Q2. (a) MLE: $\hat{\Sigma} = S_0 = \begin{bmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{bmatrix}$. $\frac{n-1}{n} = \begin{bmatrix} 0.01406 & 0.01142 \\ 0.01142 & 0.01425 \end{bmatrix}$

(b) $\frac{n-p}{p} (\bar{x}-\mu)' S_0^{-1} (\bar{x}-\mu) \sim F_{p,n-p}$, $n=42$, $p=2$ $F_{2,40}$.

Find $A = \{ \mu \in \mathbb{R}^2 : (\bar{X}-\mu)' S_0^{-1} (\bar{X}-\mu) < \frac{2}{40} F_{1-0.05;2,40} = 0.1615 \}$

Set $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \Rightarrow R_1 = \{ \vec{\mu} \mid 26.844 - 32.909\mu_1 - 58.263\mu_2 - 326.572\mu_1\mu_2$

R_1 is 95% CI for μ. $+ 203.751\mu_1^2 + 201.033\mu_2^2 < 0.1615 \}$.

(c) $\mu = \begin{pmatrix} 0.6 \\ 0.58 \end{pmatrix} \Rightarrow LHS = 0.63 > 0.1615 \therefore \mu_1 \notin R_1$, not in R_1 confident interval

(d) $R_2 = \{ a^T\mu \mid a^T\bar{x} \pm \sqrt{h(\alpha) a^T S a} \}$, $h(\alpha) = \frac{41 \cdot 2}{40} F_{1-0.05;2,40} \doteq 6.615$

$\mu_2: a^T\bar{x} \pm \sqrt{h(\alpha) a^T S a} = 0.603 \pm 0.0484 \Rightarrow R_2 = (0.555, 0.651) \Rightarrow R_2 = \begin{bmatrix} 0.516, 0.612 \\ 0.555, 0.651 \end{bmatrix}$

(e) μ_0 known, Σ unknown $\Rightarrow W = n\log|S_0 + \delta\delta'| - n\log|S_0|$

$\delta^T S^{-1} \delta = 42 \cdot \begin{pmatrix} 0.555-0.6 \\ 0.651-0.58 \end{pmatrix} \begin{pmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{pmatrix}^{-1} \begin{pmatrix} \cancel{0.495} -0.045 \\ 0.071 \end{pmatrix}^T = 26.3$

$P(F_{2,40} \geq 26.3) = 0.000 < 0.05$ $\therefore$ Thus we can reject H_0 and accept $\mu = \mu_0$
p-value

Q3 (a) $\bar{X}_n \sim N_p(\mu, \frac{\Sigma}{n}) \Rightarrow a_i^T\bar{X}_n \sim N(a_i^T\mu, \frac{a_i^T\Sigma a_i}{n})$ $\left(\frac{a_i^T\bar{X}_n - a_i^T\mu}{\sqrt{\frac{a_i^T\Sigma a_i}{n}}} \right) \sim N(0,1)$ 0.

$\frac{\sqrt{n}\left(\frac{a_i^T\bar{x}_n - a_i^T\mu}{\sqrt{a_i^T\Sigma a_i}}\right)}{\sqrt{\frac{a_i^T S a_i}{a_i^T\Sigma a_i}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1) \overset{def}{\Rightarrow} P\left(T = \frac{\sqrt{n}(a_i^T\bar{x}_n - a_i^T\mu)}{\sqrt{a_i^T S a_i}} \,\middle|\, T \in [-t_{n-1}(\frac{\alpha_i}{2}), t_{n-1}(\frac{\alpha_i}{2})] \right) = 1-\alpha_i$

$\Rightarrow C_i(\alpha_i)$ is a confidence Interval of α_i

(b) $m=3$ $P(a_i^T\mu \in C_i, i=1,2,3) = 1 - P(\text{at least } 1\ i, a_i^T\mu \notin C_i)$ $\cancel{\geq 1 - \sum P(a_i^T\mu \notin C_i)}$

$= 1 - P(\bigcup \{a_i^T\mu \notin C_i\}) \geq 1 - \sum P(a_i^T\mu \notin C_i) = 1 - (\alpha_1 + \alpha_2 + \alpha_3)$

