Assignment 02

Answers start at the second page, this page only shows how I loaded the data.

```r
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

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

```r
library(matrixStats)
```

```
##
## Attaching package: 'matrixStats'

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

```r
survival <- read_excel("C:/Users/John/Desktop/STAT 445/Data/survival_data.xlsx", col_names = F)
```

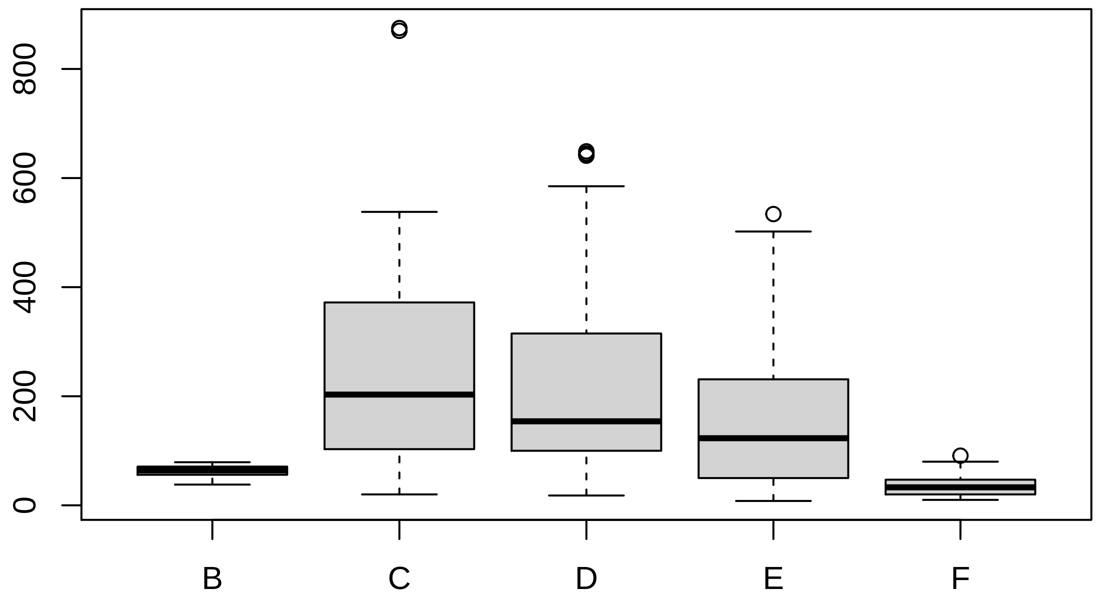
```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ...
```

```r
colnames(survival)=c("Group","B","C","D","E","F")
```

