STA437 Assignment #1

Rui Qiu #999292509

2016-01-31

Problem 1

(a) Solution:
We know that the density function of univariate Normal Distribution is:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

The marginal distributions of $X_1, \cdots, X_5$ are:

$$f_1(x_1) = \frac{1}{\sigma_1\sqrt{2\pi}}\, exp(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2})$$
$$f_2(x_2) = \frac{1}{\sigma_2\sqrt{2\pi}}\, exp(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2})$$
$$f_3(x_3) = \frac{1}{\sigma_3\sqrt{2\pi}}\, exp(-\frac{(x_3-\mu_3)^2}{2\sigma_3^2})$$
$$f_4(x_4) = \frac{1}{\sigma_4\sqrt{2\pi}}\, exp(-\frac{(x_4-\mu_4)^2}{2\sigma_4^2})$$
$$f_5(x_5) = \frac{1}{\sigma_5\sqrt{2\pi}}\, exp(-\frac{(x_5-\mu_5)^2}{2\sigma_5^2})$$

where
$\mu_1 = 1, \sigma_1^2 = 4.5,$
$\mu_2 = 2, \sigma_2^2 = 4.0,$
$\mu_3 = 1, \sigma_3^2 = 7.5,$
$\mu_4 = 0, \sigma_4^2 = 8.0,$
$\mu_5 = 0, \sigma_5^2 = 5.5.$

i.e. the marginal distributions of $X_1, \ldots, X_5$ are $N(1, 4.5), N(2, 4.0), N(1, 7.5), N(0, 8.0), N(0, 5.5)$.

(b) Solution:

$$\boldsymbol{x_1} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \boldsymbol{x_2} = \begin{pmatrix} x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ -1 \end{pmatrix},$$

$$\boldsymbol{\mu_1} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \boldsymbol{\mu_2} = \begin{pmatrix} \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

$$\Sigma_{11} = \begin{pmatrix} 4.5 & -2.0 \\ -2.0 & 4.0 \end{pmatrix}, \Sigma_{12} = \begin{pmatrix} -1.5 & -1.0 & -0.5 \\ 3.0 & 2.0 & 1.0 \end{pmatrix},$$

$$\Sigma_{21} = \begin{pmatrix} -1.5 & 3.0 \\ -1.0 & 2.0 \\ -0.5 & 1.0 \end{pmatrix}, \Sigma_{22} = \begin{pmatrix} 7.5 & 5.0 & 2.5 \\ 5.0 & 8.0 & 4.0 \\ 2.5 & 4.0 & 5.5 \end{pmatrix}.$$

And the conditional mean and covariance are:

$$\begin{aligned} \overline{\boldsymbol{\mu}} &= \boldsymbol{\mu_1} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x_2} - \boldsymbol{\mu_2}) \\ \overline{\Sigma} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ f(x_1, x_2 | x_3, x_4, x_5) &= \frac{f(x_1, x_2, x_3, x_4, x_5)}{f(x_3, x_4, x_5)} \\ &\sim N(\overline{\boldsymbol{\mu}}, \overline{\Sigma}) \end{aligned}$$

where $\overline{\boldsymbol{\mu}}, \overline{\Sigma}$ are calculated by:

```
> x2 <- matrix(c(2,3,-1),ncol=1)
> mu1 <- matrix(c(1,2),ncol=1)
> mu2 <- matrix(c(1,0,0),ncol=1)
> C <- matrix(c(4.5,-2,-1.5,-1,-0.5,-2,4,3,2,1,-1.5,3,7.5,5,2.5,-1,2,5,8,4,-0.5,1,2.5,4,5.5
),ncol=5)
> sigma11 <- C[1:2,1:2]
> sigma21 <- C[3:5,1:2]
> sigma12 <- C[1:2,3:5]
> sigma22 <- C[3:5,3:5]
> mubar <- mu1 + sigma12 %*% solve(sigma22) %*% (x2 - mu2); mubar
     [,1]
[1,]  0.8
[2,]  2.4
> sigmabar <- sigma11 -sigma12 %*% solve(sigma22) %*% sigma21; sigmabar
     [,1] [,2]
[1,]  4.2 -1.4
```

```
[2,] -1.4  2.8
```

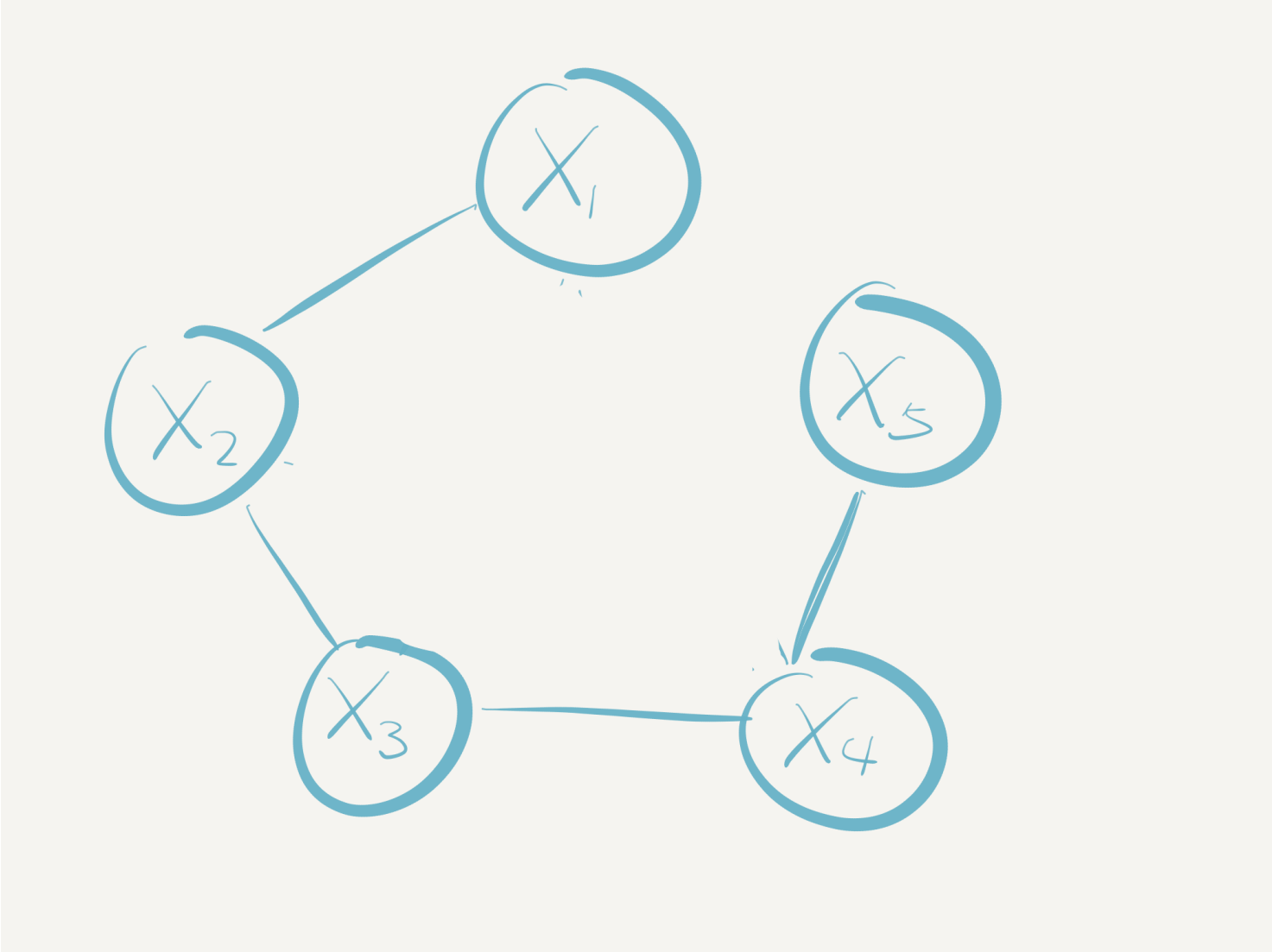
So the conditional distribution of (X_1, X_2) given $X_3 = 2, X_4 = 3, X_5 = -1$ is:

$$N(\begin{pmatrix} 0.8 \\ 2.4 \end{pmatrix}, \begin{pmatrix} 4.2 & -1.4 \\ -1.4 & 2.8 \end{pmatrix}).$$