(c) By lose (b), we have $P(a_i^T\mu \in C_i(\frac{\alpha}{m}), \forall i) \geq 1 - (\underbrace{\frac{\alpha}{m} + \frac{\alpha}{m} + \cdots}_{\#m}) = 1 - \alpha$

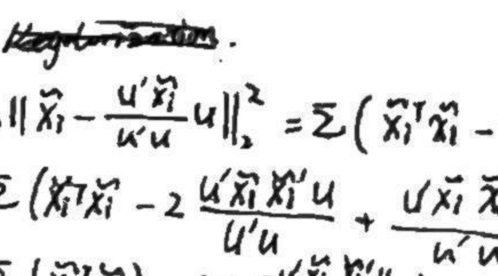
(d) $m=2$, $I_{\mu_1} = \cancel{0.555}$ $(1,0)\begin{pmatrix} 0.564 \\ 0.603 \end{pmatrix} \pm \underbrace{t\left(\frac{0.05}{2\times 2}\right)}_{2.326} \cdot \left(\sqrt{\frac{(1,0)S\binom{1}{0}}{n}}\right) = 0.564 \pm 0.043$

$I_{\mu_1} = (0.521, 0.607)$

$I_{\mu_2} = (0,1)\begin{pmatrix} 0.564 \\ 0.603 \end{pmatrix} \pm 2.306 \cdot \sqrt{\frac{(0,1)S(0,1)^T}{42}} = 0.603 \pm 0.043 = (0.560, 0.646)$

I_{μ_1}, I_{μ_2} are confidence interval of μ $R_3 = \begin{pmatrix} I_{\mu_1} \\ I_{\mu_2} \end{pmatrix}$

(e)



R_3, precise and easy for calculate.

Q4: ~~Regularization~~.

$SSR = \sum \|\tilde{x}_i - \frac{u'\tilde{x}_i}{u'u} u\|_2^2 = \sum\left(\tilde{x}_i^T\tilde{x}_i - \tilde{x}_i^T\frac{u'\tilde{x}_i}{u'u}u - \frac{\tilde{x}_i^T u' \tilde{x}_i}{u'u} u + \frac{u'\tilde{x}_i u u'\tilde{x}_i u}{u'u u'u} \right)$

$= \sum\left(\tilde{x}_i^T\tilde{x}_i - 2\frac{u'\tilde{x}_i \tilde{x}_i' u}{u'u} + \frac{u'\tilde{x}_i \tilde{x}_i' u}{u'u} \right)$

$= \sum(\tilde{x}_i^T\tilde{x}_i) - \sum\left(\frac{u'\tilde{x}_i\tilde{x}_i'u}{u'u}\right)$

$\underset{\min}{\cancel{\max}}\ SSR \iff \max \sum\left(\frac{u'\tilde{x}_i\tilde{x}_i'u}{u'u}\right)$, i.e. max Sample Variance: $\max \frac{1}{n-1}\sum\left(u'\tilde{x}_i - \frac{\sum u'\tilde{x}_i}{n}\right)^2$

$= \frac{1}{n-1}\sum\left(u'\tilde{x}_i - \underbrace{u'\frac{\sum\tilde{x}_i}{n}}_{=0}\right)^2 = \frac{1}{n-1}\sum(u'\tilde{x}_i)^2$

$\therefore$ max Sample Var $\iff$ min RSS.

STAT 2221 Multivariate: HW 3

Ji Changcheng

Spring 2024

(a) Write down the equation for the logistic regression model of LI on remission in cancer patients (using the parameters β0 and β1).

$\pi(x) = exp(\beta_0 + \beta_1 x)/(1 + exp(\beta_0 + \beta_1 x))$ The parameters are obtained from R code result
$\beta_0 = -3.77714, \beta_1 = 0.14486$

Question 1

```
df <- read.csv("homeless_smalldata.csv", header=T,row.names=1)
```

(a)