Question 01

(a) Boxplot of the data

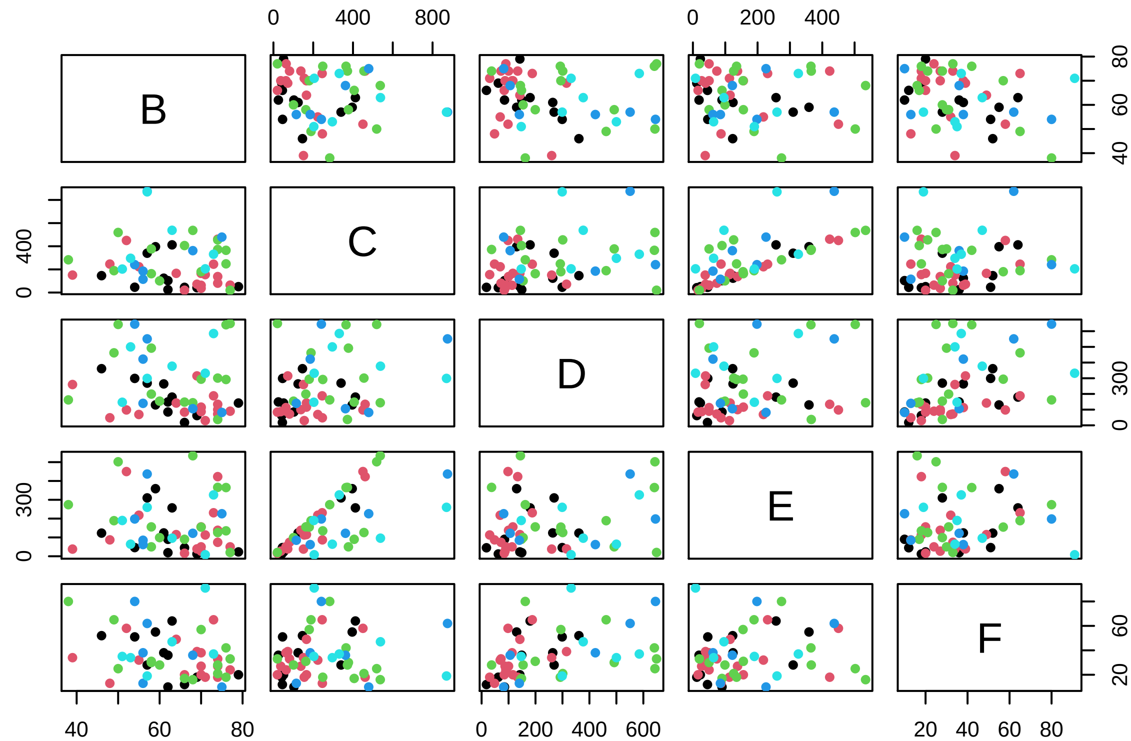
```
boxplot(survival[,2:6])
```



Scale can be an issue as value ranges in B and F are much smaller than the rest of the group. And in group C, D, and E there are values that are significantly larger than the rest of the group, means those outliers can potentially affect the correlations.

(b) Bivariate scatter matrix plot

```
plot(survival[,2:6], col=survival$Group, pch = 16)
```



Legend: B = black, C = red , D = green, E = blue, F = cyan

(c) Sample mean vector :

$$\tilde{\bar{x}} = \begin{pmatrix} 63.24528 \\ 252.50943 \\ 239.79245 \\ 165.43396 \\ 35.69811 \end{pmatrix} \tag{1}$$

Sample standard deviation vector:

$$\tilde{s} = \begin{pmatrix} 10.15629 \\ 195.90614 \\ 182.24427 \\ 139.15425 \\ 19.21396 \end{pmatrix} \tag{2}$$

(d)Matrix $\bar{X}^*$ of samples means using the standardized values

$$\bar{X}^* = \begin{bmatrix} -0.15841606 & -0.4861344 & -0.3385651 & -0.26901056 & -0.04106505 \\ 0.20969439 & -0.4154256 & -0.6484015 & -0.18951245 & -0.18271157 \\ 0.06024443 & 0.2478840 & 0.4941976 & 0.38237349 & -0.03261611 \\ -0.22107316 & 0.6312065 & 0.4638877 & 0.16456107 & 0.21521954 \\ -0.18825277 & 0.7894456 & 0.7327576 & -0.05821331 & 0.42340149 \end{bmatrix} \tag{3}$$

(e) Distance matrix $\bar{D^*}$

$$\bar{D^*} = \begin{bmatrix} 0.0000000 & 0.5127254 & 1.3055503 & 1.4662787 & 1.7423809 \\ 0.5127254 & 0.0000000 & 1.4551391 & 1.6738889 & 1.9754213 \\ 1.3055503 & 1.4551391 & 0.0000000 & 0.5795336 & 0.9022303 \\ 1.4662787 & 1.6738889 & 0.5795336 & 0.0000000 & 0.4374654 \\ 1.7423809 & 1.9754213 & 0.9022303 & 0.4374654 & 0.0000000 \end{bmatrix} \tag{4}$$

(f) From the distance matrix we can see that, there is a relatively low separation between group 1 and 2, group 3 and 4, group 4 and 5. There is a relatively large separation between group 1 and 5, group 2 and 4, group 2 and 5.