(c) Solution:
Since all the denpendency information hides in the concentration matrix $K = C^{-1}$. So we observe C^{-1}:

```
> options(digits=5)
> solve(C)
          [,1]        [,2]        [,3]        [,4]     [,5]
[1,]  2.8571e-01  1.4286e-01 -1.2336e-17  1.7623e-18  0.00000
[2,]  1.4286e-01  4.2857e-01 -1.4286e-01  1.7843e-17  0.00000
[3,] -3.9651e-18 -1.4286e-01  2.8571e-01 -1.4286e-01  0.00000
[4,]  5.9476e-18  1.1895e-17 -1.4286e-01  2.8571e-01 -0.14286
[5,]  0.0000e+00  0.0000e+00  0.0000e+00 -1.4286e-01  0.28571
```

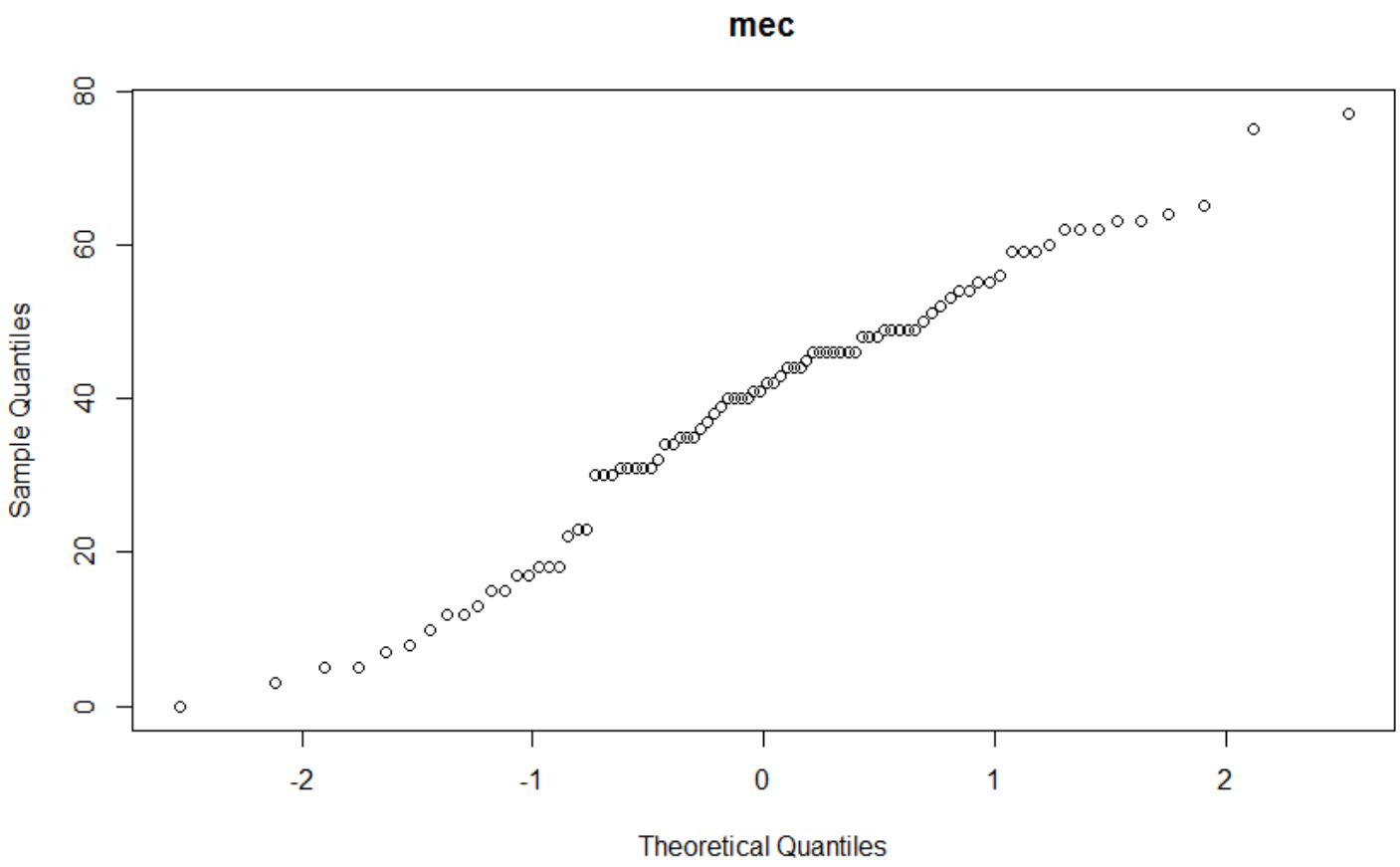
We find that $k_{15} = k_{25} = k_{35} = 0$ (and by symmetry, $k_{51} = k_{52} = k_{53} = 0$), so there is no edge between X_5 and X_1, X_2, X_3 respectively. Additionally, $k_{31} = k_{13} = k_{41} = k_{14} = k_{24} = k_{42} \approx 0$. Therefore, the graph structure is shown below:

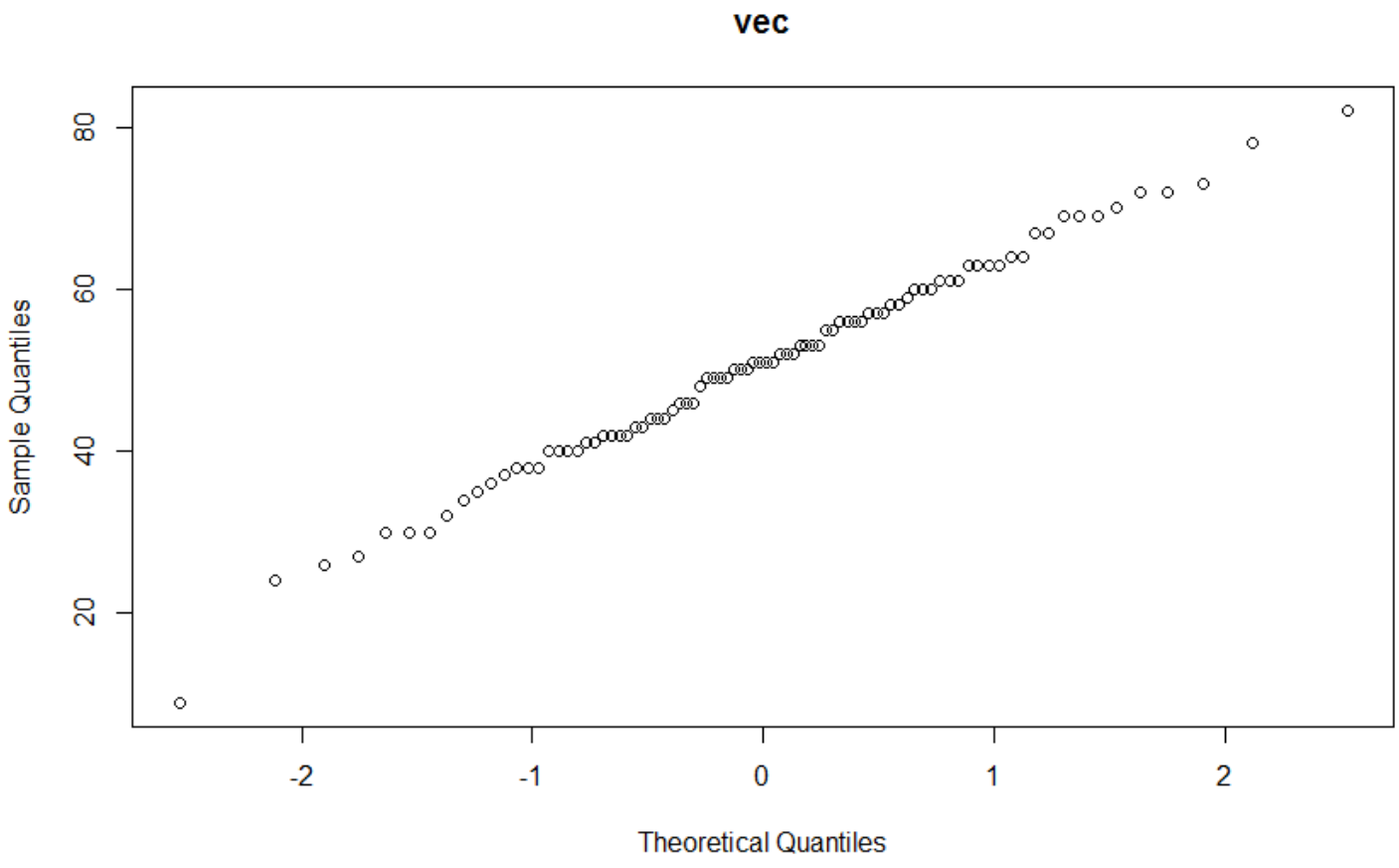


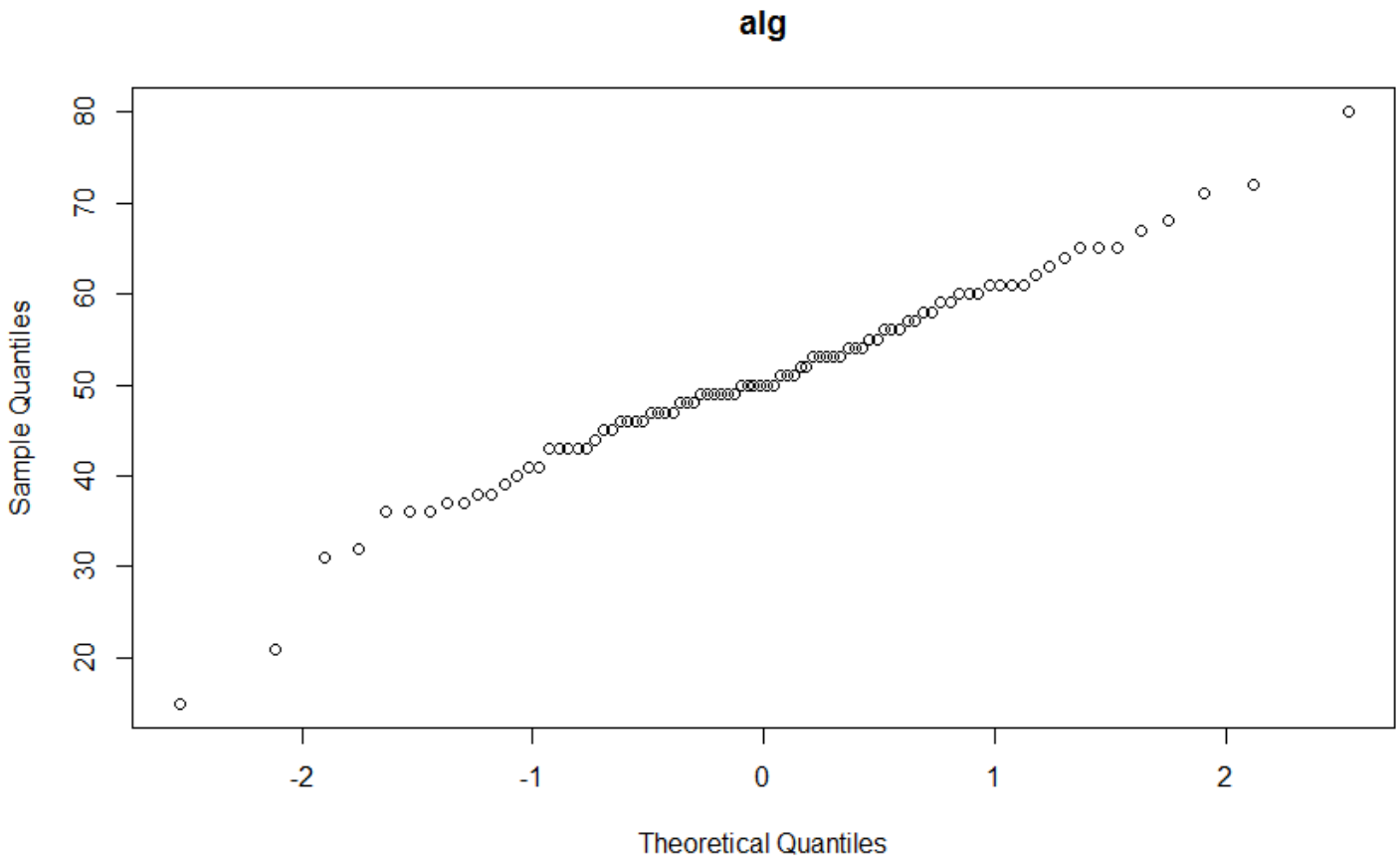
Problem 2

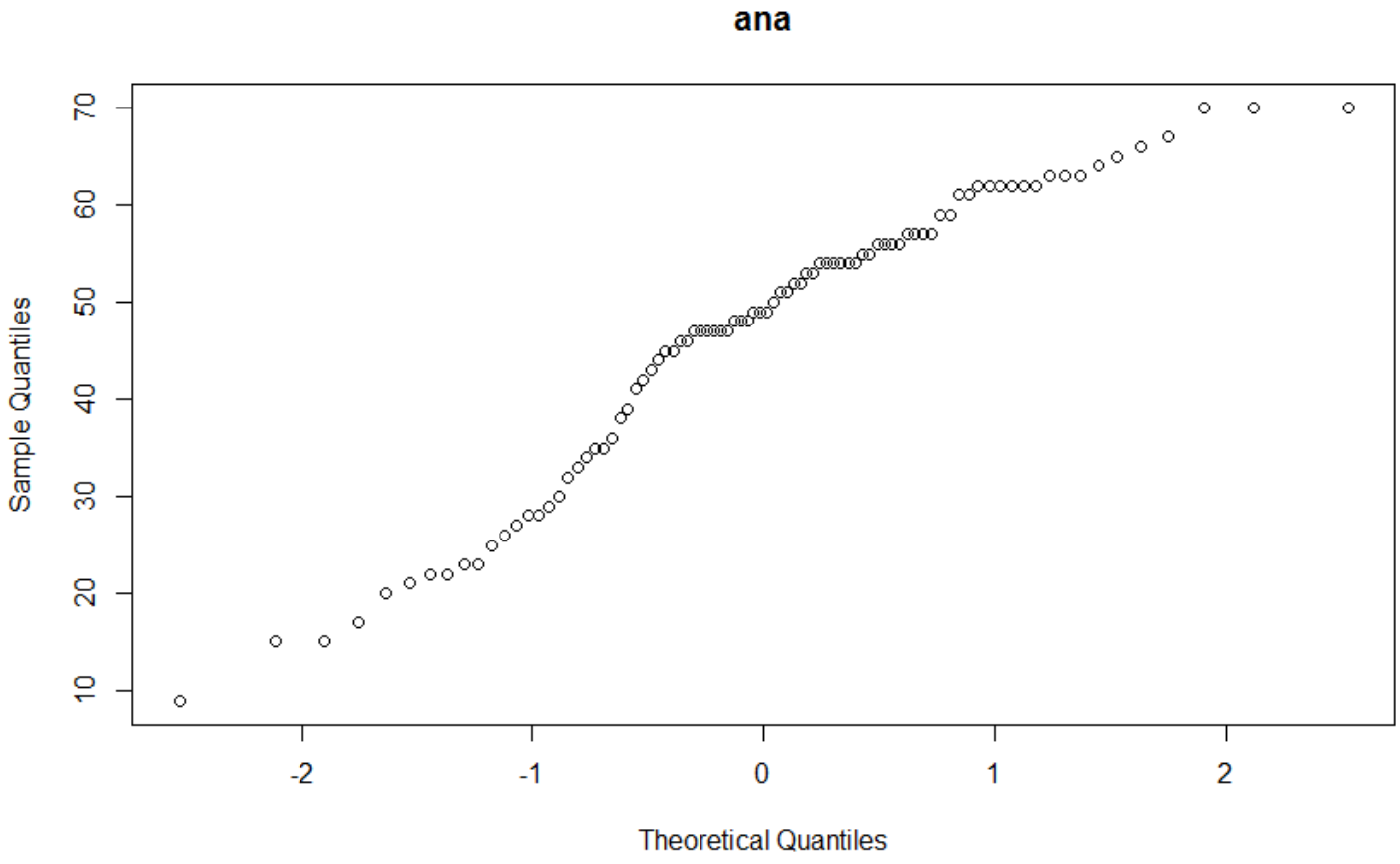
(a) Solution:

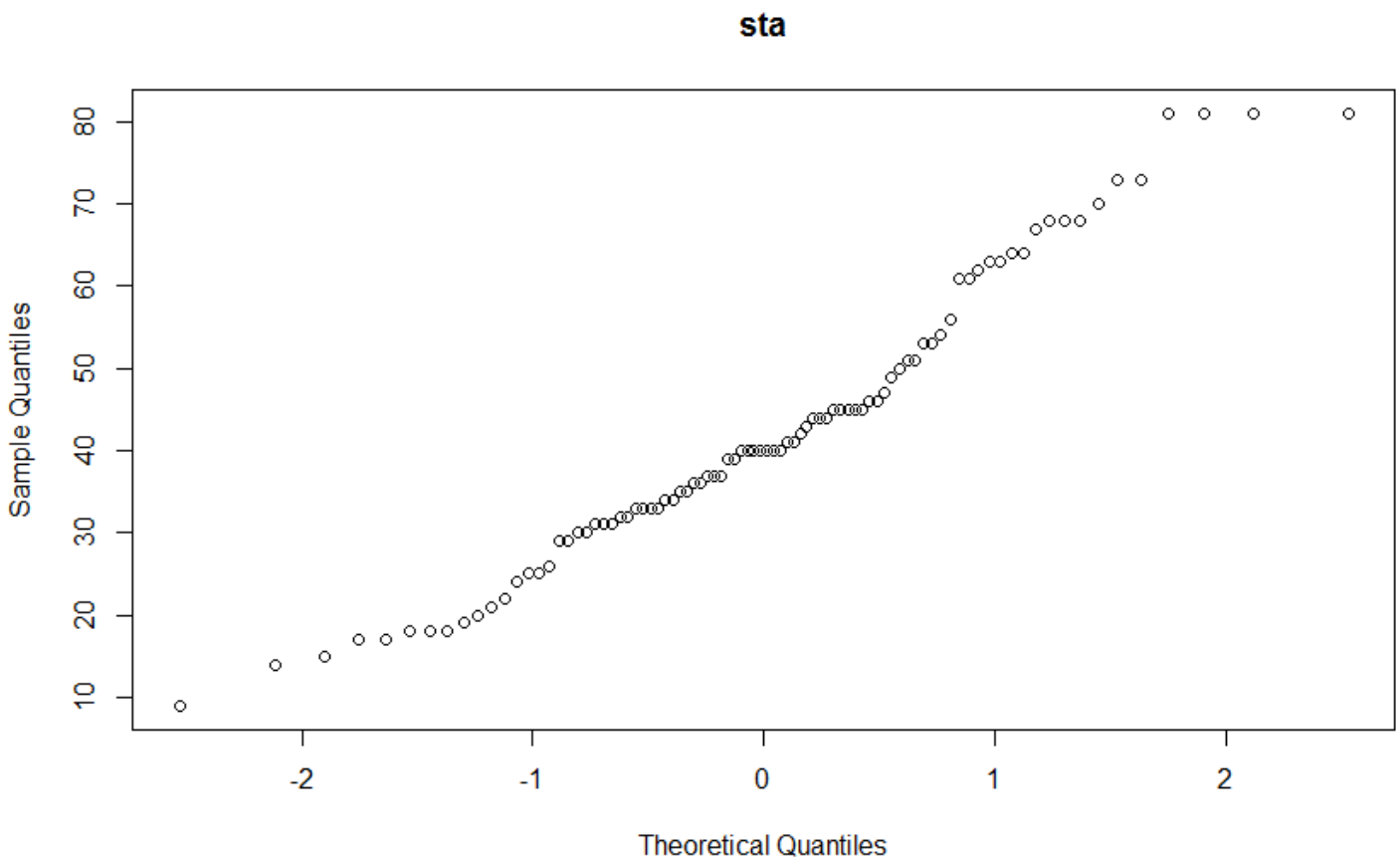
```
> exam <- scan("marks.txt",what=list(0,0,0,0,0))
Read 88 records
> mec <- exam[[1]]
> vec <- exam[[2]]
> alg <- exam[[3]]
> ana <- exam[[4]]
> sta <- exam[[5]]
> qqnorm(mec,main="mec")
> qqnorm(vec,main="vec")
> qqnorm(alg,main="alg")
> qqnorm(ana,main="ana")
> qqnorm(sta,main="sta")
```


mec
Sample Quantiles
Theoretical Quantiles
0
20
40
60
80
-2
-1
0
1
2


vec
Sample Quantiles
Theoretical Quantiles
20
40
60
80
-2
-1
0
1
2


alg
Sample Quantiles
Theoretical Quantiles
20
30
40
50
60
70
80
-2
-1
0
1
2


ana
Sample Quantiles
Theoretical Quantiles
10
20
30
40
50
60
70
-2
-1
0
1
2




```
> # Conduct Shapiro-Wilk test on each numeric vector.
> shapiro.test(mec)

    Shapiro-Wilk normality test

data:  mec
W = 0.97241, p-value = 0.05708

> shapiro.test(vec)

    Shapiro-Wilk normality test

data:  vec
W = 0.99305, p-value = 0.9276

> shapiro.test(alg)

    Shapiro-Wilk normality test

data:  alg
W = 0.98206, p-value = 0.2637

> shapiro.test(ana)

    Shapiro-Wilk normality test

data:  ana
```

