503 HW 5

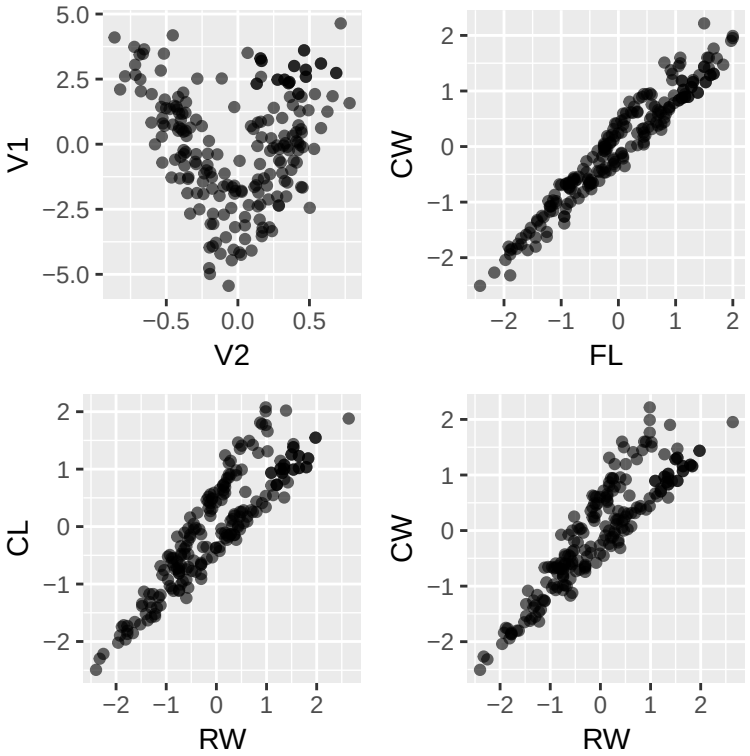
Naomi Giertych

4/16/2018

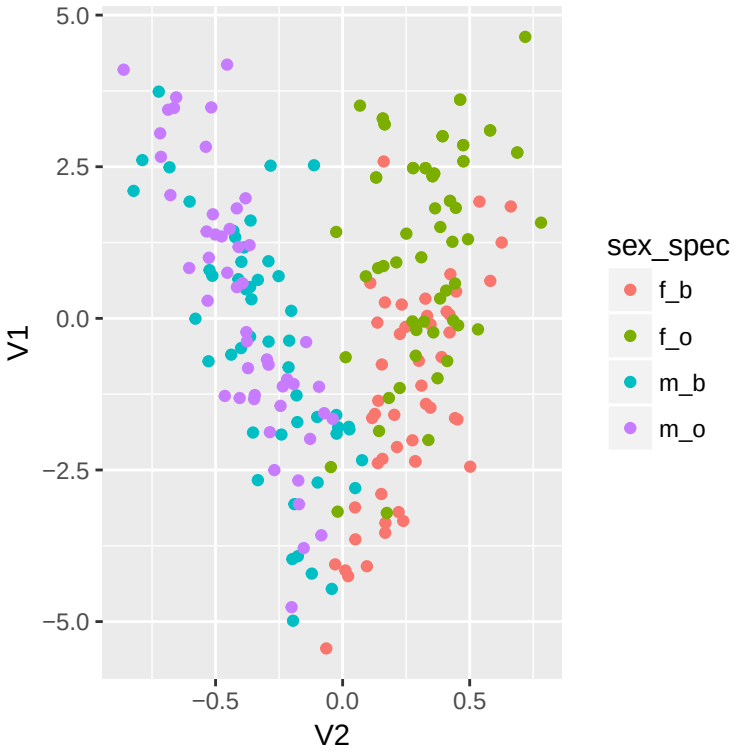
Question 1

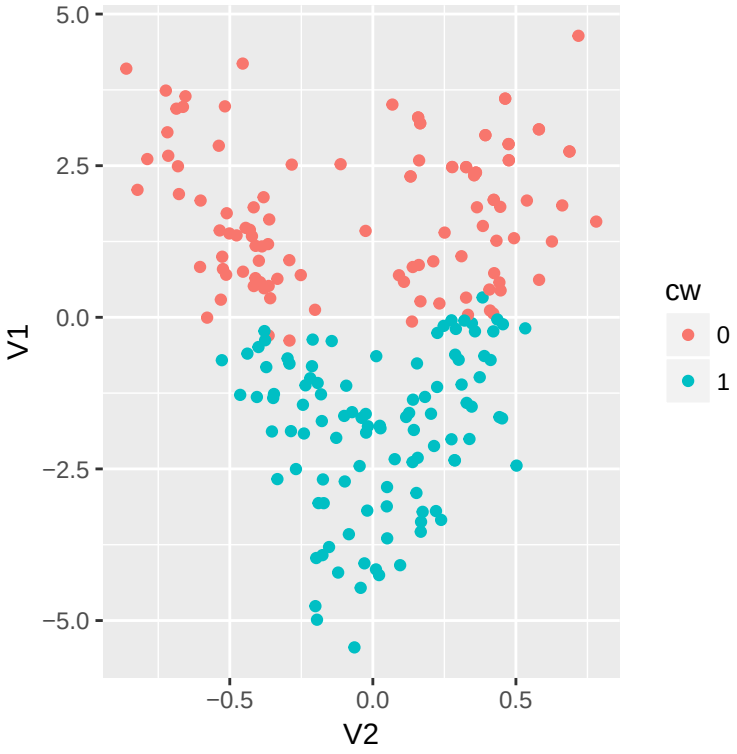
I perform a clustering analysis on a dataset containing 5 morphological measurements of blue and orange crabs of both sexes of the species *Leptograpsus variegatus* collected in Australia. I explore hierarchical clustering, K-means, and mixture modeling using different values of k. I then choose the best value of k, k^* for the mixture model using the BIC criterion. Finally, I compare the results between the K-means and the mixture model using this k^*.

Before performing any analysis, I centered and scaled the dataset. I then calculated a distance matrix of the data and performed multidimensional scaling (MDS) to reduce the data to two dimensions for visualization purposes. Below I plot the variables obtained from the MDS in the top left graph and a few of the variables from the scaled dataset in the other graphs. "FL" stands for frontal lobe size; "CW" is the carapace width; "CL" is the carapace length; and "RW" is the rear width.