Code used to solve (c), (d), (e)

```
# c
sample_mean <- colMeans(survival[,2:6])
sample_mean
```

```
##         B         C         D         E         F
##  63.24528 252.50943 239.79245 165.43396  35.69811
```

```
sample_sd <- colSds(as.matrix(survival[,2:6]))
sample_sd
```

```
## [1]  10.15629 195.90614 182.24427 139.15425  19.21396
```

```
# d
X_bar <- matrix(data = NA, nrow = 5,ncol = 5)
for (i in 1:5) {
  group_data <- filter(survival, Group == i)
  X_bar[i,] <- colMeans(group_data[,2:6])
}
X_star <- matrix(data=NA, nrow=5, ncol=5)
for (j in 1:5) { # j represents each variable of X1,..,X9
  X_star[,j] <- (X_bar[,j]-sample_mean[j])/sample_sd[j]
}
X_star
```

```
##             [,1]       [,2]       [,3]        [,4]        [,5]
## [1,] -0.15841606 -0.4861344 -0.3385651 -0.26901056 -0.04106505
## [2,]  0.20969439 -0.4154256 -0.6484015 -0.18951245 -0.18271157
## [3,]  0.06024443  0.2478840  0.4941976  0.38237349 -0.03261611
## [4,] -0.22107316  0.6312065  0.4638877  0.16456107  0.21521954
## [5,] -0.18825277  0.7894456  0.7327576 -0.05821331  0.42340149
```

```
# e
dist(X_star, method="euclidean",diag=T,upper=T)
```

```
##           1         2         3         4         5
## 1 0.0000000 0.5127254 1.3055503 1.4662787 1.7423809
## 2 0.5127254 0.0000000 1.4551391 1.6738889 1.9754213
## 3 1.3055503 1.4551391 0.0000000 0.5795336 0.9022303
## 4 1.4662787 1.6738889 0.5795336 0.0000000 0.4374654
## 5 1.7423809 1.9754213 0.9022303 0.4374654 0.0000000
```

Question 2

(a) $A = corr(B)$

$$A = \begin{bmatrix} 1.0000000 & 0.9410886 & 0.8707802 & 0.8091758 & 0.7815510 & 0.7278784 & 0.6689597 \\ 0.9410886 & 1.0000000 & 0.9088096 & 0.8198258 & 0.8013282 & 0.7318546 & 0.6799537 \\ 0.8707802 & 0.9088096 & 1.0000000 & 0.8057904 & 0.7197996 & 0.6737991 & 0.6769384 \\ 0.8091758 & 0.8198258 & 0.8057904 & 1.0000000 & 0.9050509 & 0.8665732 & 0.8539900 \\ 0.7815510 & 0.8013282 & 0.7197996 & 0.9050509 & 1.0000000 & 0.9733801 & 0.7905565 \\ 0.7278784 & 0.7318546 & 0.6737991 & 0.8665732 & 0.9733801 & 1.0000000 & 0.7987302 \\ 0.6689597 & 0.6799537 & 0.6769384 & 0.8539900 & 0.7905565 & 0.7987302 & 1.0000000 \end{bmatrix} \tag{5}$$

(b) $det(A) = 9.011147 \times 10^{-6}$
(c) Since A is the correlation matrix of the data, as $det(A)$ is getting closer to 0, it shows there's a stronger relationship between the variables. When A is an identity matrix, that is, all the correlations equals $0(r_{ij} = 0$ for $i \neq j)$, $det(A)$ has the maximum value at 1.
(d)

$$A^{-1}\tilde{b} = \begin{pmatrix} 0.29123846 \\ 0.01544922 \\ 0.41095061 \\ -0.31259312 \\ -0.17544947 \\ 0.56607556 \\ 0.46999267 \end{pmatrix} \tag{6}$$

(e)

$$\tilde{y} = \begin{pmatrix} 144.4946 \\ 146.5773 \\ 140.7780 \\ 150.9731 \\ 148.8515 \\ 143.8101 \\ 135.8276 \end{pmatrix} \tag{7}$$

(f) Projection of $\tilde{x}$ onto $\tilde{y}$

$$\begin{pmatrix} 0.1270371 \\ 0.1288682 \\ 0.1237696 \\ 0.1327329 \\ 0.1308677 \\ 0.1264354 \\ 0.1194173 \end{pmatrix} \tag{8}$$

