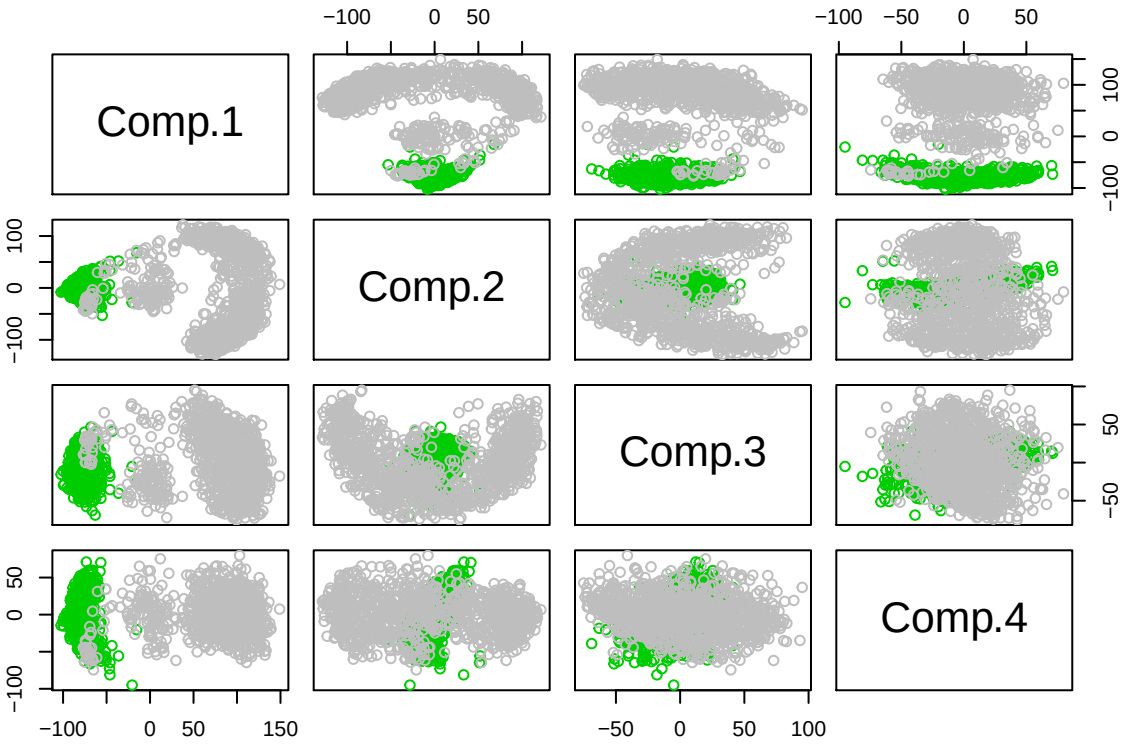
```
Xbar <- colMeans(x)
Xcentered <- x - matrix(Xbar, nrow=1055, ncol=16, byrow=TRUE)
xdec <- svd(Xcentered)
par(mfrow=c(4,7), mar=rep(1.5,4), oma=rep(0,4))
for (j in 1:4){
  for (k in -3:3){
    draw_digit(Xbar + k * xdec$d[j]*xdec$v[,j] / sqrt(1055),
               main = paste0("PC",j,": ",k,"sd"))
  }
}
```



- PC1 shows the position of 1st, 2nd points. (focus on volume of upper part / lower part)
- PC2 shows the position of 3rd, 4th points. (focus on position of upper part / lower part)
- PC3 shows the shape of 3. (focus on waist of 3.)
- PC4 shows the shape of 3. (focus on first-last points' positions)

(f)

```
pendigit8 <- read.table('pendigit8.txt',sep=",")
pendigit <- rbind(pendigit3, pendigit8)
pairs(princomp(pendigit[,-17])$scores[,1:4], col=pendigit[,17])
```



Some parts of digit 3 are mixed with digit 8 area. It is hard to separate nicely.