STA437 Assignment #4

Rui Qiu #999292509

2016-03-31

Problem 1

(a) Proof:
For single linkage clustering, we have:

$$d(U, V) = \min\{d(x_i, x_j) : x_i \in U, x_j \in V\}$$

So we want to find $\min\{d(x_i, x_j) : x_i \in A, x_j \in B \cup C\}$.
There are two possibilities where the x_j could be located:

- Suppose we have the minimum distance $d(x_i, x_j)$, and $x_i \in A$, $x_j \in B \subset B \cup C$, i.e. $d(A, B) < d(A, C)$.

$$\begin{aligned}\therefore d(A, B \cup C) &= \min\{d(A, B), d(A, C)\} = \frac{1}{2}\, d(A, B) + \frac{1}{2}\, d(A, C) + \frac{1}{2}\, d(A, B) - \frac{1}{2}\, d(A, C) = d(A, B)\\ &= \frac{1}{2}\, d(A, B) + \frac{1}{2}\, d(A, C) - \frac{1}{2}\, |d(A, B) - d(A, C)|\end{aligned}$$

- Similarly, if we suppose the minimum distance $d(x_i, x_j)$ has $x_i \in A, x_j \in C \in B \cup C$, i.e. $d(A, B) > d(A, C)$.

$$\begin{aligned}\therefore d(A, B \cup C) &= \min\{d(A, B), d(A, C)\} = \frac{1}{2}\, d(A, B) + \frac{1}{2}\, d(A, C) - \frac{1}{2}\, d(A, B) + \frac{1}{2}\, d(A, C) = d(A, C)\\ &= \frac{1}{2}\, d(A, B) + \frac{1}{2}\, d(A, C) - \frac{1}{2}\, |d(A, B) - d(A, C)|\end{aligned}$$

(b) Proof:
For complete linkage clustering, we have:

$$d(U, V) = \max\{d(x_i, x_j) : x_i \in U, x_j \in V\}$$

So we want to find $\max\{d(x_i, x_j) : x_i \in A, x_j \in B \cup C\}$.
There are two possibilities where the x_j could be located:

- Suppose we have the maximum distance $d(x_i, x_j)$, and $x_i \in A$, $x_j \in B \subset B \cup C$, i.e. $d(A, B) > d(A, C)$.

$$\begin{aligned}\therefore d(A, B \cup C) &= \max\{d(A, B), d(A, C)\} = \frac{1}{2}\, d(A, B) + \frac{1}{2}\, d(A, C) + \frac{1}{2}\, d(A, B) - \frac{1}{2}\, d(A, C) = d(A, B)\\ &= \frac{1}{2}\, d(A, B) + \frac{1}{2}\, d(A, C) + \frac{1}{2}\, |d(A, B) - d(A, C)|\end{aligned}$$

- Similarly, if we suppose the maximum distance $d(x_i, x_j)$ has $x_i \in A, x_j \in C \in B \cup C$, i.e.

$d(A,B) < d(A,C)$.

$$\therefore d(A,B \cup C) = \max\{d(A,B), d(A,C)\} = \frac{1}{2}d(A,B) + \frac{1}{2}d(A,C) - \frac{1}{2}d(A,B) + \frac{1}{2}d(A,C) = d(A,C$$
$$= \frac{1}{2}d(A,B) + \frac{1}{2}d(A,C) + \frac{1}{2}|d(A,B) - d(A,C)|$$

(c) Proof:

$$d(A,C) = \max\{d(a,c) : a \in A, c \in C\}$$
$$d(A,B) = \max\{d(a,b) : a \in A, b \in B\}$$
$$d(B,C) = \max\{d(b,c) : b \in B, c \in C\}$$