$H_0 : \mu_1 = \mu_2$
$H_\alpha : \mu_1 \neq \mu_2$

(b) Determine the dimension p and the sample sizes n1, n2.

```
cat("Dimension p:", ncol(df)-1, "\n")
```

```
## Dimension p: 3
```

```
cat("Sample size nonhomeless n1 (homeless=0):", sum(df$homeless == 0), "\n")
```

```
## Sample size nonhomeless n1 (homeless=0): 244
```

```
cat("Sample size homeless n2 (homeless=1):", sum(df$homeless == 1), "\n")
```

```
## Sample size homeless n2 (homeless=1): 209
```

(c) Conduct the test.

```
non_homeless_data <- df[df$homeless == 0, ]
homeless_data <- df[df$homeless == 1, ]
p <- 3
n1 <- nrow(non_homeless_data)
n2 <- nrow(homeless_data)


mean1 <- colMeans(non_homeless_data[,c(1,2,3)])
mean2 <- colMeans(homeless_data[,c(1,2,3)])
covmat <- cov(df[,c(1,2,3)])

cat("Means of non homeless: ",mean1,"\n")
```

```
## Means of non homeless:  49.00083 32.48683 31.84016
```

```
cat("Means of homeless: ",mean2,"\n")
```

```
## Means of homeless:  46.93678 30.73085 34.02392
```

```
cat("Covariance of vars: \n")
```

```
## Covariance of vars:
```

```
covmat
```

```
##             pcs        mcs       cesd
## pcs  116.30766   15.29467  -39.50384
## mcs   15.29467  164.84858 -109.56974
## cesd -39.50384 -109.56974  156.61170
```

```
tval <- (n1+n2-p-1)*(n1*n2)/((n1+n2)*p*(n1+n2-2))* t(mean1-mean2) %*% solve(covmat) %*% (mean1-mean2)
cat("T statistic = ",tval,"\n")
```

```
## T statistic =  2.012177
```

```
p_value <- 1 - pf(tval, p, n1+n2)
cat("p value = ",p_value,"\n")
```

```
## p value =  0.1114513
```

p-value > .05, thus we can not reject the H0 that they have the same means.

(d) Scatter points

```
library(dplyr)
```

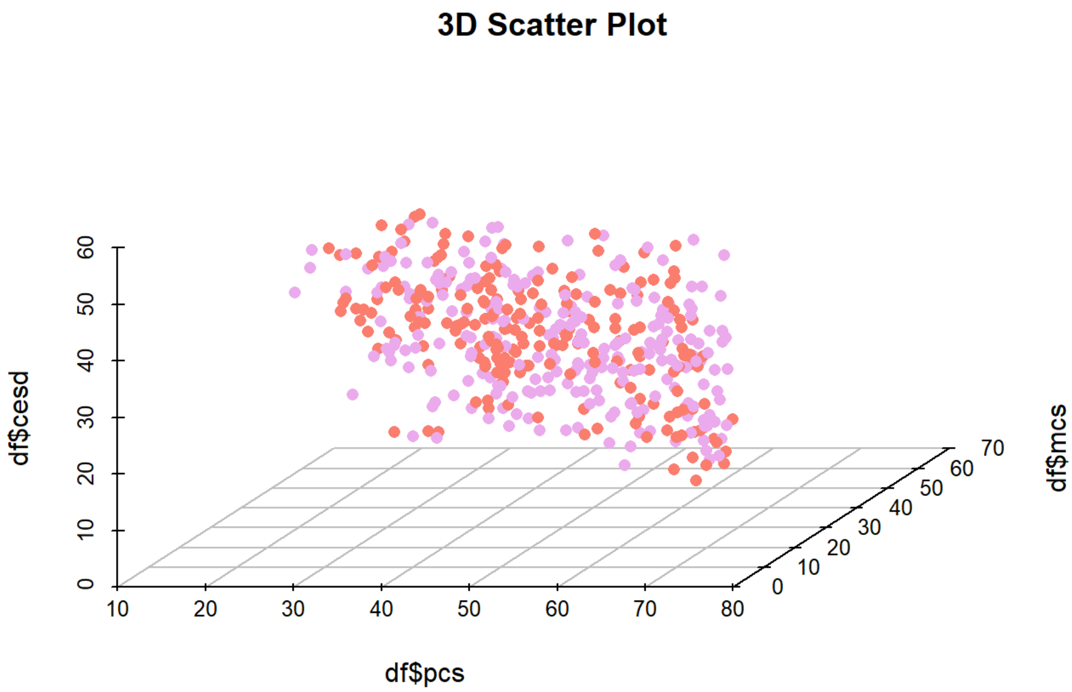
```
## Warning: 程辑包'dplyr'是用R版本4.3.2 来建造的
```

```
##
## 载入程辑包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scatterplot3d)
colors <- c("plum2", "salmon")
colors <- colors[as.factor(df$homeless)]
scatterplot3d(df$pcs, df$mcs, df$cesd,pch = 16, main = "3D Scatter Plot", angle = 45, scale.y = 0.7, pty = "s", box = FALSE,
              color=colors)
```

