Chemicals Clustering

Guannan Shen

January 9, 2019

Contents

```
####################### Set up workspace
rm(list = ls())
library(knitr)
library(tidyverse)
library(magrittr)
library(stats)
options(stringsAsFactors = F)
options(dplyr.width = Inf)
getwd()
```

```
## [1] "/home/guanshim/Documents/gitlab/ECCHO_github/DataRaw"
```

```
## not in function
"%nin%" <- Negate("%in%")

# ######## clean memory ##################### rm(list =
# ls()) gc() is(dds) slotNames(dds)
```

1 Unsupervised

```
## import the chems data
phthal <- read.csv("phthal_di.csv")
chems <- t(phthal[, -1])
colnames(chems) <- phthal$PID

## take a look
head(chems[, 1:5])
```

```
##               10611 10645 10652 10654 10661
## MCOCH_ng_ml     0.2   0.0   0.0   0.0   0.0
## MHBP_ng_ml      1.7   0.7   0.6   0.1   5.3
## MHiBP_ng_ml     2.3   0.7   1.4   0.3  29.5
## MHiNCH_ng_ml    0.1   0.3   0.2   0.1   0.4
## MBP_ng_ml      14.1   3.1   3.9   0.4  37.9
## MBzP2_ng_ml    39.7   1.7   1.1   0.2   5.9
```
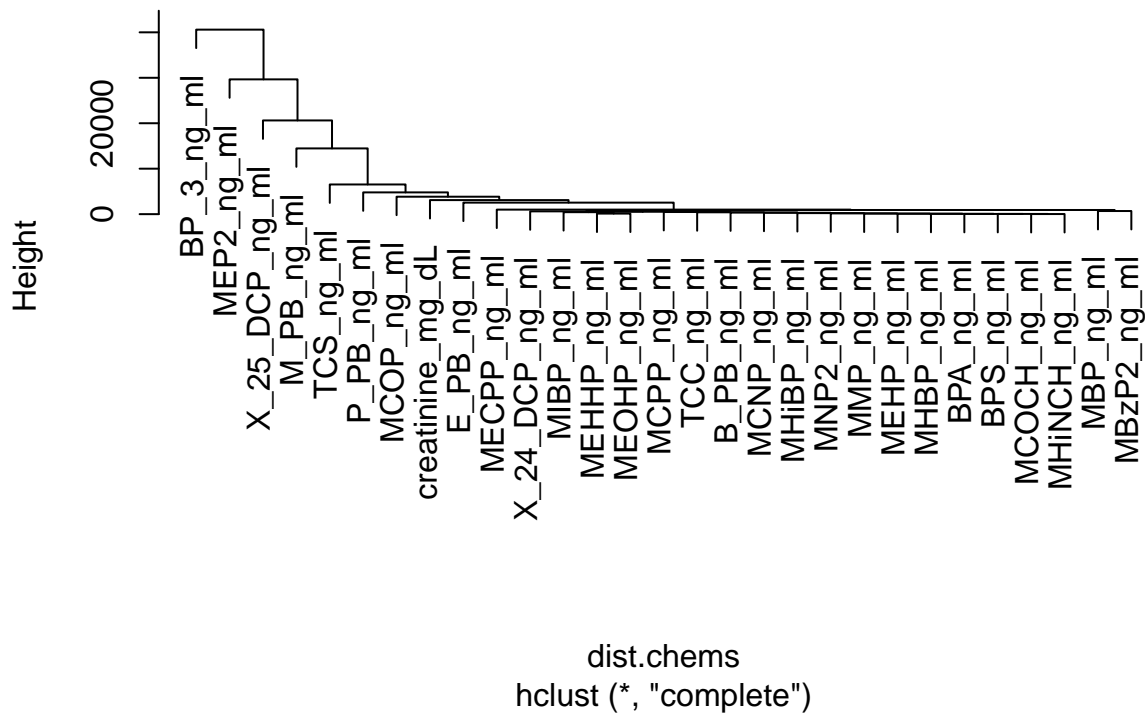
```
dim(chems)
```

```
## [1]  29 446
```

```
length(phthal$PID)
```

```
## [1] 446
```

```
## for hier clustering
chems <- as.matrix(chems)
dist.chems <- dist(chems, method = "euclidean")
plot(hclust(dist.chems, method = "complete"), main = "Cluster Dendrogram of Chemicals")
```
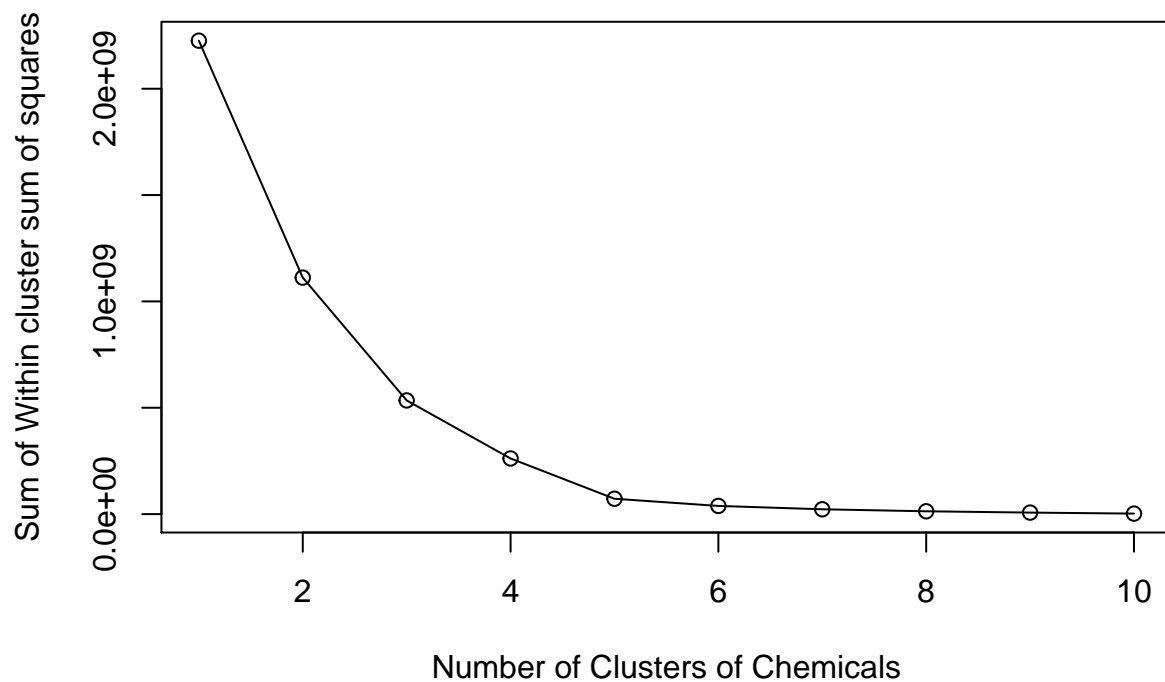
Cluster Dendrogram of Chemicals