I added back in the sex and species data for the purposes of visualizing the results of the MDS results. As we can see the second variable of MDS is easily interpreted as the sex of the crab. Similarly, the second variable is easily interpreted as whether or not the carapace width is below average (1) or above average (0).



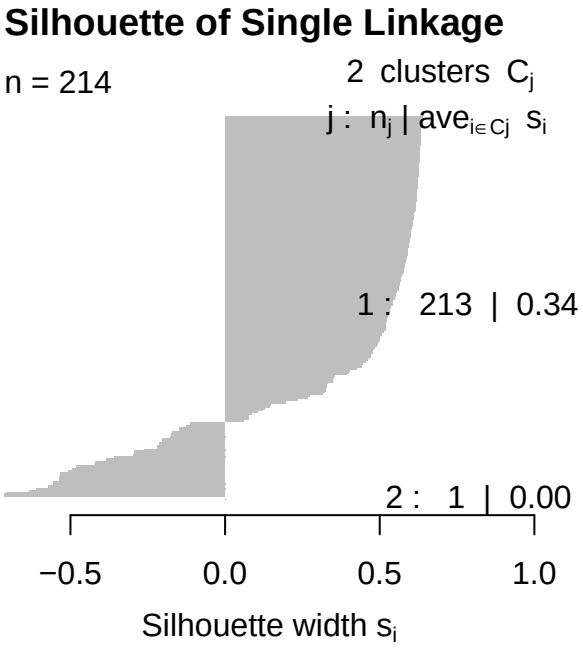


Additionally, I added back sex and species to the scaled dataset to get a better understanding of how

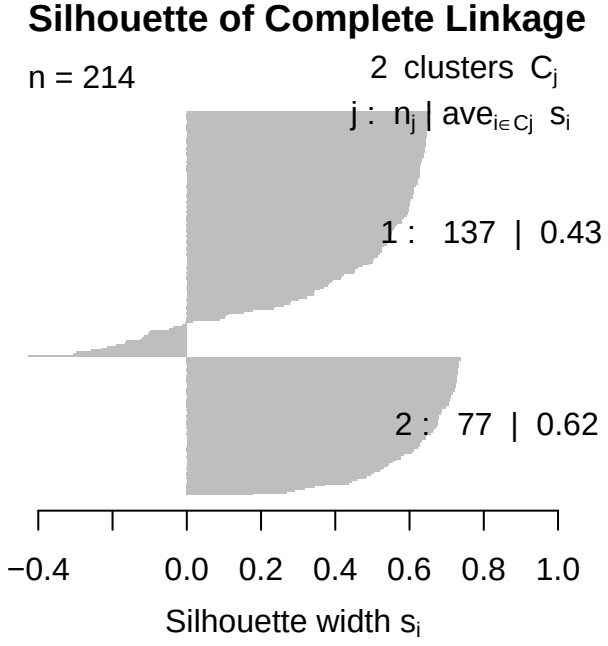
As we can see from the graphs above, the crabs dataset appears to have two clusters with some separation along carapace length and width. A combination of the carapace length and width may be the defining factor for the first variable from the MDS above.

Hierarchical clustering

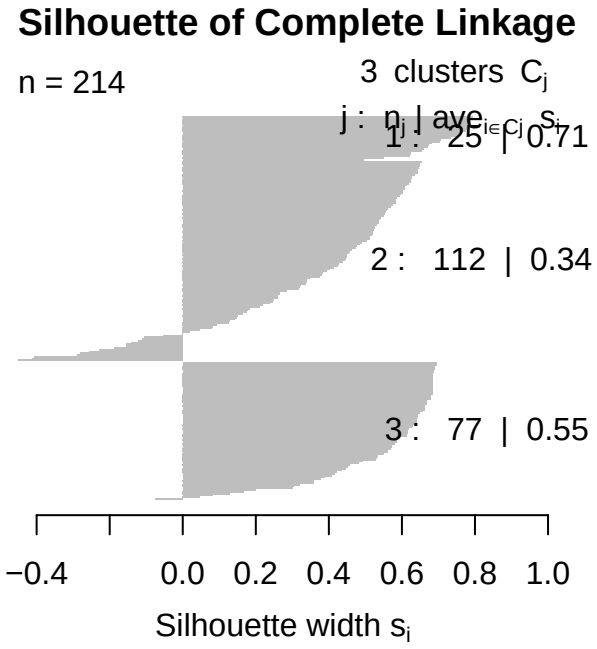
I perform hierarchical clustering using single, complete, and average linkage. Instead of using the dissimilarity matrix, I use the scaled data and specify the method. Below I plot the silhouette of each of these methods.

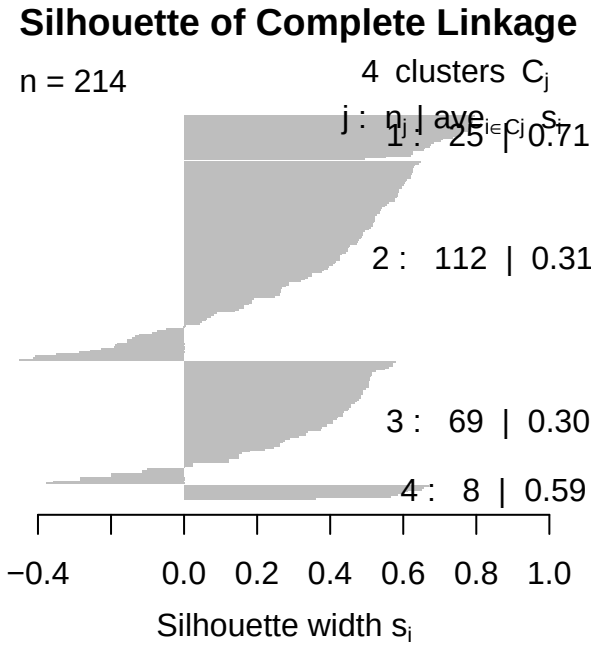


From the silhouette plot above for single linkage, we can see that group 1 only has a weak structure and group 2 has no structure. It does not make sense to continue grouping the objects for single linkage.



