Team 1 Chip2 Analysis

2025-03-03

Our team was given the task: Clustering/Co-clustering of Ige and link with the clinic: can we explain the groups in relation to the clinic?

Prepare data

```
# Reading the dataset
chip2 <- read.csv("Data Set/ACC_2023_Chip2.csv")
```

To process the data we split the dataset into allergens and clinical data. We imputed missing values with 9's.

```
# raw data
chip2 <- chip2 %>% select(-COFACTEURS.généraux,
                          -Clinique.par.Aliment,
                          -ATCD.d.anaphylaxie.alimentaire..AA.)

# chip 2 data
chip2_IgE <- chip2%>%
  select(`Act.d.1`:`Can.f.6`) %>% {
    .[is.na(.)] <- 9
    .
  }

# clinical data
chip2_clinical <- chip2 %>%
  select('Code.local':'TPO_severite')
```

Perform the PCA

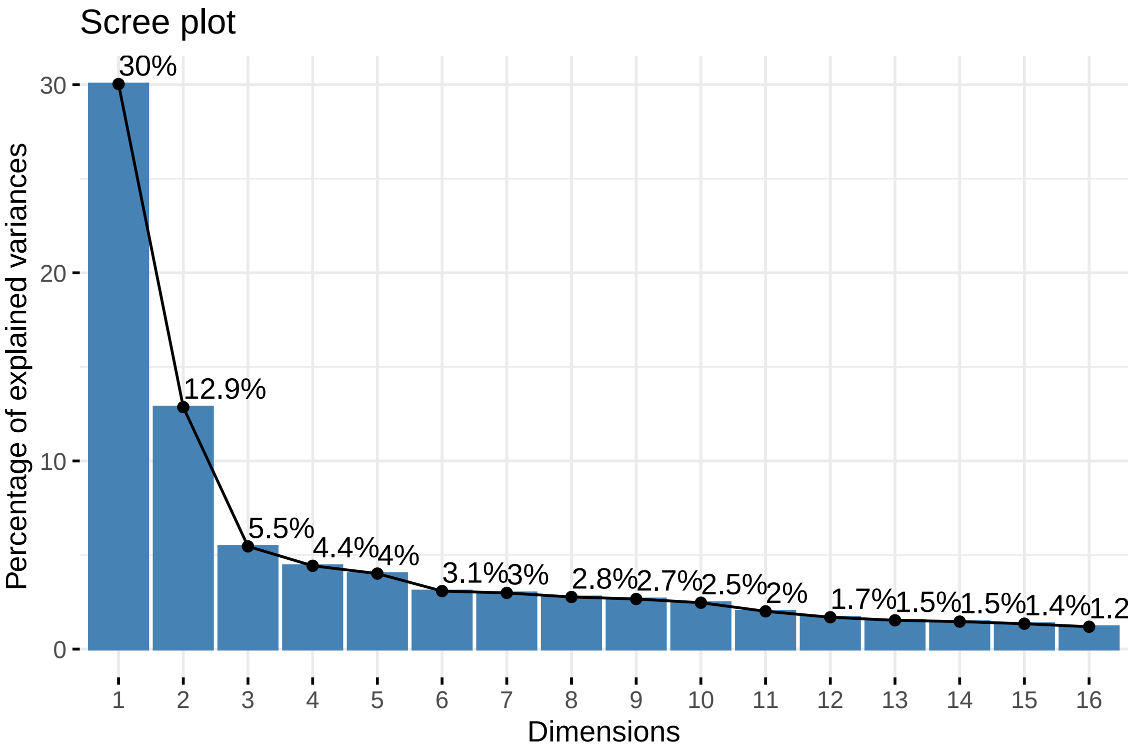
We did PCA using the following code

```
res <- PCA(chip2_IgE, graph = FALSE, ncp = 16)
```