Code used to compute the answers for (a) to (f):

```
B <- read_excel("C:/Users/John/Desktop/STAT 445/Data/w-nat-track-rec.xlsx", col_names = F)
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ...
```

```
# a
A <- cor(B)
A
```

```
##           ...1      ...2      ...3      ...4      ...5      ...6      ...7
## ...1 1.0000000 0.9410886 0.8707802 0.8091758 0.7815510 0.7278784 0.6689597
## ...2 0.9410886 1.0000000 0.9088096 0.8198258 0.8013282 0.7318546 0.6799537
## ...3 0.8707802 0.9088096 1.0000000 0.8057904 0.7197996 0.6737991 0.6769384
## ...4 0.8091758 0.8198258 0.8057904 1.0000000 0.9050509 0.8665732 0.8539900
## ...5 0.7815510 0.8013282 0.7197996 0.9050509 1.0000000 0.9733801 0.7905565
## ...6 0.7278784 0.7318546 0.6737991 0.8665732 0.9733801 1.0000000 0.7987302
## ...7 0.6689597 0.6799537 0.6769384 0.8539900 0.7905565 0.7987302 1.0000000
```

```
# b
det(A)
```

```
## [1] 9.011147e-06
```

```
# d
b = c(1,1,1,1,1,1,1)
solve(A)%*%b
```

```
##             [,1]
## ...1  0.29123846
## ...2  0.01544922
## ...3  0.41095061
## ...4 -0.31259312
## ...5 -0.17544947
## ...6  0.56607556
## ...7  0.46999267
```

```
# e
x = c(1,-1,1,-1,1,-1,1)
y <- A%*%A%*%A%*%A%*%x
y
```

```
##          [,1]
## ...1 144.4946
## ...2 146.5773
```

```
## ...3 140.7780
## ...4 150.9731
## ...5 148.8515
## ...6 143.8101
## ...7 135.8276
```

```r
# f
# sum(x*y) is the dot product of x and y, sum(y^2) is the square of the norm of y
projection <-(sum(x*y)/sum(y^2))* y
projection
```

```
##           [,1]
## ...1 0.1270371
## ...2 0.1288682
## ...3 0.1237696
## ...4 0.1327329
## ...5 0.1308677
## ...6 0.1264354
## ...7 0.1194173
```

Question 3

(a) Please note that the columns in matrix C are ordered by their corresponding eigenvalues in descending order

$$C = \begin{bmatrix} -0.3777657 & -0.4071756 & -0.1405803 & 0.58706293 & -0.16706891 & 0.53969730 & 0.08893934 \\ -0.3832103 & -0.4136291 & -0.1007833 & 0.19407501 & 0.09350016 & -0.74493139 & -0.26565662 \\ -0.3680361 & -0.4593531 & 0.2370255 & -0.64543118 & 0.32727328 & 0.24009405 & 0.12660435 \\ -0.3947810 & 0.1612459 & 0.1475424 & -0.29520804 & -0.81905467 & -0.01650651 & -0.19521315 \\ -0.3892610 & 0.3090877 & -0.4219855 & -0.06669044 & 0.02613100 & -0.18898771 & 0.73076817 \\ -0.3760945 & 0.4231899 & -0.4060627 & -0.08015699 & 0.35169796 & 0.24049968 & -0.57150644 \\ -0.3552031 & 0.3892153 & 0.7410610 & 0.32107640 & 0.24700821 & -0.04826992 & 0.08208401 \end{bmatrix} \tag{9}$$

$$D = \begin{bmatrix} 5.807624 & 0.0000000 & 0.0000000 & 0.0000000 & 0.00000000 & 0.00000000 & 0.00000000 \\ 0.000000 & 0.6286934 & 0.0000000 & 0.0000000 & 0.00000000 & 0.00000000 & 0.00000000 \\ 0.000000 & 0.0000000 & 0.2793346 & 0.0000000 & 0.00000000 & 0.00000000 & 0.00000000 \\ 0.000000 & 0.0000000 & 0.0000000 & 0.1245547 & 0.00000000 & 0.00000000 & 0.00000000 \\ 0.000000 & 0.0000000 & 0.0000000 & 0.0000000 & 0.09097174 & 0.00000000 & 0.00000000 \\ 0.000000 & 0.0000000 & 0.0000000 & 0.0000000 & 0.00000000 & 0.05451882 & 0.00000000 \\ 0.000000 & 0.0000000 & 0.0000000 & 0.0000000 & 0.00000000 & 0.00000000 & 0.01430226 \end{bmatrix} \tag{10}$$