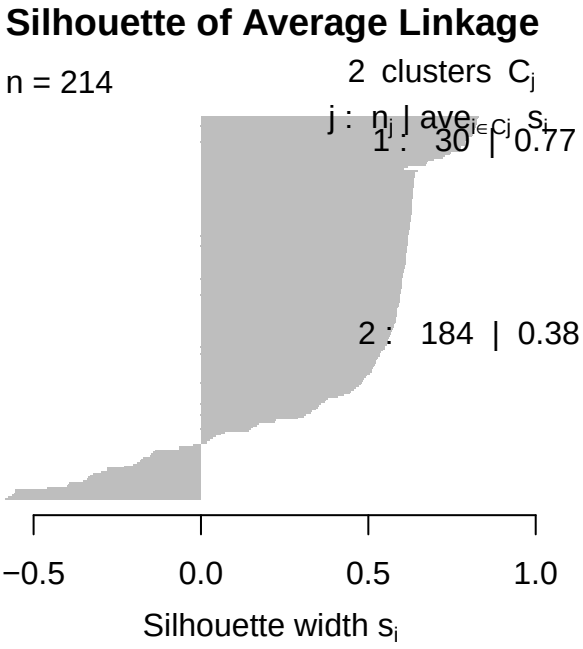
From the silhouette plot above for complete linkage, we can see that group 1 still has a fairly weak structure. However, group 2 now has a somewhat reasonable amount of structure. Additionally, there are no outliers in the groups. Below I explore a couple of other groupings from complete linkage.



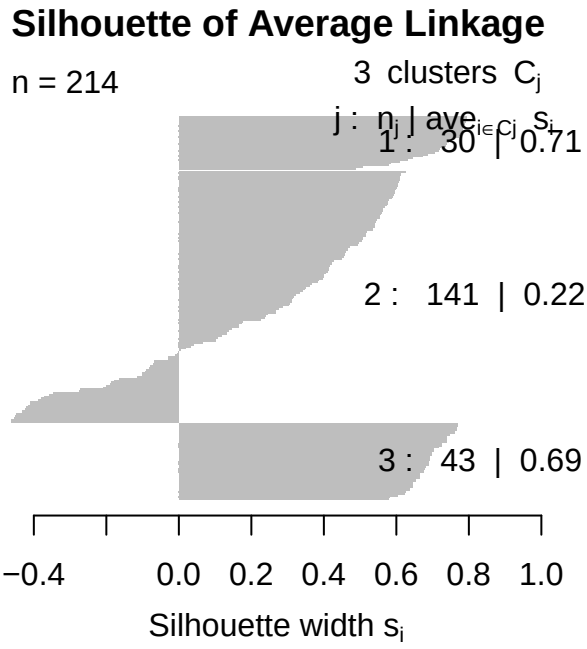


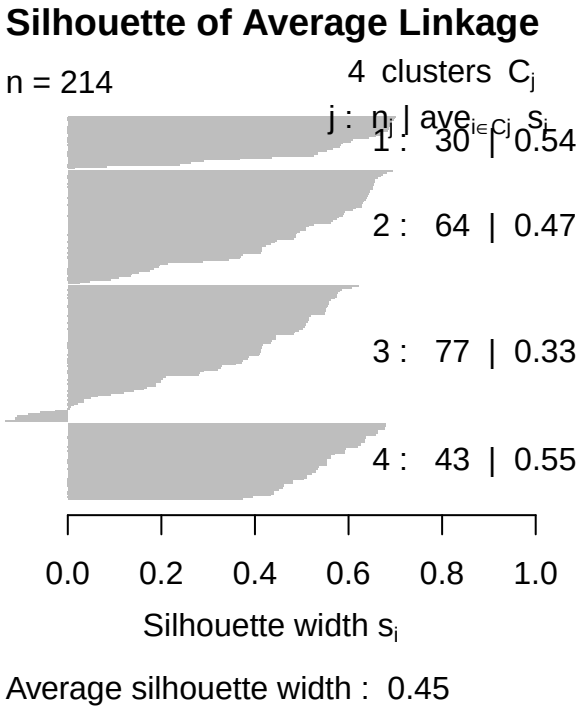
As we can see from the additional complete linkage silhouette plots, the first group now has a fairly strong structure and the last group (either third or fourth) has a moderate amount of structure. The remaining groups only have a small amount of structure and contain several outliers that do not fit within the group as indicated by the bars in the negative direction.

Finally, I examine the silhouette plots for the average linkage again starting with only two clusters.



As we can see from the average linkage silhouette plot above, the first group has a fairly strong structure. However, the second group has a very week structure with many outliers. Below I explore a couple of other groupings of the average linkage.





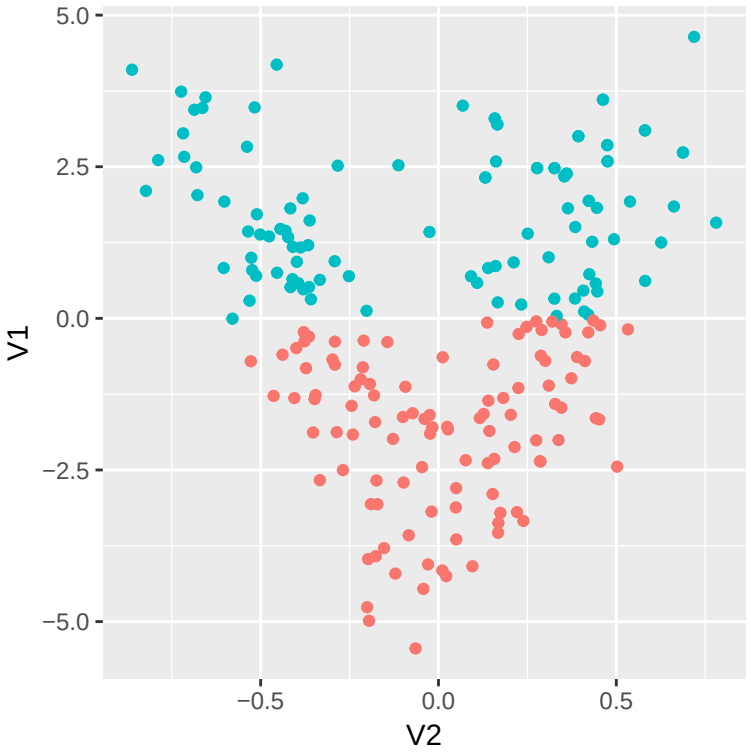
Similar to the complete linkage silhouette plots, the first group now has a fairly strong structure and the last group (either third or fourth) has a moderate amount of structure whereas the remaining groups have a small or moderate amount of structure.