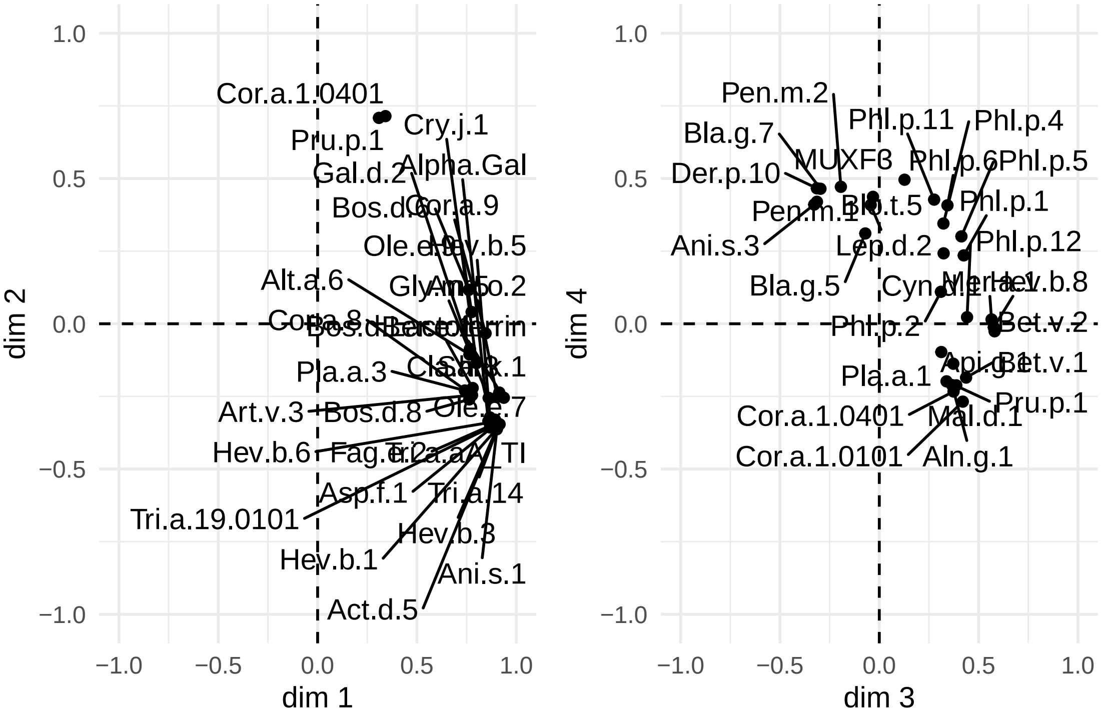
and the we checked how many components should we use to explain at least 80% of the variance. The screeplot shows that all the PCs explain a small portion of the variability, in fact we need 16 PCs to explain 80% of the variance. Eg: the first four PCs only explain 52.8% of the variance.

```
n_vars <- which(cumsum(as.vector(res$eig[,2])) > 80)[1]

factoextra::fviz_screeplot(res, addlabels = TRUE, ncp = n_vars)
```