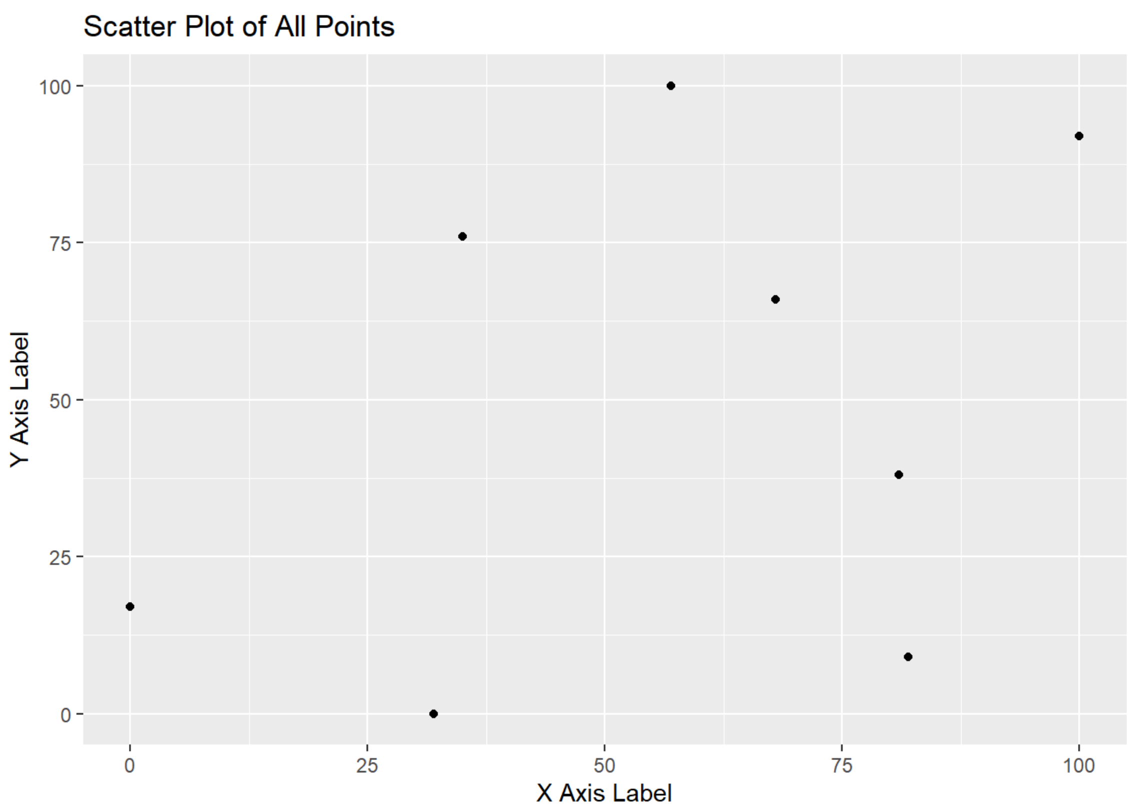


Data are generally distributed in the same way.