K-means

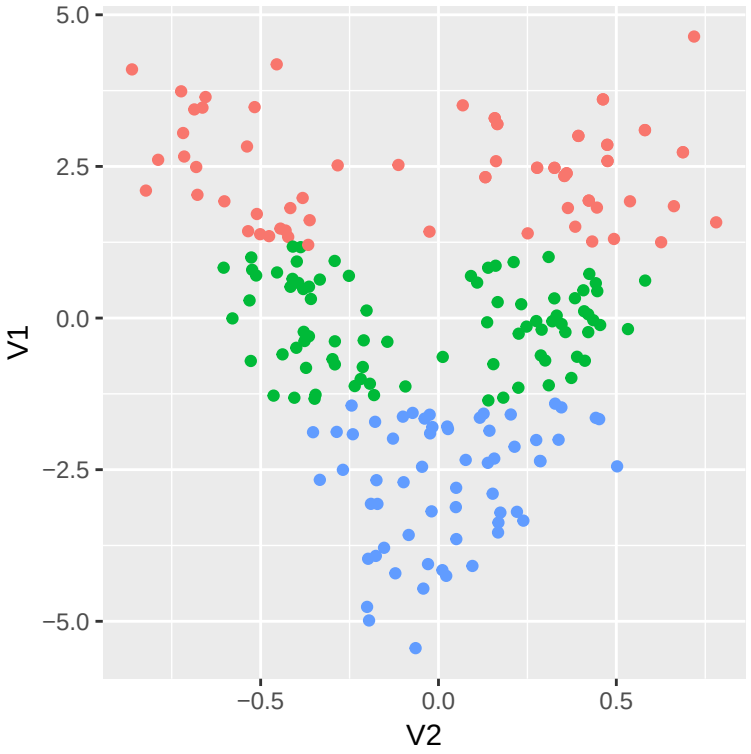
The K-means algorithm is performed as follows:

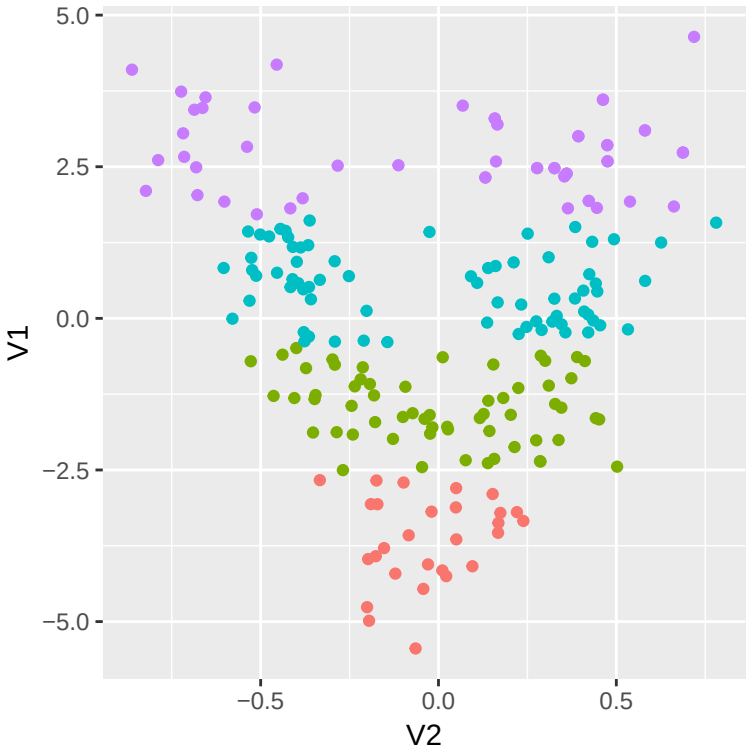
1) Specify the number of clusters, K.
2) Randomly select K of the objects from the data set as the initial cluster centroids.
3) Assign each observation to the closest centroid based on the Euclidean distance between the observation and the current centroid assignment.
4) For each of the clusters update the cluster centroid by calculating the new mean values of the observations within the cluster. Note the the center of the cluster is the $p \times 1$ vector that contains the mean of each variable for the observations. In our case, the centroids would be a 5×1 vector of the means for FL, RW, CL, CW, and BD of the crabs in the cluster.
5) Iterate steps 3 and 4 until cluster assignments converge and minimize the total within sum of squares.

K-means is very sensitive to initial values, so as a rule we always use multiple starts and choose the best one.



As we can see from the graph above, our K-means cluster analysis splits the data into two groups along the first variable of the MDS analysis above which we determined before to be whether the carapace width is above or below average. Below, I try 3 and 4 clusters.