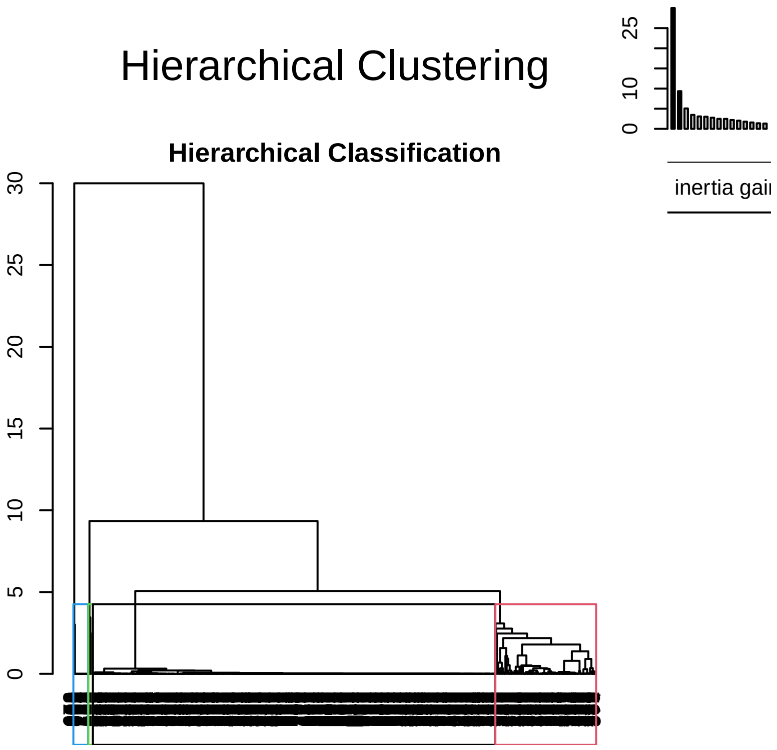
The first principal component shows us that all variables are positively correlated. By analysing the second, third and fourth components, we managed to identify some clusters of IgEs. The biggest cluster that we identified had high value on the first PC. further investigation needs to be done in order to understand which are the clusters that are identified by PCA and if the IgEs that belong to those clusters belong to the same families.

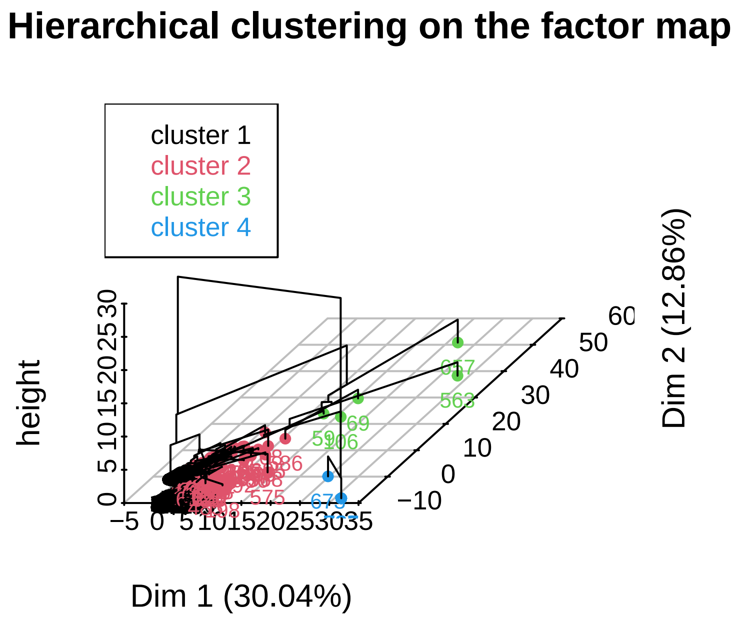


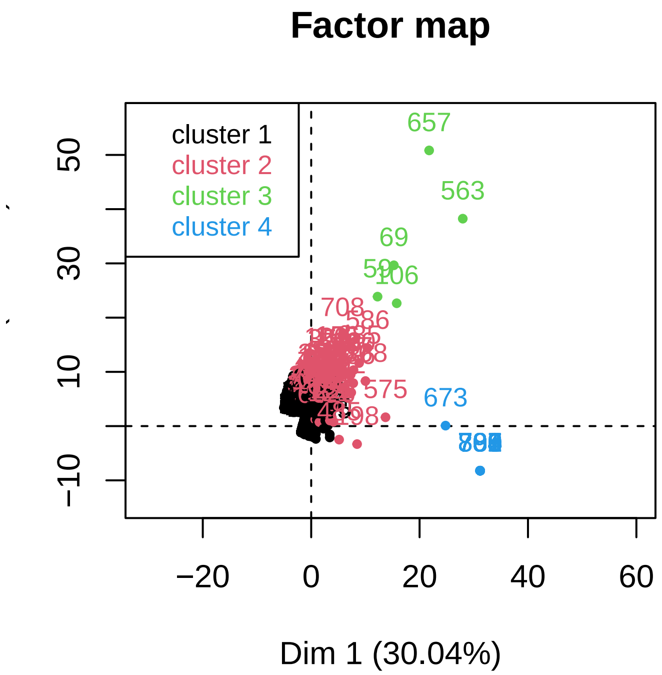
Hierarchical clustering

We then performed hierarchical clustering of individuals based on their IgEs values. The plot shows the emergence of a big cluster and three smaller clusters. Further investigation showed that the big cluster was made of individuals with zero to low values in almost all IgEs. While the other two smaller clusters were each composed of individuals with high values for similar IgEs. The third cluster had too few observations, so we decided to remove it afterwards.