dist.chems
hclust (*, "complete")

```
## kmeans cluster

## kmeans and try k
maxk <- 10
k_c <- 1:maxk
k_sws <- NULL

k_wsm <- NULL
for (i in k_c) {
    km <- kmeans(chems, i, iter.max = 10)
    k_sws[i] <- sum(km$withinss)
    if (i == 1) {
        k_wsm <- c(km$withinss)
    } else {
        k_wsm <- c(k_wsm, km$withinss)
    }
}
plot(k_c, k_sws, type = "o", xlab = "Number of Clusters of Chemicals",
    ylab = "Sum of Within cluster sum of squares")
```

```
km.chem5 <- kmeans(chems, 5, iter.max = 10)
km.chem5$cluster
```

```
##         MCOCH_ng_ml           MHBP_ng_ml          MHiBP_ng_ml         MHiNCH_ng_ml
##                   4                    4                    4                    4
##          MBP_ng_ml          MBzP2_ng_ml           MCNP_ng_ml           MCOP_ng_ml
##                   4                    4                    4                    4
##          MCPP_ng_ml          MECPP_ng_ml          MEHHP_ng_ml           MEHP_ng_ml
##                   4                    4                    4                    4
##         MEOHP_ng_ml           MEP2_ng_ml            MMP_ng_ml           MNP2_ng_ml
##                   4                    5                    4                    4
##         MIBP_ng_ml  creatinine_mg_dL      X_24_DCP_ng_ml       X_25_DCP_ng_ml
##                   4                    4                    4                    2
##          B_PB_ng_ml           BP_3_ng_ml            BPA_ng_ml            BPS_ng_ml
##                   4                    1                    4                    4
##          E_PB_ng_ml           M_PB_ng_ml           P_PB_ng_ml            TCC_ng_ml
##                   4                    3                    4                    4
##           TCS_ng_ml
##                   4
```