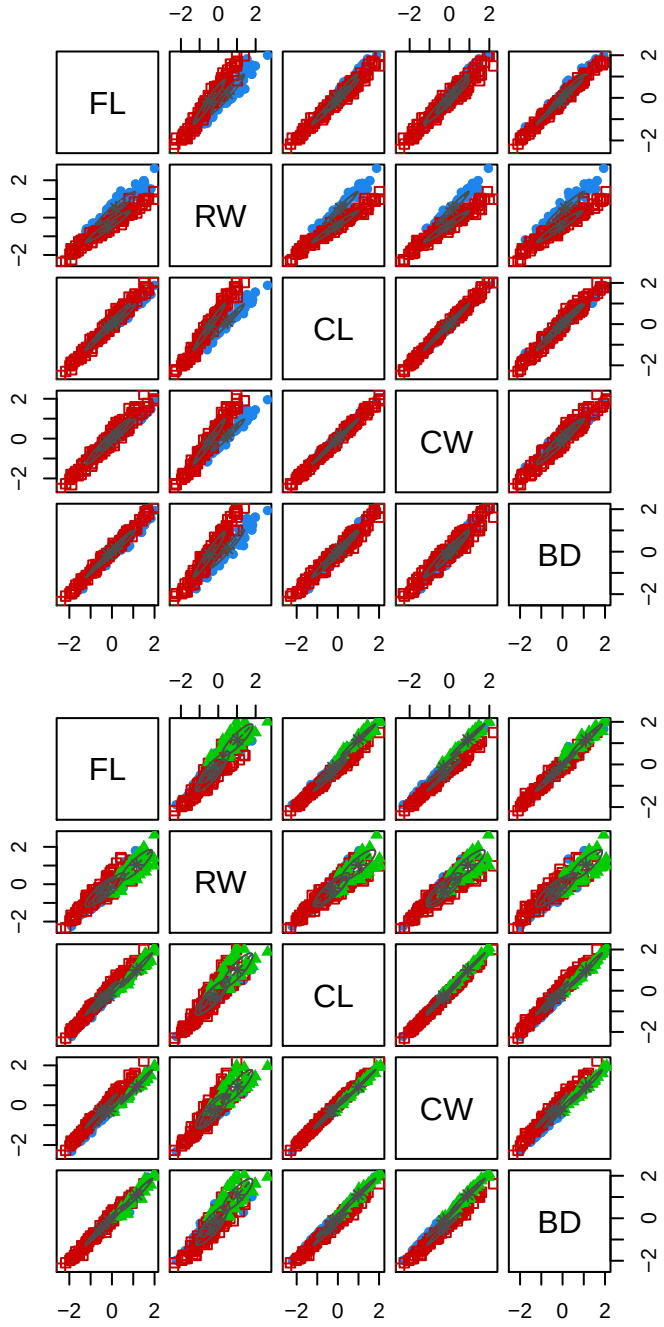


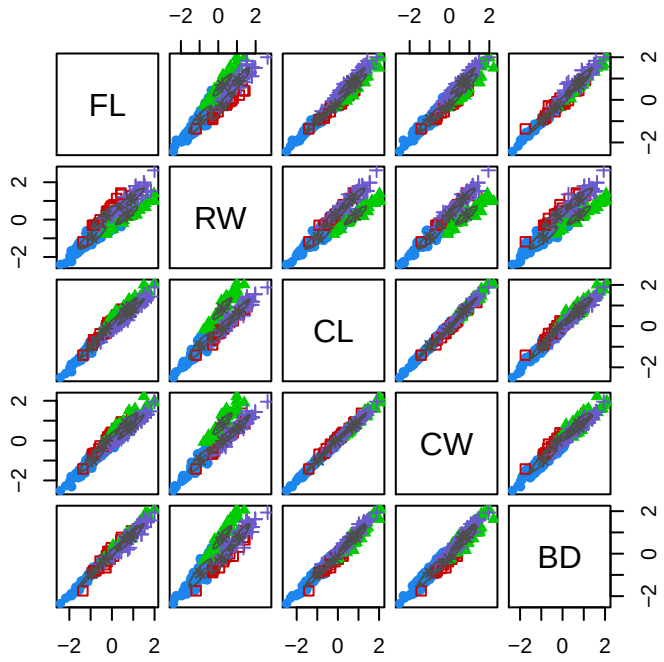
Of all of the potential K-means clusters, four clusters appears to the best since it captures the group at the bottom of the "V" that has a tight distribution around V2 and V1 but maintains the group with the widest distribution on V2. It is also appealing since it might have a natural relationship with the species and sex of each crab.

Mixture Model

Next, I fit a Gaussian mixture model to attempt to capture the clusters within the crab data. Bayes Gaussian mixture modeling assumes that there is a latent variable, Z_n that determines the class of each observation and each observation is normally distributed with some class-specific mean and variance given this latent variable, i.e. $X_n|Z_n \sim N(\mu_k, \sum_k)$. We then assume that Z_n has a distribution given by π, $Z_n|\pi \sim Discrete(\pi)$ and π has a prior $Dir(\alpha)$ distribution. Finally, we assume that $\mu_k|(\mu_0, \sum_0) \sim N(\mu_0, \sum_0)$ and $\sum_k|\Psi \sim IW(\Psi, m)$. Note, $(\alpha, \mu_0, \sum_0, \Psi)$ are not random for our purposes.

The final Gaussian mixture model is fit through an EM algorithm until convergence. We use the mixture model that obtains the lowest BIC. Before determining the final model, I fit a Gaussian mixture model to the crabs dataset with 2, 3, and 4 initial clusters.

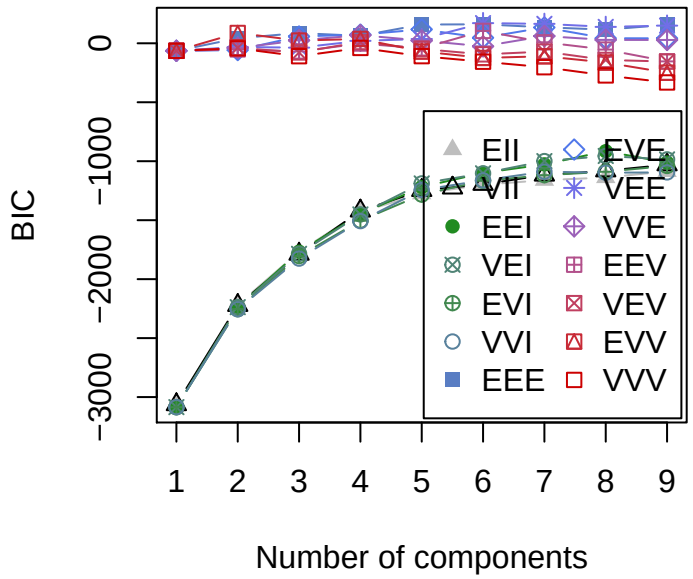

FL
RW
CL
CW
BD
−2 0 2



Based on the figures above, 2 or 4 clusters seems to match the data the best. Below, I plot the BIC curves for the different potential models.

```
## [1] "VEE"
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEE (ellipsoidal, equal shape and orientation) model with 6 components:
##
##  log.likelihood   n df      BIC      ICL
##        233.0476 214 55 170.9666 154.4732
##
## Clustering table:
##  1  2  3  4  5  6
## 21 28 36 59 27 43
```