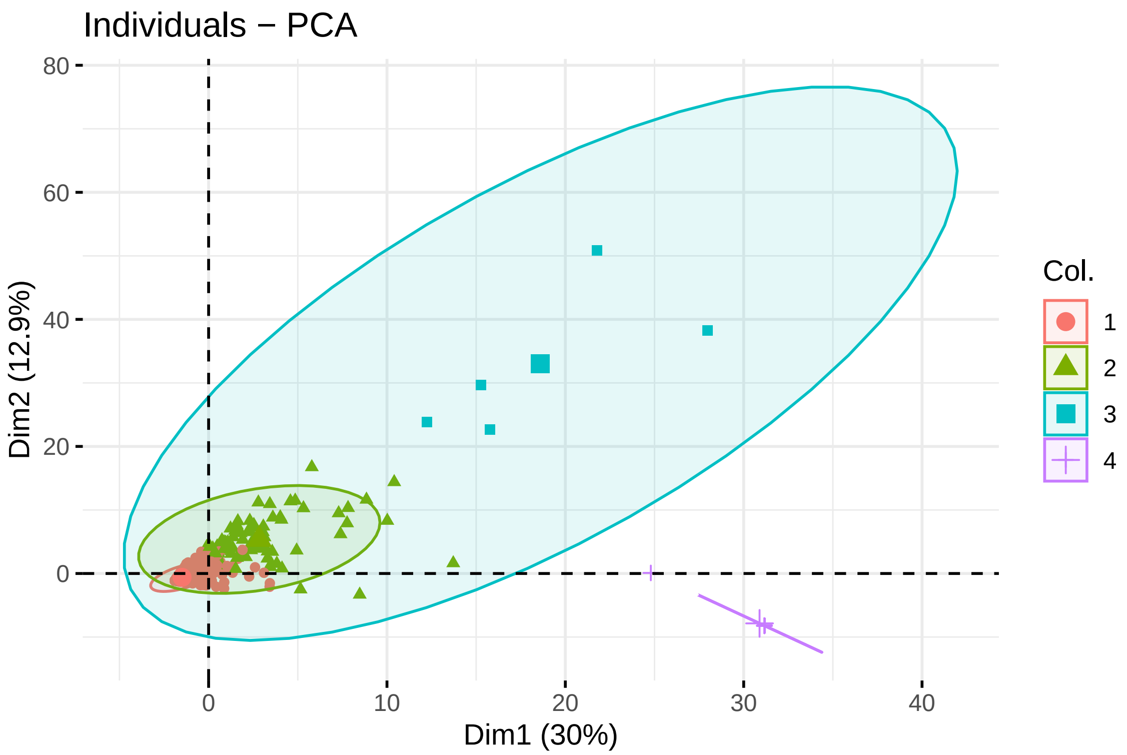
```
res.hcpc <- HCPC(res, nb.clust = 4)
```


Hierarchical Clustering
25
10
0
inertia gai
Hierarchical Classification
30
25
20
15
10
5
0