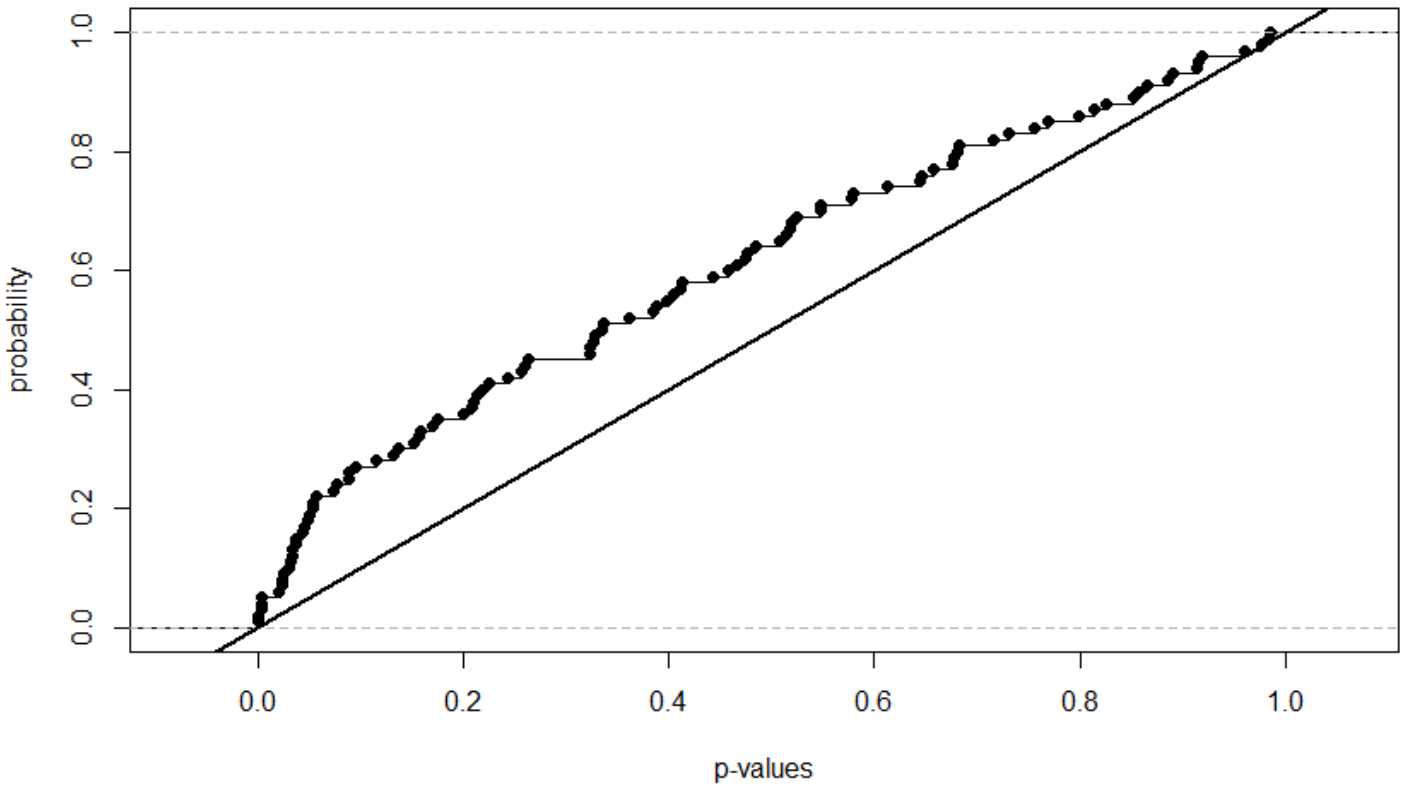
```
W = 0.94245, p-value = 0.0006896

> shapiro.test(sta)

    Shapiro-Wilk normality test

data:  sta
W = 0.96448, p-value = 0.01633
```

If we choose the significant level $\alpha = 0.05$ and compare it with all the p-value calculated above, considering the null hypothesis to be the population (marks of each subject) is normally distributed. So by this standard, The subject `Analysis` and `Statistics` are not normally distributed, but `Mechanics`, `Vector` and `Algebra` are normally distributed. However, if we only observe the normal qq-plot, it is rather difficult to find non-normality.

(b) Solution:

```
> qqmultinorm <- function(x,nproj=50,scale=T,plot.qq=F,plot.edf=F) {
+     p <- ncol(x)
+     if(scale) x <- scale(x,center=T,scale=T)
+     if(plot.qq) devAskNewPage(ask = T)
+     pvals <- NULL
+     for (i in 1:nproj) {
+         a <- rnorm(p)
+         a <- a/sqrt(sum(a^2))
+         y <- as.vector(x%*%a)
+         if (plot.qq) qqnorm(y)
+         r <- shapiro.test(y)
+         pvals <- c(pvals,r$p.value)
+     }
+     if (plot.edf) {
+         plot(ecdf(pvals),xlab="p-values",ylab="probability",
+              main=" ")
+         abline(0,1,lwd=2)
+     }
+     pvals
+ }
> qqmultinorm(cbind(mec,vec,alg,ana,sta),nproj=100,plot.edf = T)
  [1] 0.1702885479 0.0039282470 0.3370129277 0.7564195574 0.0036681689 0.9147497216 0.54827
83163 0.7697682982 0.2002148575
 [10] 0.7982440173 0.5479766005 0.0042656488 0.3872610138 0.4433911852 0.2095891181 0.17466
87856 0.2560393357 0.3229468406
 [19] 0.6826149058 0.0535594856 0.1143479294 0.8849812542 0.6766734457 0.5250354827 0.71557
09210 0.4106310046 0.3982922999
 [28] 0.6443565785 0.6452138342 0.0949228237 0.0309165043 0.6772934316 0.2595925665 0.98600
38310 0.9607452824 0.4849393279
 [37] 0.0334650807 0.7311569296 0.3843914649 0.0451430512 0.0729071900 0.0195461688 0.21741
84042 0.3617523234 0.9194818087
 [46] 0.5192814766 0.0762661473 0.6124619653 0.2627656677 0.8655291585 0.1362055917 0.03679
72484 0.2252608624 0.9146698749
 [55] 0.1587596702 0.5074427074 0.0884876065 0.4659083656 0.6570778383 0.4749472426 0.05373
01775 0.3275417363 0.2123803414
 [64] 0.0883848755 0.0259198266 0.3270835859 0.4125493339 0.4575065140 0.0336635252 0.00046
```

```
72151 0.0437661258 0.9776657175
 [73] 0.8145561709 0.8251221635 0.1321112163 0.8897860969 0.0480875811 0.0226081164 0.02347
39051 0.0368628309 0.6813812589
 [82] 0.1559373348 0.0315421817 0.5795950304 0.0566768329 0.8525610780 0.9832225452 0.40508
02453 0.2424252911 0.0508759390
 [91] 0.8574450372 0.0002604204 0.4761556069 0.3234476322 0.1522855208 0.5150665047 0.51739
29418 0.3351673406 0.5774913612
[100] 0.2079817122
```