Question 4

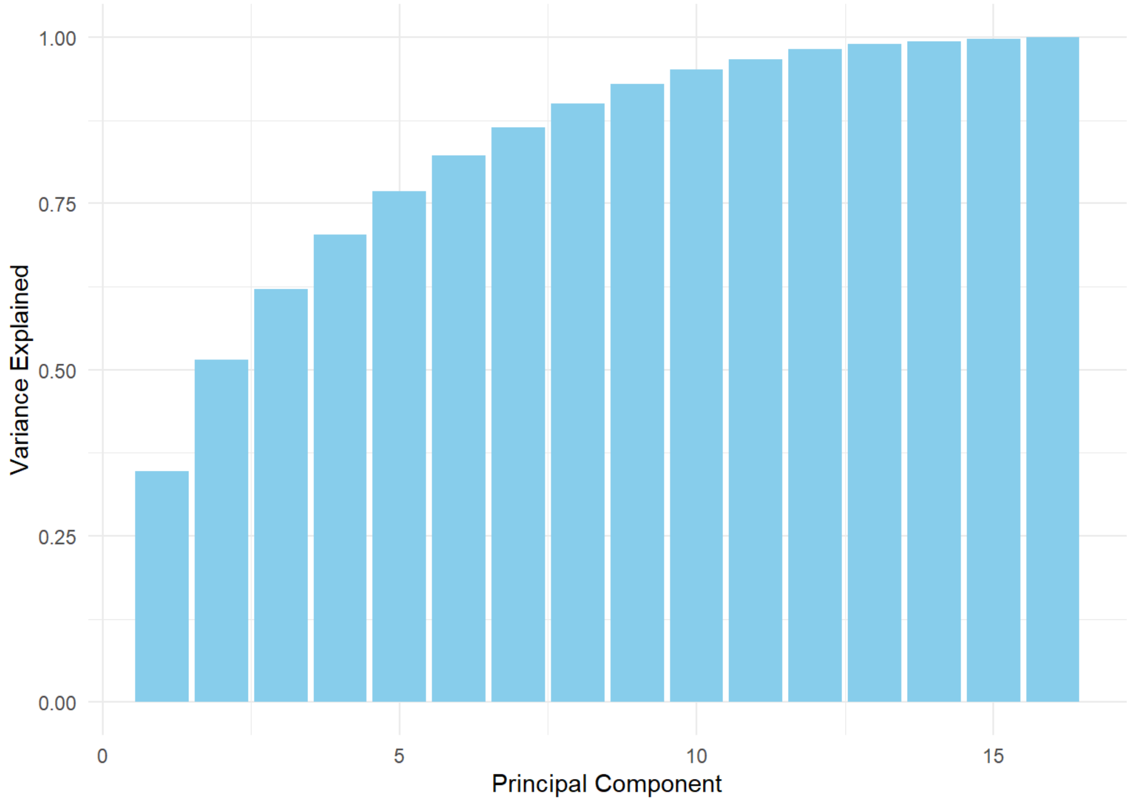
(a)

```
library(ggplot2)
df2 <- read.table("pendigit3.txt", sep = ",")
df2 <- df2[, -17]
obs<-data.frame(x = as.numeric(df2[1, seq(1, 16, by = 2)]), y = as.numeric(df2[1, seq(2, 16, by = 2)]))
ggplot(obs, aes(x = x, y = y)) +
  geom_point() +
  ggtitle("Scatter Plot of All Points") +
  xlab("X Axis Label") +
  ylab("Y Axis Label")
```