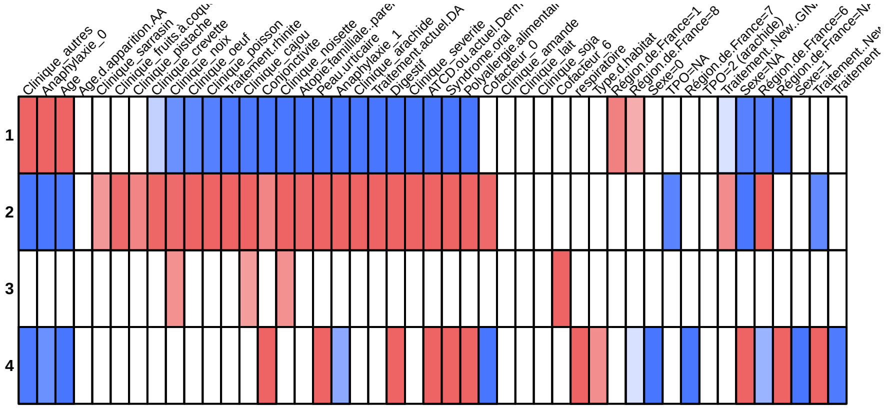

Hierarchical clustering on the factor map
cluster 1
cluster 2
cluster 3
cluster 4
height
0 5 10 15 20 25 30
−5 0 5 10 15 20 25 30 35
−10 0 10 20 30 40 50 60
Dim 1 (30.04%)
Dim 2 (12.86%)
617
563
69
59
106
686
575
679



```
clust <- res.hcpc$data.clust[,"clust"]
```



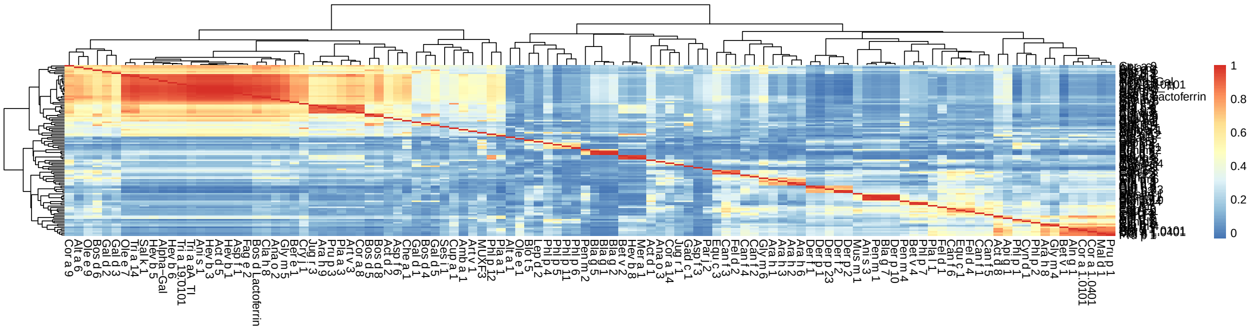
We then examined the values of clinical variables for each cluster. If a variable is coloured blue for a cluster, it indicates that the average value of the variable in that cluster is lower than the overall average. If it's coloured red, the average value of the variable in that cluster is higher than the total average. We found that the big cluster has low values on almost all clinical variables and high values on age. Conversely, the other two clusters of interest (second and forth) have lower age on average but higher values on clinical variables: the second cluster has overall higher values on all clinical symptoms in particular anaphylaxis and skin reactions, while the fourth cluster had higher values in digestive symptoms, rhinitis and conjunctivitis.



Clustering on correlation matrix

We first computed the correlation matrix and presented the corresponding heatmap to understand how the variables of IgE are related to each other. We can see that there are mainly three groups: one with highly correlated variables (red square in the upper diagonal), another with non-correlated variables (small blue square in the middle diagonal) and low correlated variables (light blue and orange square in the lower diagonal).

```r
# 1. Compute the correlation matrix and plot
cor_matrix <- cor(chip2_IgE)
pheatmap(cor_matrix)
```