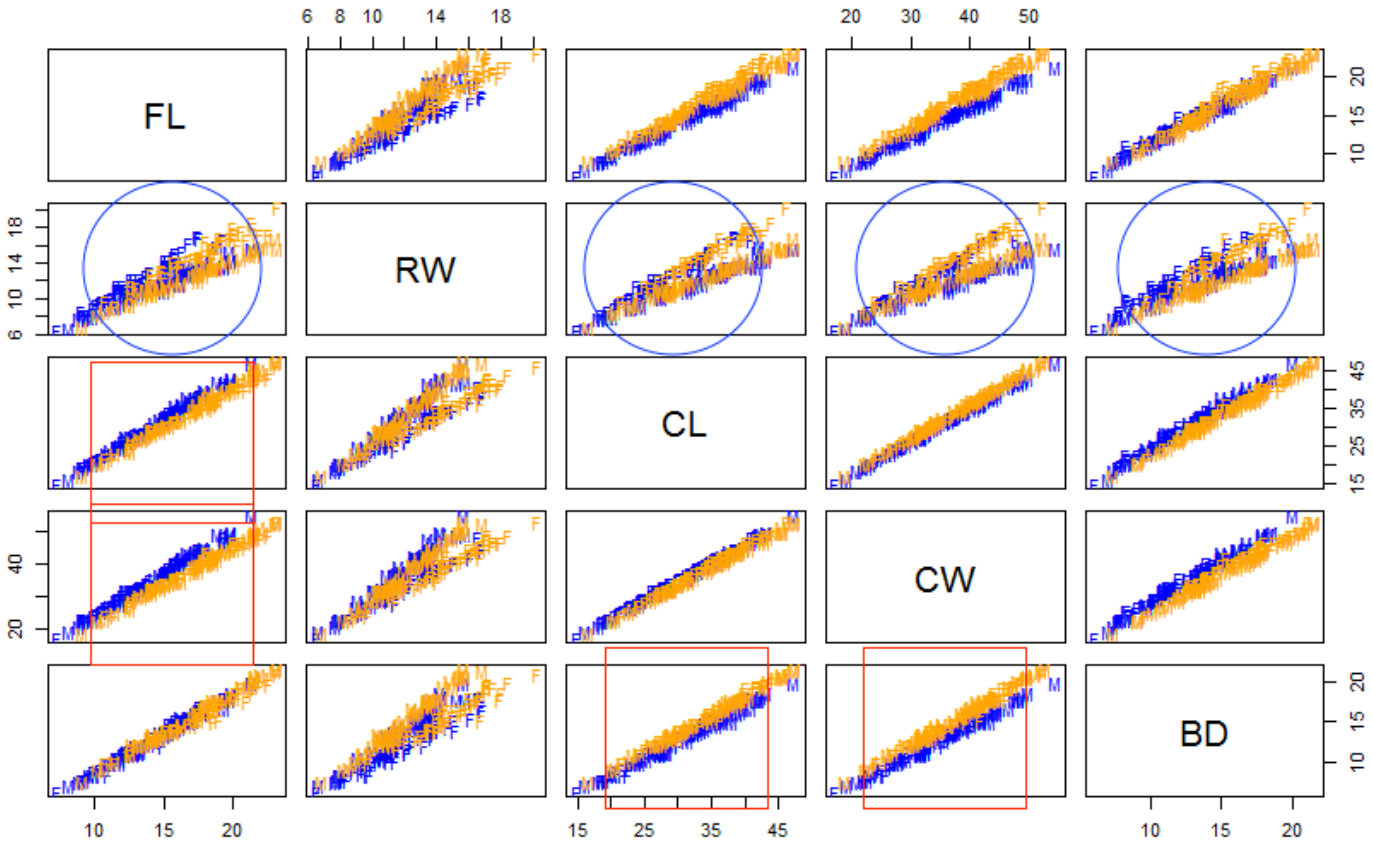
Based on the plot, the data seems to be not multivariate normal, since it's clearly above the diagonal line. And we all know points on diagonal line indicate normality.

Problem 3

Since the resolution of printed version is limited, the real plot is more distinguishable.

```
> x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
Read 200 records
> colour <- ifelse(x[[1]]=="B","blue","orange")
> sex <- x[[2]]
> FL <- x[[4]]
> RW <- x[[5]]
> CL <- x[[6]]
> CW <- x[[7]]
> BD <- x[[8]]
```

```
> pairs(cbind(FL,RW,CL,CW,BD),pch=sex,col=colour)
```



(a) Solution:
We are looking for those pairwise scatterplots whose overlap between two species is as little as possible. By observing, the scatterplots of

- `FL vs CL`,
- `FL vs CW`,
- `BD vs CL` and
- `BD vs CW`
  are particularly effective for "separating" the two species (identified by red squares).

(b) Solution:
Similarly, the scatterplots of

- `RW vs FL`
- `RW vs CL`
- `RW vs CW`
- `RW vs BD`
  are particularly effectice for "separating" the two sexes (identified by blue circles). (If we look carefully, there is an obvious bifurcation in these scatterplots, and each brach mainly consists one sex.)

(c) Solution:
First we would ask, how many such 3D scatterplots do we need to draw? Since the number of variables is 5, each 3D scatterplots has 3 dimension, so the answer is `choose(5,3)` which is 10.