(b)

$$CC^T = \begin{bmatrix} 1.000000 & 0.0000000 & -0.0000000 & 0.0000000 & 0.00000000 & 0.00000000 & -0.00000000 \\ 0.000000 & 1.0000000 & -0.0000000 & 0.0000000 & 0.00000000 & 0.00000000 & 0.00000000 \\ -0.000000 & -0.0000000 & 1.0000000 & 0.0000000 & 0.00000000 & 0.00000000 & 0.00000000 \\ 0.000000 & 0.0000000 & 0.0000000 & 1.0000000 & 0.00000000 & -0.00000000 & 0.00000000 \\ 0.000000 & 0.0000000 & 0.0000000 & 0.0000000 & 1.00000000 & -0.00000000 & -0.00000000 \\ 0.000000 & 0.0000000 & 0.0000000 & -0.0000000 & -0.00000000 & 1.00000000 & -0.00000000 \\ -0.000000 & 0.0000000 & 0.0000000 & 0.0000000 & -0.00000000 & -0.00000000 & 1.00000000 \end{bmatrix} \tag{11}$$

(c)

$$C_1 = \begin{bmatrix} -0.3777657 & -0.4071756 & -0.1405803 & 0.58706293 \\ -0.3832103 & -0.4136291 & -0.1007833 & 0.19407501 \\ -0.3680361 & -0.4593531 & 0.2370255 & -0.64543118 \\ -0.3947810 & 0.1612459 & 0.1475424 & -0.29520804 \\ -0.3892610 & 0.3090877 & -0.4219855 & -0.06669044 \\ -0.3760945 & 0.4231899 & -0.4060627 & -0.08015699 \\ -0.3552031 & 0.3892153 & 0.7410610 & 0.32107640 \end{bmatrix} \tag{12}$$

$$D_1 = \begin{bmatrix} 5.807624 & 0.0000000 & 0.0000000 & 0.0000000 \\ 0.000000 & 0.6286934 & 0.0000000 & 0.0000000 \\ 0.000000 & 0.0000000 & 0.2793346 & 0.0000000 \\ 0.000000 & 0.0000000 & 0.0000000 & 0.1245547 \end{bmatrix} \tag{13}$$

(d)

$$A - C_1DC_1^T = \begin{bmatrix} 0.0185322 & -0.0236776 & 0.0022514 & 0.0117144 & -0.0050283 & 0.0010041 & -0.0050700 \\ -0.0236776 & 0.0320584 & -0.0074482 & -0.0055547 & 0.0051210 & -0.0046044 & 0.0037495 \\ 0.0022514 & -0.0074482 & 0.0131158 & -0.0249549 & -0.0003726 & 0.0125842 & 0.0068709 \\ 0.0117144 & -0.0055547 & -0.0249549 & 0.0615883 & -0.0038173 & -0.0248261 & -0.0185905 \\ -0.0050283 & 0.0051210 & -0.0003726 & -0.0038173 & 0.0096471 & -0.0076151 & 0.0019424 \\ 0.0010041 & -0.0046044 & 0.0125842 & -0.0248261 & -0.0076151 & 0.0190772 & 0.0065991 \\ -0.0050700 & 0.0037495 & 0.0068709 & -0.0185905 & 0.0019424 & 0.0065991 & 0.0057739 \end{bmatrix} \tag{14}$$