Show

$$d(A,C) \leq d(A,B) + d(B,C)$$

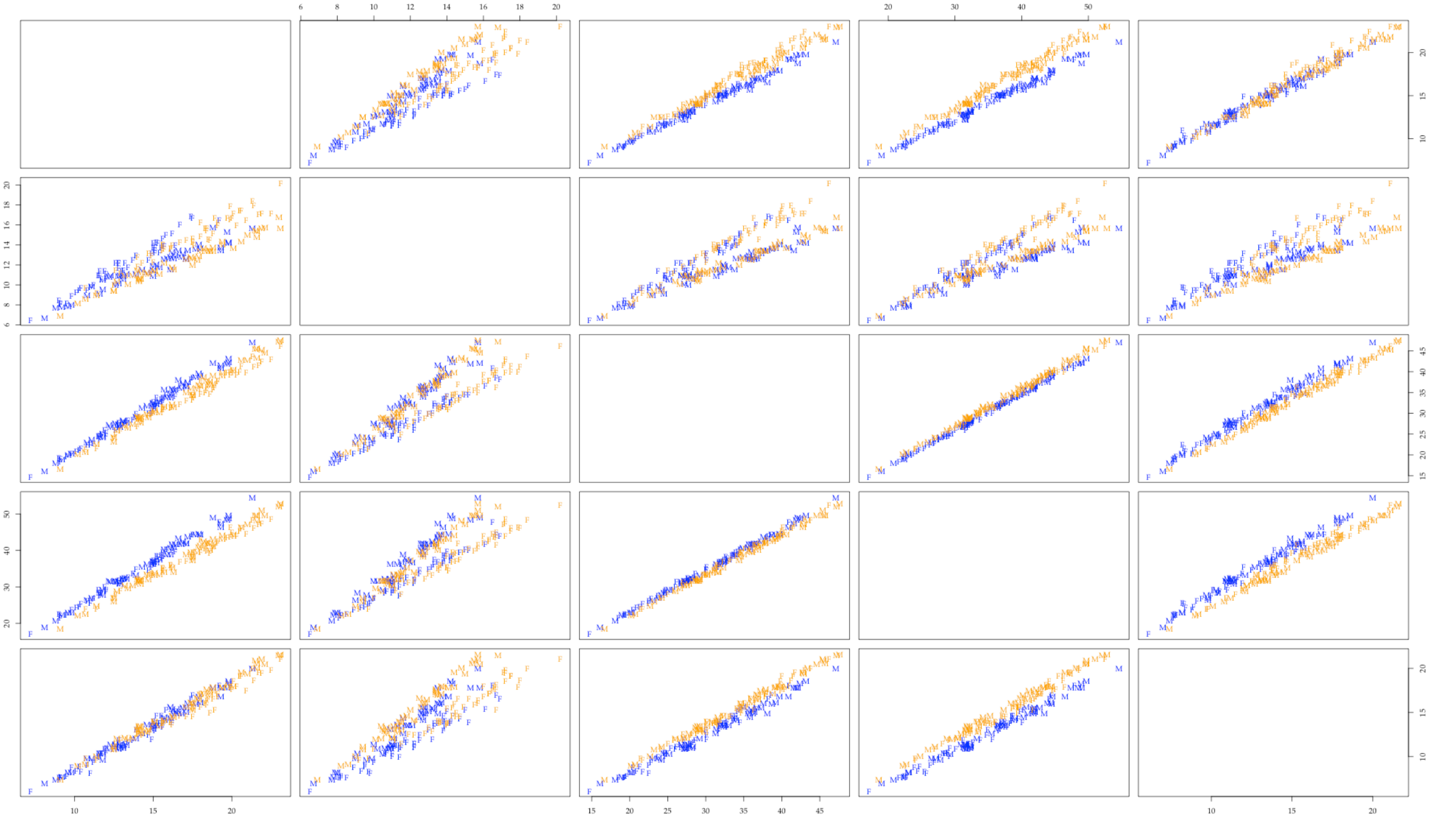
Suppose $d(A,C) = d(a_1,c_1), d(A,B) = d(a_2,b_2), d(B,C) = d(b_3,c_3)$.
Since $d(A,C)$ is the largest direct distance between any a, c, which is from a_1 to c_1.
But according to triangle inequality $d(a,c) \leq d(a,b) + d(b,c)$, we could have a indirect distance larger than $d(a_1,c_1)$:

$$\begin{aligned} d(a_1,c_1) &\leq d(a_1,b_0) + d(b_0,c_1) \\ &\leq d(a_2,b_2) + d(b_3,c_3) \end{aligned}$$