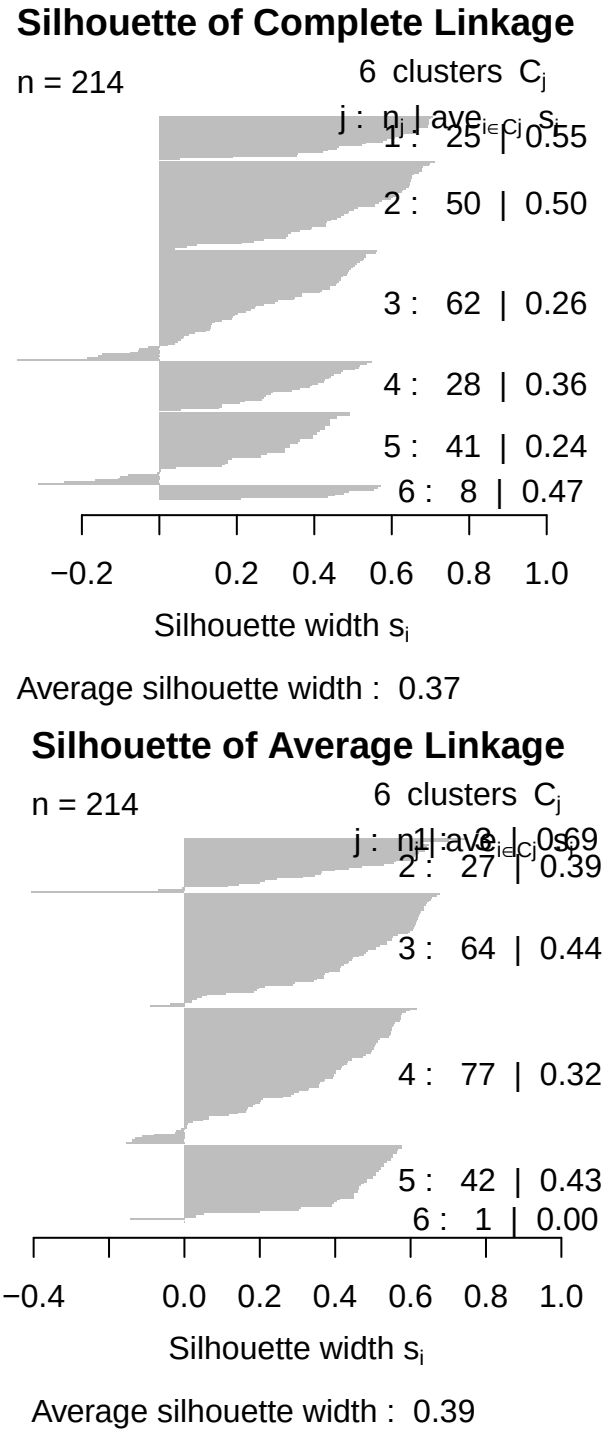


Based on the results of the plot above, I should use the "VEE" model which is defined as having a Gaussian mixture with the individual Gaussians having ellipsoidal, equal shape and orientation Gaussians. Additionally, the plot suggests that I use 6 clusters since this has the smallest BIC with the constraint of only 6 components. A potential explanation for having 6 clusters could be that age of the crab was not accounted for

Comparison

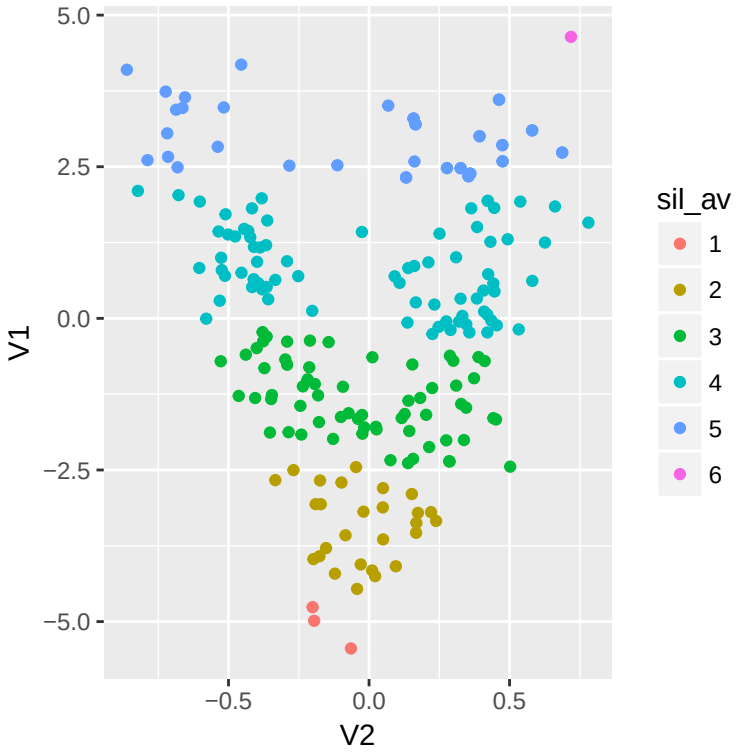
Finally, I refit the hierarchical clustering method, K-means and the Gaussian mixture model to have 6 clusters (from the BIC graph above). Since the selection criterion have suggested 6 clusters, I don't expect these to line up well with the pre-defined categories of Sex and Species.

I only re-examine the complete linkage and the average linkage for the hierarchical clustering method because those linkages were able to detect some amount of clustering.

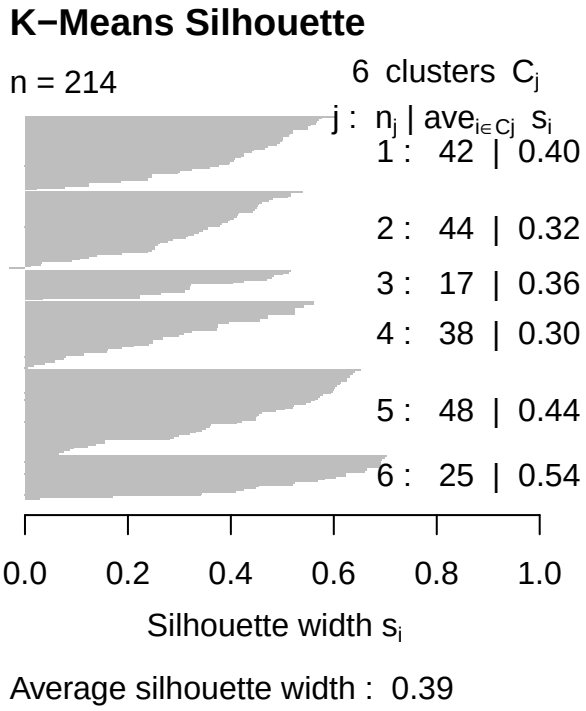


Not unexpectedly, 6 clusters seems to be high for the number of clusters in the crabs dataset. Only the first and second the clusters have a significant structure in the complete linkage; similarly, only the first cluster in the average linkage has a significant amount of structure. Average linkage appears to perform slightly better however with an average silhouette width of 0.39 as opposed to 0.37.