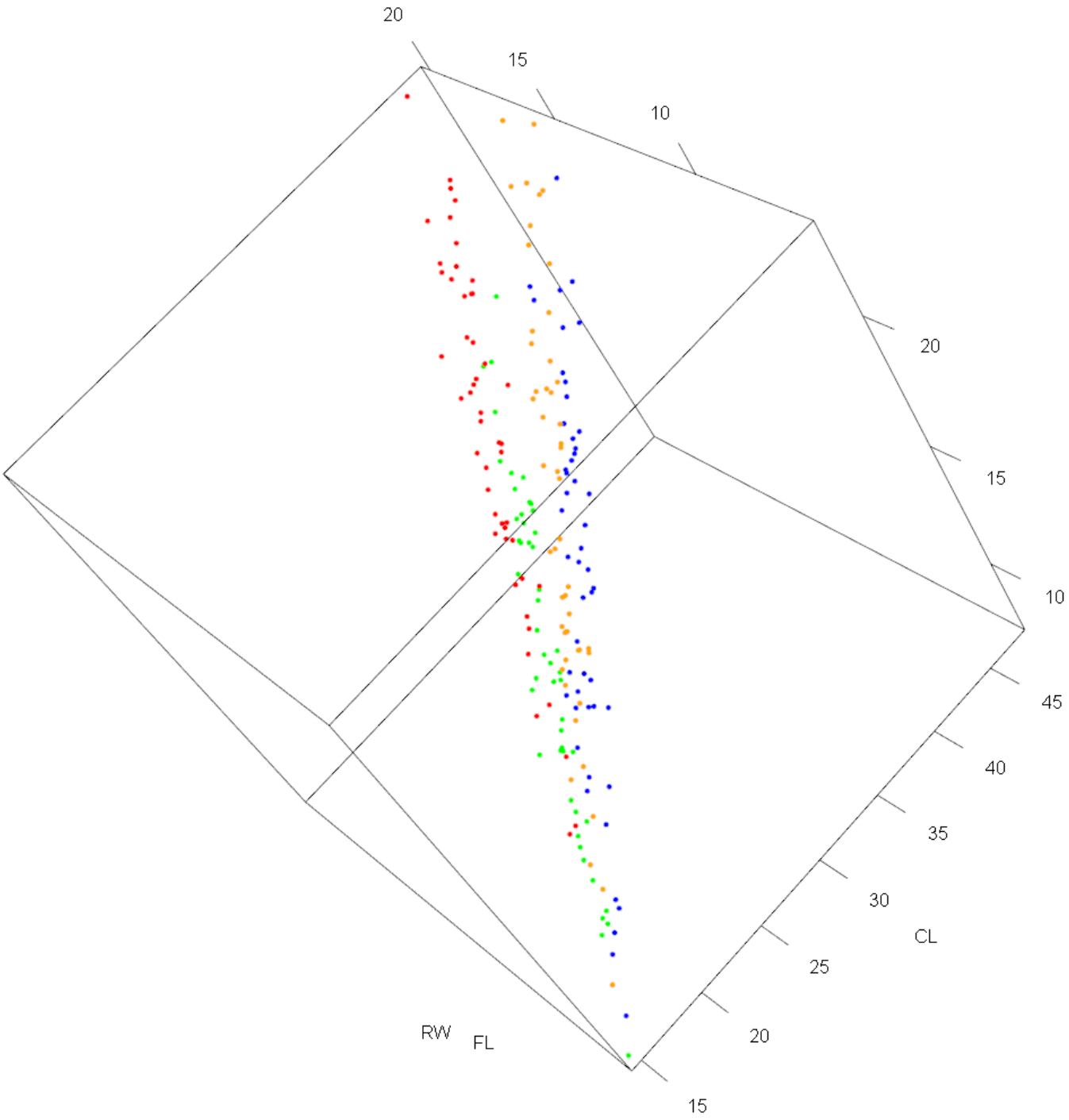
But in fact, consider that pairs in part (a) and part (b) already gave us good pairs to separating species and sexes. So we suppose we can only consider those 3-variable combinations of 1 solution in part (a) and 1 solution in part (b):

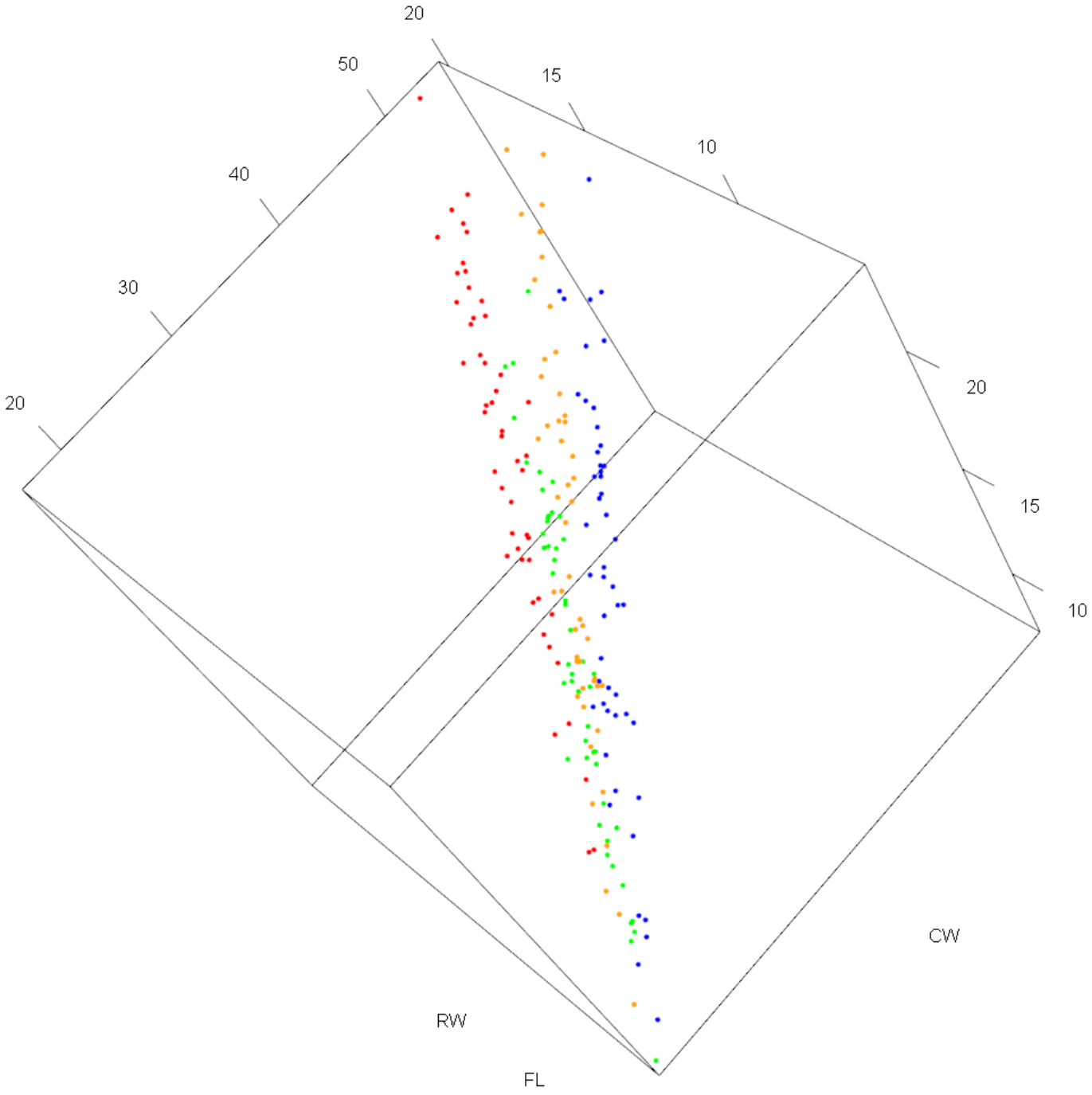
- `FL, CL, RW`
- `FL, CW, RW`
- `BD, CL, RW`
- `BD, CW, RW`