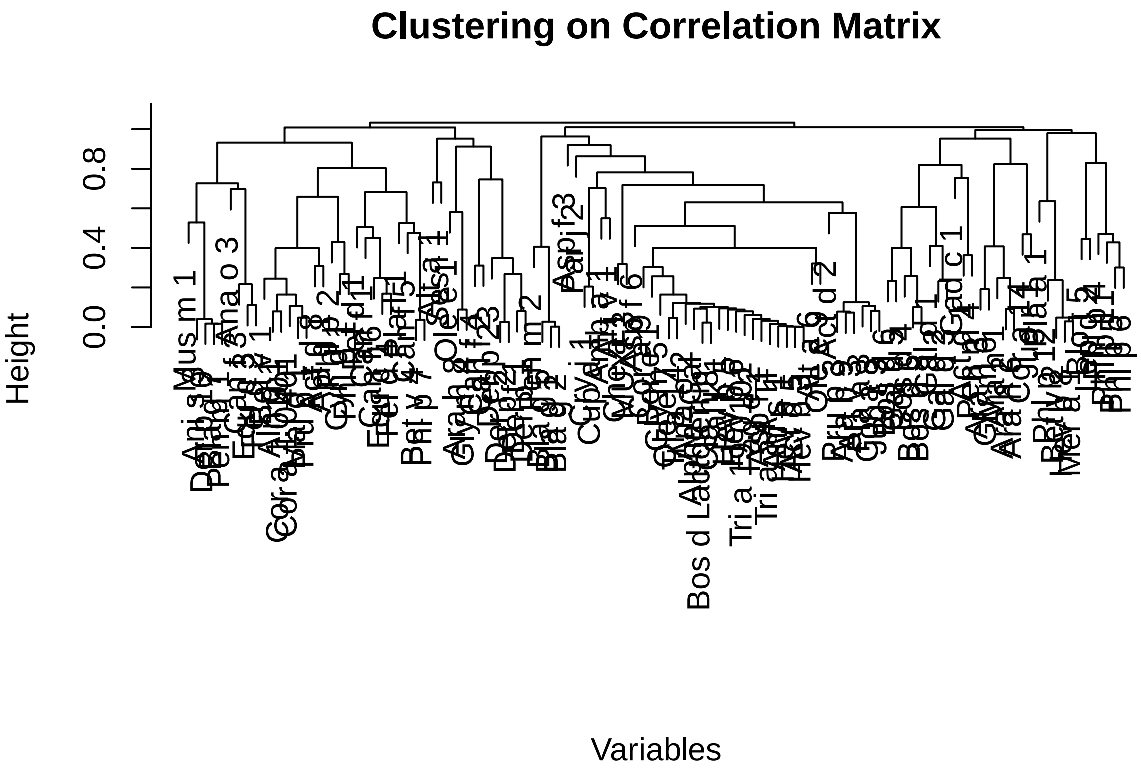
Second, we converted the correlation matrix into a distance matrix and performed the hierarchical clustering with three groups as input. So we can understand how these three variable groups are formed.

```
# 2. Convert to a Distance Matrix
# Using 1 - correlation as a distance measure
dist_matrix <- as.dist(1 - cor_matrix)

# 3. Perform Hierarchical Clustering
hc <- hclust(dist_matrix, method = "complete")
clusters_variables <- cutree(hc, k =3)

clusters_var <- data.frame(Variables=clusters_variables%>%names(), Clusters=clusters_variables%>%as.vec

# 4. Visualize the Dendrogram
plot(hc, main = "Clustering on Correlation Matrix",
     xlab = "Variables",
     sub = "")
```



Co-clustering

This step consists in performing clusters for the variables and individuals instead of clusters only for variables and only for individuals. This would help us understand the relationship between the two. Since before we discovered 3 main clusters for variables and 3 main clusters for individuals, we decided to set the number of row clusters and the number of column clusters equal to 3.