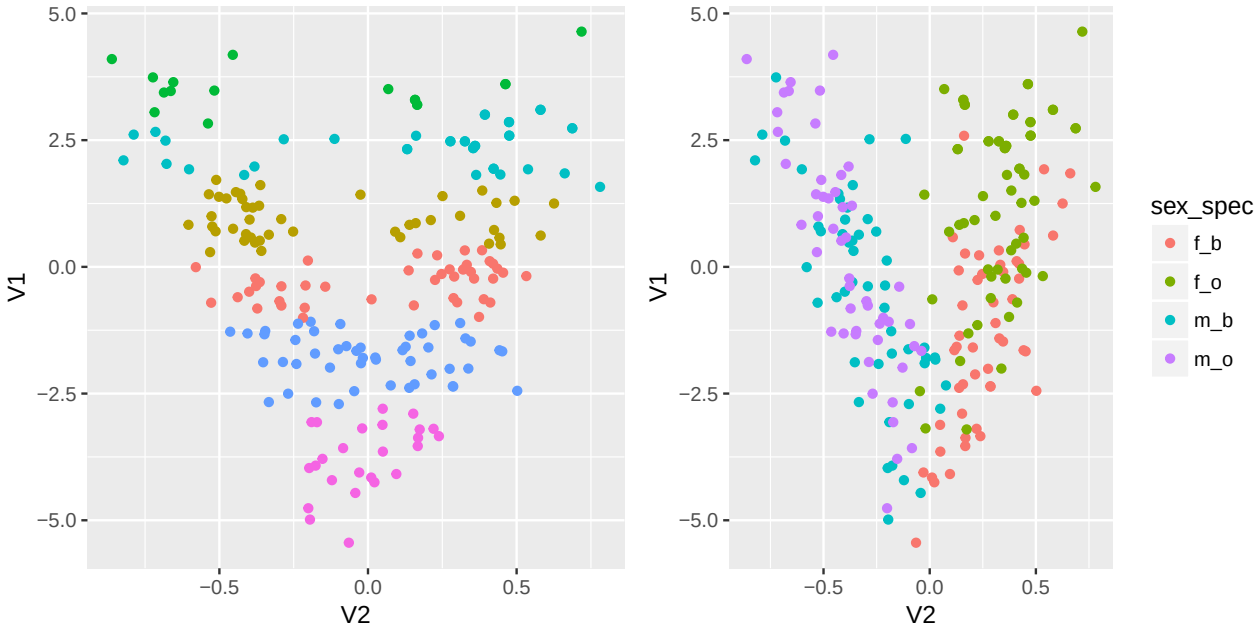
To compare to the natural clustering, I graph the results of the hierarchical cluster.



Next, I refit the K-means algorithm.

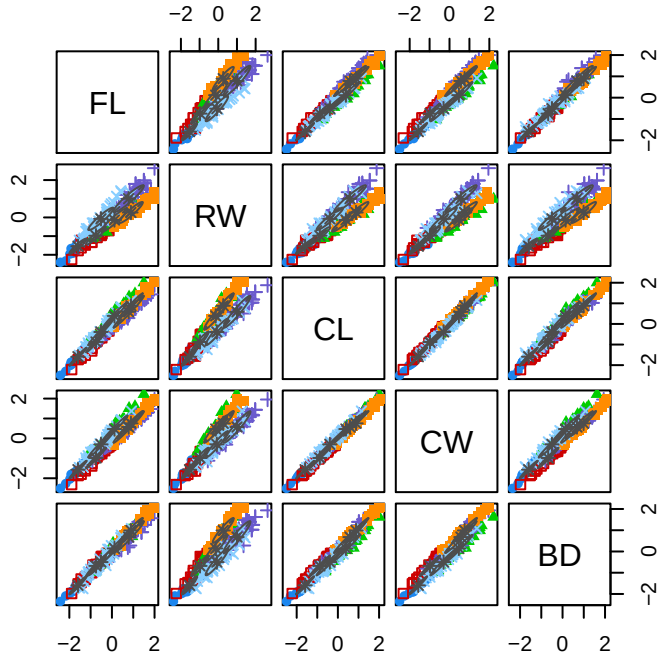


Looking at the silhouette plot of K-means above, all of the groups have a moderate structure and only one of the groups have outliers.



Looking at the graph of the clusters above for K-means, it does not align well with the predefined categories of Sex and Species. However, it would fit better with the carapace width cut off explored above.

Finally, I plot the results of a mixture model with 6 hidden clusters.