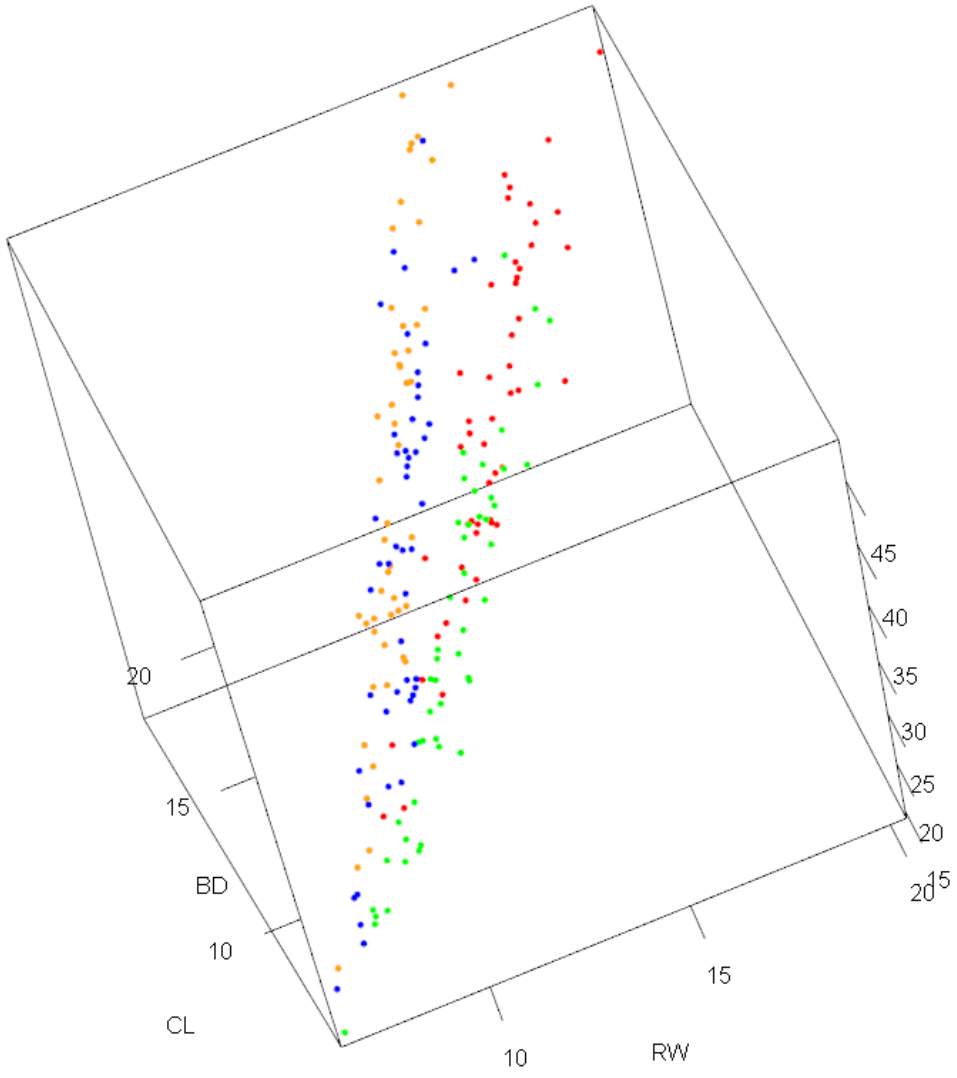
I tried to use package `scatterplot3d` to draw the plots but it turned out that the graph is not easy to read in practice. So I took a detour, using package `rgl` and function `plot3d()` instead. But before the real plotting, I had to change colours of female crabs, because `plot3d()` does not support `pch`.

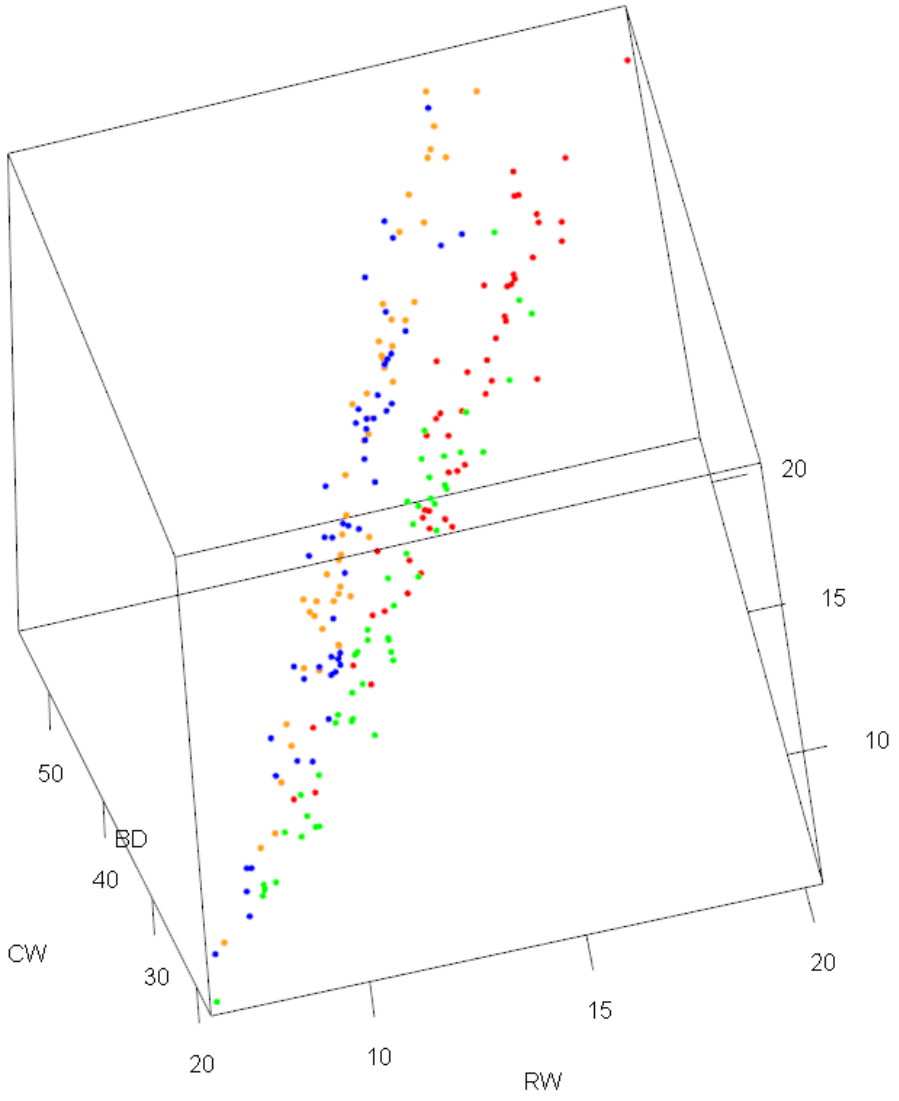
```
> library(rgl)
> for (i in 1:200) {if (sex[i]=="F" && colour[i]=="blue") {colour[i] <- "green"}}
> for (i in 1:200) {if (sex[i]=="F" && colour[i]=="orange") {colour[i] <- "red"}}
> plot3d(FL, CL, RW, col=colour, size=5)
> plot3d(FL, CW, RW, col=colour, size=5)
> plot3d(BD, CL, RW, col=colour, size=5)
> plot3d(BD, CW, RW, col=colour, size=5)
```

So we have the following 4 plots which can be zoom-in/out/rotated very easily (but cannot be shown on paper):


20
15
10
20
15
10
45
40
35
30
CL
25
20
15
RW
FL


20
50
15
40
10
30
20
20
15
10
CW
RW
FL


45
40
35
30
25
20
15
20
15
10
CL
10
15
20
RW
BD



By rotating and playing around a little bit with them, the 2nd and 4th plots have 4 obvious branches while the 1st and 3rd plots have 2 obvious branches of species but much overlap in sexes. So the following 2 combinations of variables effectively separate 4 species/sex groups, i.e.:

- `FL, CW, RW` and
- `BD, CW, RW`