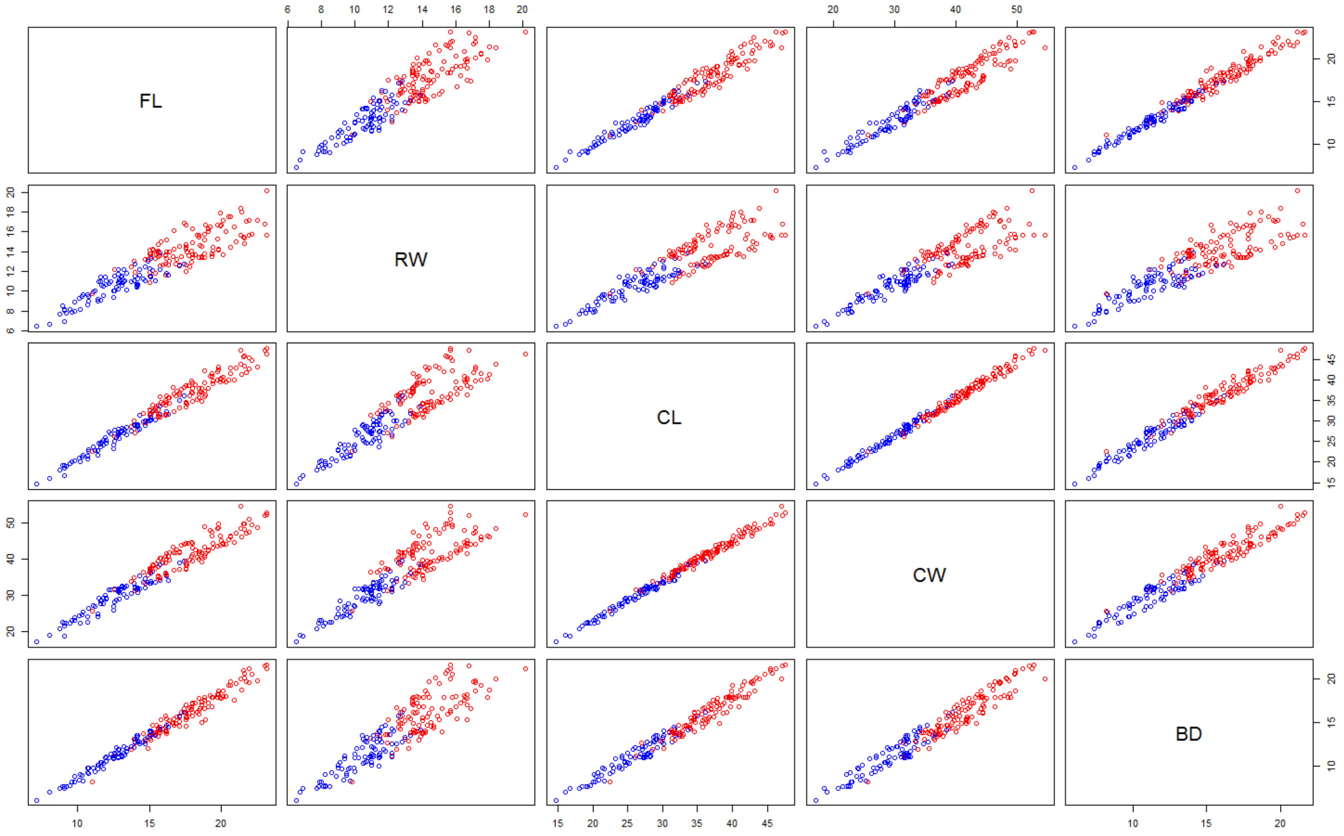
Therefore, $d(A,C) \leq d(A,B) + d(B,C)$.

Problem 2

(a) Solution:

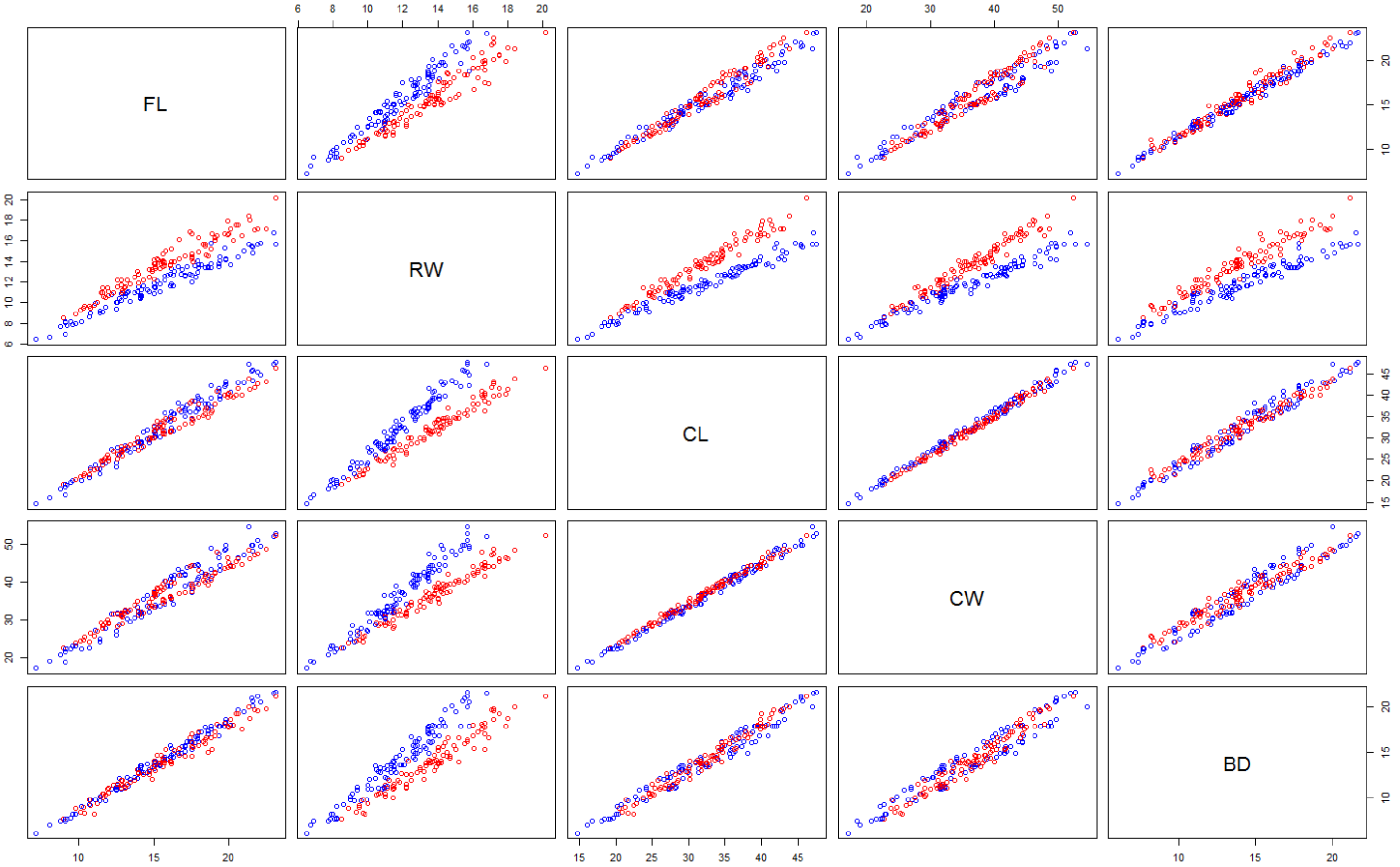


- First we recall the pairwise plot in Homework 1.