As we can see from the plot above, the mixture model does capture the separation seen with respect to the rear width of the crabs. It does somewhat capture the groupings of sex and species for the crabs; however, it appears to split up the sex and species category into more "refined" categories which lack interpretability.

```
library(knitr)
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = F, fig.align = "center",
                      fig.height = 4, fig.width = 4)

setwd("~/Documents/UMich/Classes/2018 Spring/STATS 503/Homework/HW5/")
```

```r
require(cluster)
library(factoextra)
library(mclust)
library(ggplot2)
library(GGally)
library(gridExtra)

crabs <- read.table("crabs.txt", header = T)

### Visualization
crabs_scaled <- as.data.frame(scale(crabs[,3:7], center = TRUE, scale = TRUE))
crabs_dis = dist(crabs_scaled)
crabs_mds = as.data.frame(cmdscale(crabs_dis, k=2))

# examine the data
q1 = ggplot(crabs_mds, aes(x=V2, y=V1)) + geom_point(alpha=0.6)
q2 = ggplot(crabs_scaled, aes(x=FL, y=CW)) + geom_point(alpha=0.6)
q3 = ggplot(crabs_scaled, aes(x=RW, y=CL)) + geom_point(alpha=0.6)
q4 = ggplot(crabs_scaled, aes(x=RW, y=CW)) + geom_point(alpha=0.6)
grid.arrange(q1, q2, q3, q4, ncol=2)


  # add back in the sex and species of the crabs for purposes of visualizing the mds results
  crabs_mds_vis <- crabs_mds
  crabs_mds_vis$sex <- crabs$Sex
  crabs_mds_vis$sex_l[crabs_mds_vis$sex == 1] <- "m"
  crabs_mds_vis$sex_l[crabs_mds_vis$sex == 2] <- "f"
  crabs_mds_vis$species <- crabs$Species
  crabs_mds_vis$species_l[crabs_mds_vis$species == 1] <- "b"
  crabs_mds_vis$species_l[crabs_mds_vis$species == 2] <- "o"
  crabs_mds_vis$sex_spec <- paste(crabs_mds_vis$sex_l,crabs_mds_vis$species_l, sep ="_")
  crabs_mds_vis$sex_spec <- as.factor(crabs_mds_vis$sex_spec)
  crabs_mds_sex <- ggplot(crabs_mds_vis, aes(x=V2, y=V1, colour = sex_spec)) + geom_point()
  crabs_mds_sex

  crabs_mds_vis$cw <- (crabs_scaled$CW <0)*1
  crabs_mds_vis$cw <- as.factor(crabs_mds_vis$cw)
  crabs_mds_cw <- ggplot(crabs_mds_vis, aes(x=V2, y=V1, colour = cw)) + geom_point()
  crabs_mds_cw


  crabs_scaled_vis <- crabs_scaled
  crabs_scaled_vis$sex <- crabs$Sex
  crabs_scaled_vis$sex_l[crabs_scaled_vis$sex == 1] <- "m"
  crabs_scaled_vis$sex_l[crabs_scaled_vis$sex == 2] <- "f"
  crabs_scaled_vis$species <- crabs$Species
  crabs_scaled_vis$species_l[crabs_scaled_vis$species == 1] <- "b"
  crabs_scaled_vis$species_l[crabs_scaled_vis$species == 2] <- "o"
  crabs_scaled_vis$sex_spec <- paste(crabs_scaled_vis$sex_l,crabs_scaled_vis$species_l, sep ="_")
  crabs_scaled_vis$sex_spec <- as.factor(crabs_scaled_vis$sex_spec)

  ggpairs(crabs_scaled_vis[,c(1:5,10)], mapping = aes(colour = sex_spec))
```

```
crabs_sing = agnes(crabs_scaled, diss=FALSE, method='single')
crabs_comp = agnes(crabs_scaled, diss=FALSE, method='complete')
crabs_aver = agnes(crabs_scaled, diss=FALSE, method='average')

sil_sing = silhouette(cutree(crabs_sing, k=2), crabs_dis)
plot(sil_sing, main = "Silhouette of Single Linkage")

sil_comp = silhouette(cutree(crabs_comp, k=2), crabs_dis)
plot(sil_comp, main = "Silhouette of Complete Linkage")


sil_comp = silhouette(cutree(crabs_comp, k=3), crabs_dis)
plot(sil_comp, main = "Silhouette of Complete Linkage")

sil_comp = silhouette(cutree(crabs_comp, k=4), crabs_dis)
plot(sil_comp, main = "Silhouette of Complete Linkage")


sil_av = silhouette(cutree(crabs_aver, k=2), crabs_dis)
plot(sil_av, main = "Silhouette of Average Linkage")


sil_av = silhouette(cutree(crabs_aver, k=3), crabs_dis)
plot(sil_av, main = "Silhouette of Average Linkage")
sil_av = silhouette(cutree(crabs_aver, k=4), crabs_dis)
plot(sil_av, main = "Silhouette of Average Linkage")

set.seed(77756)

crabs_clust_kmeans = kmeans(crabs_scaled, 2, nstart = 10, iter.max=100)
crabs_mds_temp = cbind(crabs_mds, as.factor(crabs_clust_kmeans$cluster))
names(crabs_mds_temp) = c('V1', 'V2', 'cluster')

ggplot(crabs_mds_temp, aes(x=V2, y=V1, color=cluster)) +
  geom_point() + theme(legend.position="none")

set.seed(77756)

crabs_mds = as.data.frame(cmdscale(crabs_dis, k=3))

crabs_clust_kmeans = kmeans(crabs_scaled, 3, nstart = 10, iter.max=100)
crabs_mds_temp = cbind(crabs_mds, as.factor(crabs_clust_kmeans$cluster))
names(crabs_mds_temp) = c('V1', 'V2','V3', 'cluster')

ggplot(crabs_mds_temp, aes(x=V2, y=V1, color=cluster)) +
  geom_point() + theme(legend.position="none")

crabs_mds = as.data.frame(cmdscale(crabs_dis, k=4))

crabs_clust_kmeans = kmeans(crabs_scaled, 4, nstart = 10, iter.max=100)
crabs_mds_temp = cbind(crabs_mds, as.factor(crabs_clust_kmeans$cluster))
names(crabs_mds_temp) = c('V1', 'V2','V3','V4', 'cluster')
```

```
ggplot(crabs_mds_temp, aes(x=V2, y=V1, color=cluster)) +
  geom_point() + theme(legend.position="none")


crabs_mix <- Mclust(crabs_scaled, G = 2)
plot(crabs_mix, what = "classification")

crabs_mix <- Mclust(crabs_scaled, G = 3)
plot(crabs_mix, what = "classification")

crabs_mix <- Mclust(crabs_scaled, G = 4)
plot(crabs_mix, what = "classification")


crabs_mix <- Mclust(crabs_scaled)
crabs_mix$modelName
summary(crabs_mix)
plot(crabs_mix, what = "BIC")


sil_comp = silhouette(cutree(crabs_comp, k=6), crabs_dis)
plot(sil_comp, main = "Silhouette of Complete Linkage")

sil_av = silhouette(cutree(crabs_aver, k=6), crabs_dis)
plot(sil_av, main = "Silhouette of Average Linkage")


crabs_mds$sil_av <- sil_av[,1]
crabs_mds$sil_av <- as.factor(crabs_mds$sil_av)
 ggplot(crabs_mds, aes(x=V2, y=V1, colour = sil_av)) + geom_point()

set.seed(77756)

crabs_mds = as.data.frame(cmdscale(crabs_dis, k=6))

crabs_clust_kmeans = kmeans(crabs_scaled, 6, nstart = 10, iter.max=100)

plot(silhouette(crabs_clust_kmeans$cluster, crabs_dis), main = "K-Means Silhouette")


crabs_mds_temp = cbind(crabs_mds, as.factor(crabs_clust_kmeans$cluster))
names(crabs_mds_temp) = c('V1', 'V2','V3','V4','V5','V6', 'cluster')

crabs_mds_final <- ggplot(crabs_mds_temp, aes(x=V2, y=V1, color=cluster)) +
  geom_point() + theme(legend.position="none")

grid.arrange(crabs_mds_final, crabs_mds_sex, nrow = 1)


crabs_mix <- Mclust(crabs_scaled, G = 6)
plot(crabs_mix, what = "classification")
```