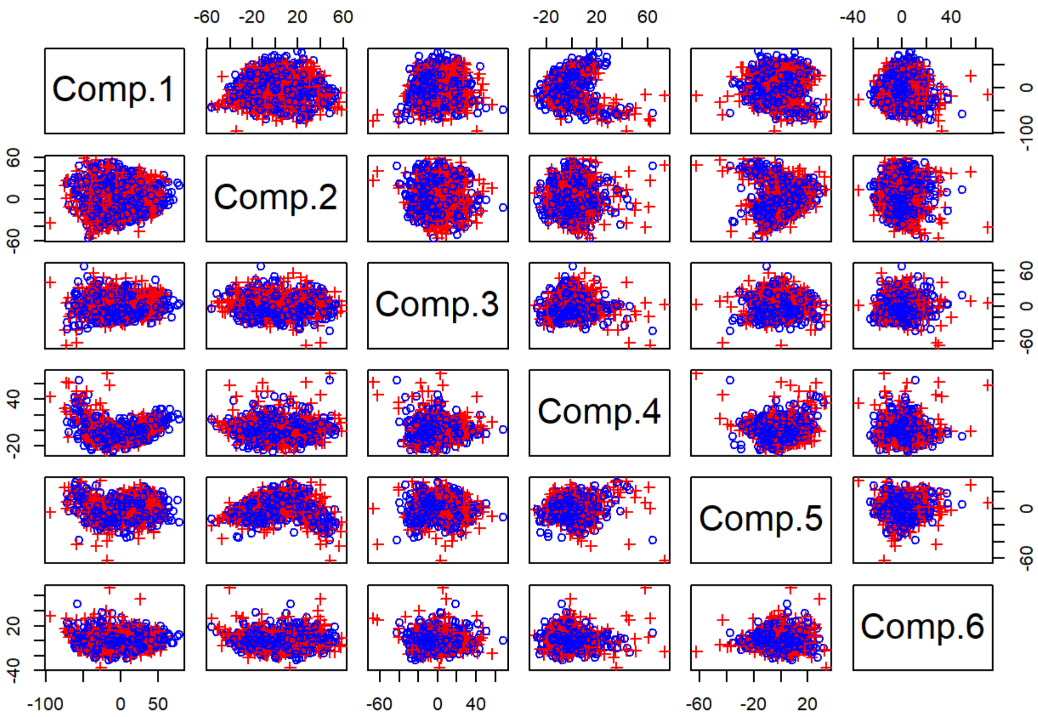


(b)

```
scaled_data <- scale(df2)

# Step 2: Perform PCA
pca_result <- prcomp(scaled_data)

v <- pca_result$sdev^2
variance_explained <- round(cumsum(v/sum(v)),3)
pca_data <- data.frame(
  PC = 1:length(variance_explained),
  VarianceExplained = variance_explained
)

ggplot(pca_data, aes(x = PC, y = variance_explained)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(x = "Principal Component", y = "Variance Explained") +
  theme_minimal()
```




```
pca_result<-princomp(df2)
Z <-pca_result$scores

# scatterplot of principal components
pairs(Z[,c(1,2,3,4,5,6)],pch=c(rep(1,100),rep(3,100)),col=c(rep("blue",100),rep("red",100)))
```



(c) The result shows strong randomness so that it is likely to be a MVN distribution

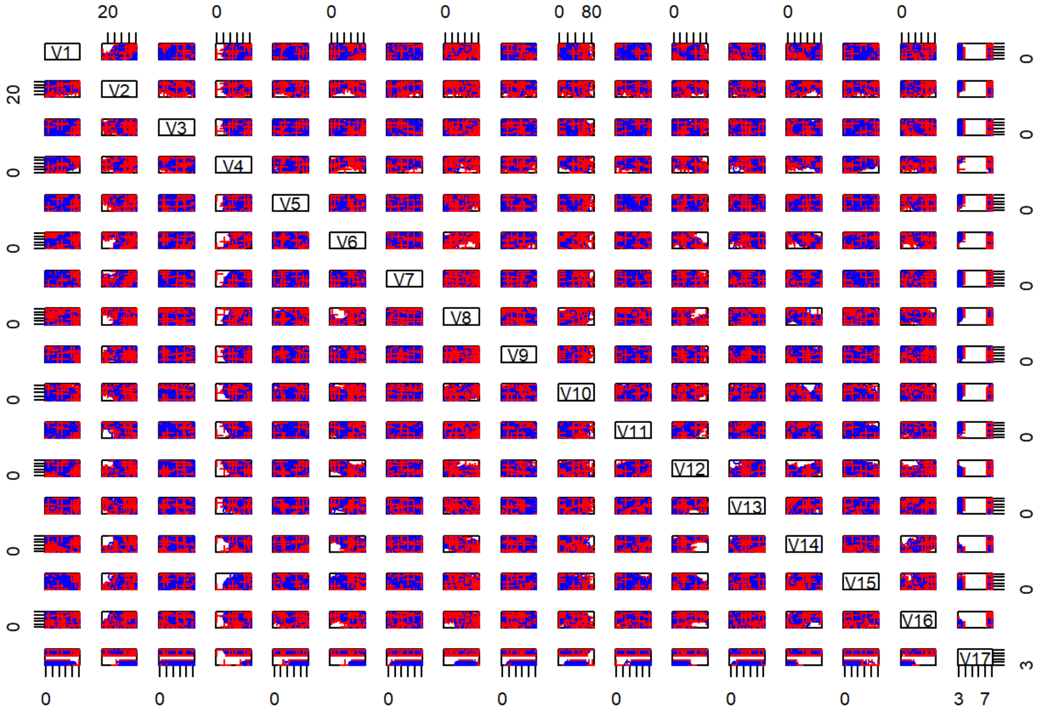
(d) I prefer to keep 6 as the explained variance reach 80%