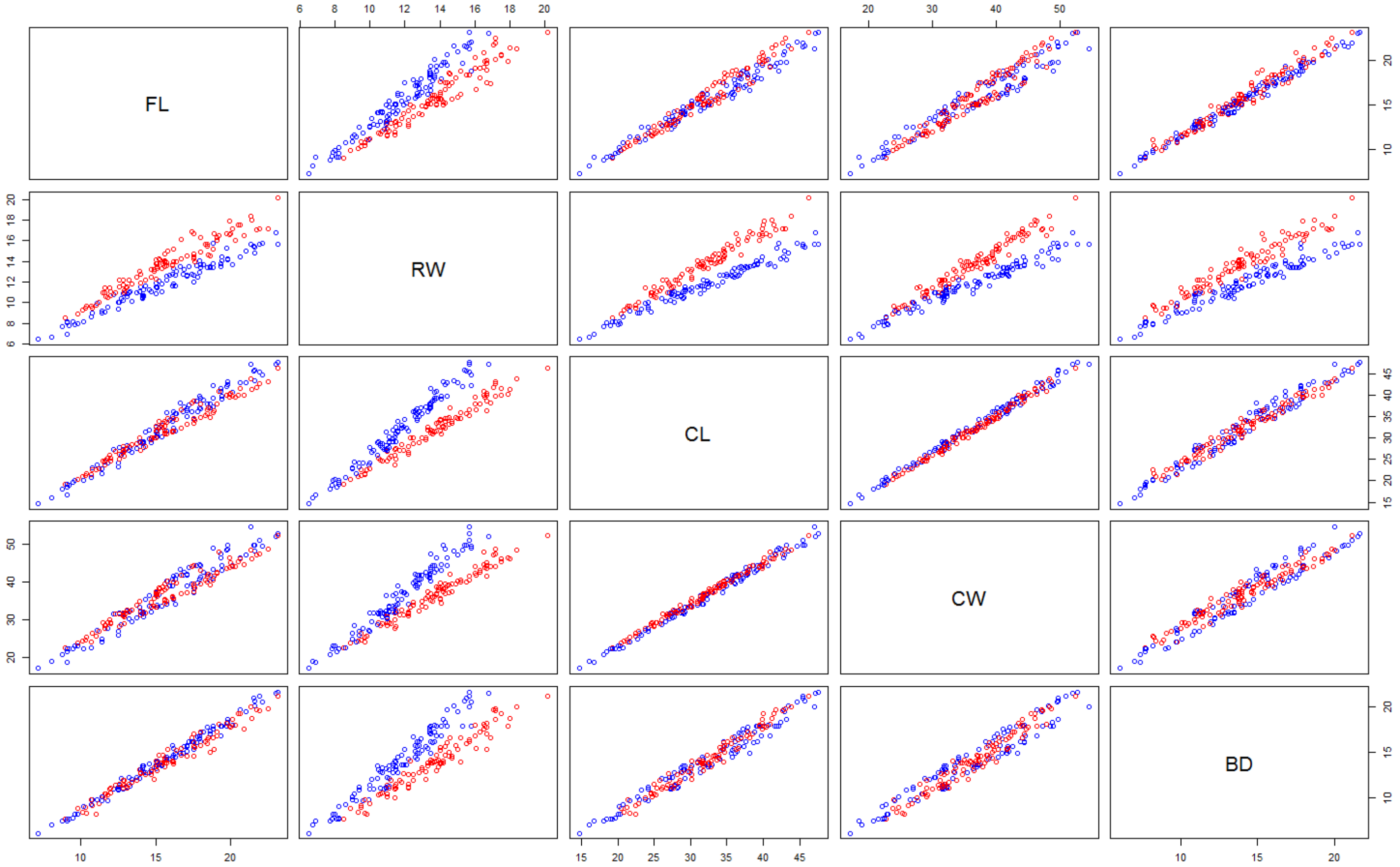
- The clusters estimated after 30 iterations seem not very plausible. Although it did distinguish two different clusters, however, it cannot be interpreted as either species or sexes.
- This is probably, constrained by the small number of iterations.

(b) Solution:



- After 100 iterations, the algorithm finds out more confound difference, which seems like the species. And clearly we can see two branches in the plot indicating two crab species in `FL vs. RW`, `CL vs. RW` etc.
- But for `CW vs. BD`, if the EM perfectly separates two species, this plot should have two clearly different coloured braches (like the one we had in HW1), but in fact we still see mixed red and blue dots in this one.
- The more iterations the algorithm runs, the more accurately we can know about the data.

(c) Solution:



- We make a final run with 1000 iterations.
- So result is very consistent with the previous 100 runs.
- The mechanism that obviously more runs still cannot separate two clusters perfectly is maybe because the data have multiple "peaks" so it will never find the real maximal for some data points.

Appendix

```
> x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
Read 200 records
> FL <- x[[4]]
> RW <- x[[5]]
> CL <- x[[6]]
> CW <- x[[7]]
> BD <- x[[8]]
> y <- cbind(FL,RW,CL,CW,BD)
> source("em.r")
> r30 <- EM(y,k=2,em.iter=30)
> r100 <- EM(y,k=2,em.iter=100)
> colour <- rep("blue",200)
> colour[r30$cluster==2] <- "red"
> pairs(y,col=colour)
> colour <- rep("blue",200)
> colour[r100$cluster==2] <- "red"
> pairs(y,col=colour)
> r1000 <- EM(y,k=2,em.iter=1000)
> colour <- rep("blue",200)
> colour[r1000$cluster==2] <- "red"
> pairs(y,col=colour)
```