$||A - C_1DC_1^T|| = 0.09097174$

Code used to compute the answers for (a) to (d):

```r
# a
spectral_decomp <- eigen(A,symmetric=T,only.values=F)
C <- spectral_decomp$vectors
C
```

```
##            [,1]       [,2]       [,3]        [,4]        [,5]        [,6]
## [1,] -0.3777657 -0.4071756 -0.1405803  0.58706293 -0.16706891  0.53969730
## [2,] -0.3832103 -0.4136291 -0.1007833  0.19407501  0.09350016 -0.74493139
## [3,] -0.3680361 -0.4593531  0.2370255 -0.64543118  0.32727328  0.24009405
## [4,] -0.3947810  0.1612459  0.1475424 -0.29520804 -0.81905467 -0.01650651
## [5,] -0.3892610  0.3090877 -0.4219855 -0.06669044  0.02613100 -0.18898771
## [6,] -0.3760945  0.4231899 -0.4060627 -0.08015699  0.35169796  0.24049968
## [7,] -0.3552031  0.3892153  0.7410610  0.32107640  0.24700821 -0.04826992
##             [,7]
## [1,]  0.08893934
## [2,] -0.26565662
## [3,]  0.12660435
## [4,] -0.19521315
## [5,]  0.73076817
## [6,] -0.57150644
## [7,]  0.08208401
```

```r
eigenvalues <- spectral_decomp$values
D <- diag(eigenvalues)
D
```

```
##          [,1]      [,2]      [,3]      [,4]       [,5]       [,6]       [,7]
## [1,] 5.807624 0.0000000 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## [2,] 0.000000 0.6286934 0.0000000 0.0000000 0.00000000 0.00000000 0.00000000
## [3,] 0.000000 0.0000000 0.2793346 0.0000000 0.00000000 0.00000000 0.00000000
## [4,] 0.000000 0.0000000 0.0000000 0.1245547 0.00000000 0.00000000 0.00000000
## [5,] 0.000000 0.0000000 0.0000000 0.0000000 0.09097174 0.00000000 0.00000000
## [6,] 0.000000 0.0000000 0.0000000 0.0000000 0.00000000 0.05451882 0.00000000
## [7,] 0.000000 0.0000000 0.0000000 0.0000000 0.00000000 0.00000000 0.01430226
```

```r
# b
round(C%*%t(C),6)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    0    0    0    0    0    0
## [2,]    0    1    0    0    0    0    0
## [3,]    0    0    1    0    0    0    0
## [4,]    0    0    0    1    0    0    0
## [5,]    0    0    0    0    1    0    0
## [6,]    0    0    0    0    0    1    0
## [7,]    0    0    0    0    0    0    1
```

```
# c
C1 <- C[,1:4]
C1
```

```
##            [,1]       [,2]       [,3]        [,4]
## [1,] -0.3777657 -0.4071756 -0.1405803  0.58706293
## [2,] -0.3832103 -0.4136291 -0.1007833  0.19407501
## [3,] -0.3680361 -0.4593531  0.2370255 -0.64543118
## [4,] -0.3947810  0.1612459  0.1475424 -0.29520804
## [5,] -0.3892610  0.3090877 -0.4219855 -0.06669044
## [6,] -0.3760945  0.4231899 -0.4060627 -0.08015699
## [7,] -0.3552031  0.3892153  0.7410610  0.32107640
```

```
D1 <- D[1:4,1:4]
D1
```

```
##          [,1]      [,2]      [,3]      [,4]
## [1,] 5.807624 0.0000000 0.0000000 0.0000000
## [2,] 0.000000 0.6286934 0.0000000 0.0000000
## [3,] 0.000000 0.0000000 0.2793346 0.0000000
## [4,] 0.000000 0.0000000 0.0000000 0.1245547
```

```
# d
A-C1%*%D1%*%t(C1)
```

```
##              ...1         ...2          ...3         ...4          ...5
## ...1  0.018532210 -0.023677600  0.0022514118  0.011714443 -0.0050283018
## ...2 -0.023677600  0.032058396 -0.0074481747 -0.005554688  0.0051210365
## ...3  0.002251412 -0.007448175  0.0131157744 -0.024954945 -0.0003725707
## ...4  0.011714443 -0.005554688 -0.0249549452  0.061588330 -0.0038172668
## ...5 -0.005028302  0.005121036 -0.0003725707 -0.003817267  0.0096470573
## ...6  0.001004107 -0.004604430  0.0125841816 -0.024826096 -0.0076150906
## ...7 -0.005070031  0.003749518  0.0068708773 -0.018590526  0.0019424398
##              ...6         ...7
## ...1  0.001004107 -0.005070031
## ...2 -0.004604430  0.003749518
## ...3  0.012584182  0.006870877
## ...4 -0.024826096 -0.018590526
## ...5 -0.007615091  0.001942440
## ...6  0.019077201  0.006599078
## ...7  0.006599078  0.005773858
```

```r
norm(A-C1%*%D1%*%t(C1),type="2")
```

```
## [1] 0.09097174
```