(e) PC1 is the linear combination that captures the maximum variance in the data. Along the direction of PC1, the variability of the data is maximized. PC2, orthogonal to PC1, have the next highest variance. Along the direction of PC2, the variability of the data is second highest.

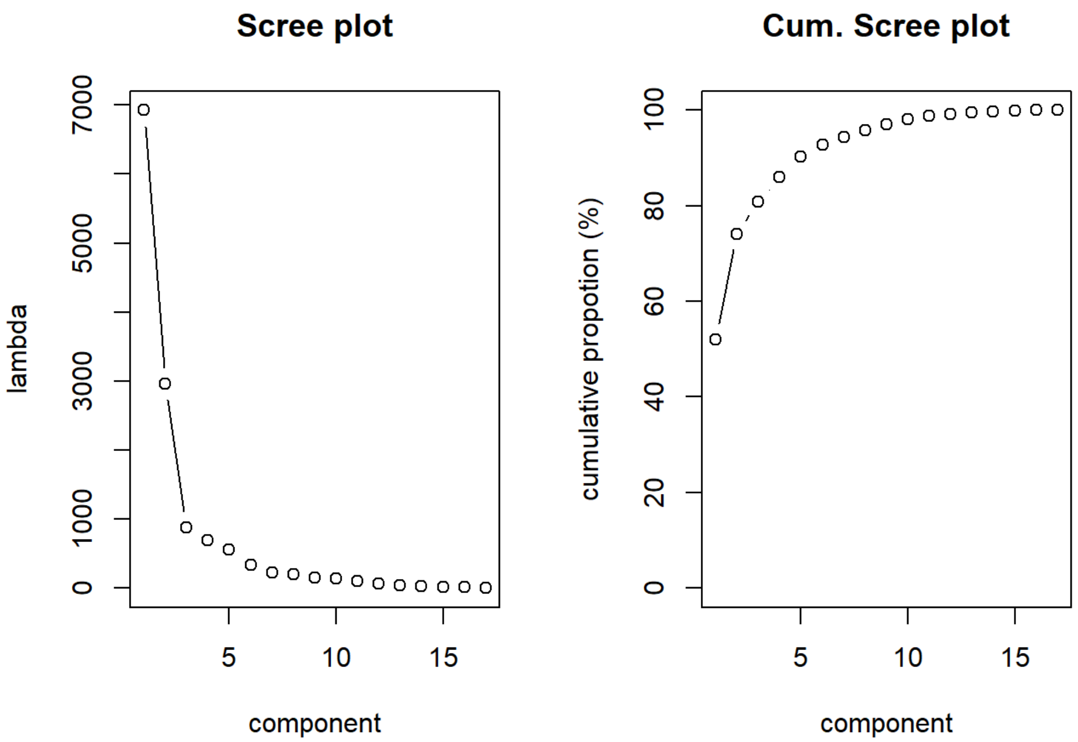
(f) From the graph, there is a nice separation between “3” and “8”.

```
df3 <- rbind(read.table("pendigit3.txt", sep = ","), read.table("pendigit8.txt", sep = ","))
```

```
spr <-princomp(df3)
U<-spr$loadings
L<-(spr$sdev)^2
Z <-spr$scores

# scatterplot of principal components
pairs(df3,pch=c(rep(1,100),rep(3,100)),col=c(rep("blue",100),rep("red",100)))
```



```
par(mfrow=c(1,2))
plot(L,type="b",xlab="component",ylab="lambda",main="Scree plot")
plot(cumsum(L)/sum(L)*100,ylim=c(0,100),type="b",xlab="component",ylab="cumulative propotion (%)",main="Cum. Scree plot")
```




```
# biplot
par(mfrow=c(1,1))
```