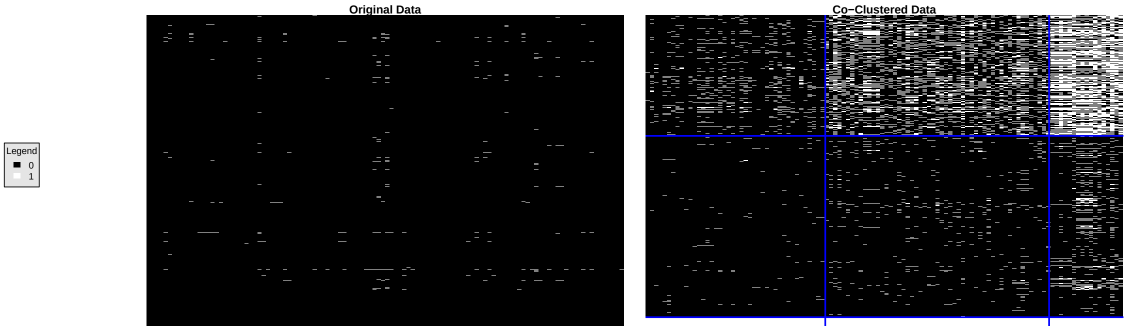
In order to implement co-clustering on binary matrices, we first had to perform a preprocessing min-max renormalization to make the IgE matrix binary. In this step, the data frame is restructured to join elements of the same cluster.

```
out <- coclusterBinary(((chip2_IgE-(min(chip2_IgE)))/
                            (max(chip2_IgE)-min(chip2_IgE))) %>%
                          as.matrix(), nbcocluster = c(3,3))
```

```
## Co-Clustering successfully terminated!
```

```
row_clusters <- out@rowclass
col_clusters <- out@colclass

clusters_var <- data.frame(Allergen=clusters_variables%>%names(), Clusters=col_clusters%>%as.vector())
plot(out)
```



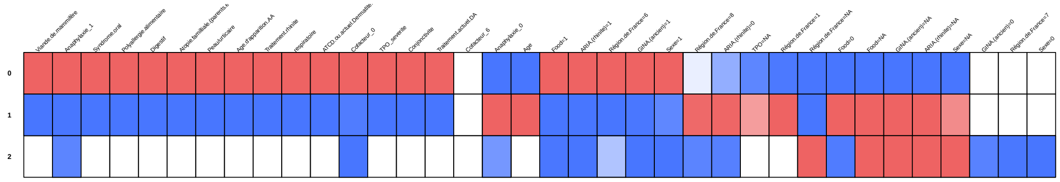
We collected the clusters on individuals generated by co-clustering and checked the clinical symptom patterns as before.

```
chip_without <- chip2_clinical[
  ,
  colSums(is.na(chip2_clinical)) != nrow(chip2_clinical)
]

chip_w <- chip_without%>% select(-`Mois du prélèvement`,
                                 -`Code local`,
                                 -`CHECK`,-`CODE PUCE`,
                                 -`Département de résidence`)

catd_res <- cbind.data.frame(chip_w, as.factor(row_clusters))

res_catdes <- catdes(catd_res,ncol(catd_res))
par(mfrow=c(1,1))
plot.catdes(res_catdes)
```



```
plt_dat <- plt_dat %>%
  left_join(clusters_var, by="Allergen")
```

We also collected the variable clusters from the co-clustering and analyzed the allergen types in the three clusters. We note that cluster 0 (line 3) is multi allergenic as it includes MUXT3. Cluster 1 (line 2) collects airway and contact allergens and finally cluster 2 is also multi allergenic collecting contact, airway and food allergies.

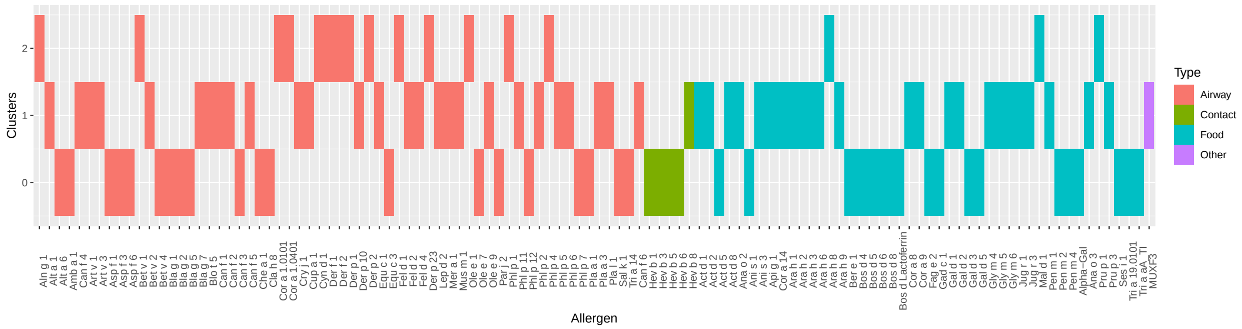
```
out <- coclusterBinary(((chip2_IgE-(min(chip2_IgE)))/
                          (max(chip2_IgE)-min(chip2_IgE))) %>%
                         as.matrix(), nbcocluster = c(3,3))
```

```
## Co-Clustering successfully terminated!
```

```
row_clusters <- out@rowclass
col_clusters <- out@colclass
clusters_var <- data.frame(Variables=clusters_variables%>%names(), Clusters=col_clusters%>%as.vector())
cols <- allergens %>%
  select(Type) %>%
  unique() %>%
  mutate(col = c("red", "blue", "yellow", "purple", "green"))
clusters_var <- clusters_var %>%
  rename(Allergen = "Variables")
plt_dat <- chip2_IgE %>%
  tidyr::gather(.,"Allergen","Value") %>%
  left_join(allergens, by = "Allergen") %>%
  left_join(cols, by = "Type") %>%
  left_join(clusters_var)
```

```
## Joining with ‘by = join_by(Allergen)‘
```

```
plt_dat %>%
  arrange(Type) %>%
  mutate(Allergen = factor(Allergen, levels = unique(Allergen))) %>%
  ggplot() +
  geom_tile(aes(x = Allergen, y = Clusters, fill = Type)) +
  theme(axis.text.x = element_text(angle = 90),
        axis.text.y = element_text())
```


Clusters
2
1
0
Aln g 1
Alt a 1
Alt a 6
Amb a 1
Can f 4
Art v 1
Art v 3
Asp f 1
Asp f 3
Asp f 6
Bet v 1
Bet v 2
Bet v 4
Bla g 1
Bla g 2
Bla g 5
Bla g 7
Blo t 5
Can f 1
Can f 2
Can f 3
Can f 5
Che a 1
Cla h 8
Cor a 1.0101
Cor a 1.0401
Cry j 1
Cup a 1
Cyn d 1
Der f 1
Der f 2
Der p 1
Der p 10
Der p 2
Equ c 1
Equ c 3
Fel d 1
Fel d 2
Fel d 4
Der p 23
Lep d 2
Mer a 1
Mus m 1
Ole e 1
Ole e 7
Ole e 9
Par j 2
Phl p 1
Phl p 11
Phl p 12
Phl p 2
Phl p 4
Phl p 5
Phl p 6
Phl p 7
Pla a 1
Pla a 3
Pla l 1
Sal k 1
Tri a 14
Can f 6
Hev b 1
Hev b 3
Hev b 5
Hev b 6
Hev b 8
Act d 1
Act d 2
Act d 5
Act d 8
Ana o 2
Ani s 1
Ani s 3
Api g 1
Cor a 14
Ara h 1
Ara h 2
Ara h 3
Ara h 6
Ara h 8
Ara h 9
Ber e 1
Bos d 4
Bos d 5
Bos d 6
Bos d 8
Bos d Lactoferrin
Cor a 8
Cor a 9
Fag e 2
Gad c 1
Gal d 1
Gal d 2
Gal d 3
Gal d 5
Gly m 4
Gly m 5
Gly m 6
Jug r 1
Jug r 3
Mal d 1
Pen m 1
Pen m 2
Pen m 4
Alpha−Gal
Ana o 3
Pru p 1
Pru p 3
Ses i 1
Tri a 19.0101
Tri a aA_TI
MUXF3
Allergen
Type
Airway
Contact
Food
Other