Case Study 1: Clustering the epileptic.qol Dataset

K means clustering using kml3d package

```
# install.packages("joineRML")
library(joineRML)
data(epileptic.qol)
# convert days to months
epileptic.qol$time_month <- epileptic.qol$time/30.25
# sort by id and time
epileptic.qol <- epileptic.qol[order(epileptic.qol$id,epileptic.qol$time_month),]

# scaling the clustering variables prior to analysis
epileptic.qol$anxiety_scale <- as.numeric(scale(epileptic.qol$anxiety))
epileptic.qol$depress_scale <- as.numeric(scale(epileptic.qol$depress))
epileptic.qol$aep_scale <- as.numeric(scale(epileptic.qol$aep))
```

k-means clustering (kml3d package)

```
# install.packages("kml3d")
library(kml3d)
```

```
## Warning: package 'rgl' was built under R version 4.2.2
```

```
# the data is in long format (each individual corresponds to multiple rows)
head(epileptic.qol)[,c(1,5,6,7,8)]
```

```
##   id time anxiety depress aep
## 1  1  147      11      14  43
## 2  1  259      12      12  51
## 3  1  519      20      21  63
## 4  1  906      17      20  53
## 5  2  134      19      13  45
## 6  2  258      21      16  50
```

```
N <- length(unique(epileptic.qol$id))    # number of individuals
n.obs <- table(epileptic.qol$id)         # number of observations
visit <- NULL
for (i in 1:N){visit <- c(visit,1:n.obs[i])}
epileptic.qol$visit <- visit
epileptic.qol <- as.data.frame(epileptic.qol)
#============================================================================================#
# kml3d package requires the data to be wide format (each individual corresponds to one row)
# the following codes transform the data from long format to wide format
#============================================================================================#
epileptic.qol.wide <- reshape(epileptic.qol[,c("id","anxiety_scale",
                                                 "depress_scale", "aep_scale","visit")],
                              idvar = "id", timevar = "visit", direction = "wide", sep="_")
#============================================================================================#
```

```r
# kml3d package requires the data to be complete (i.e., no missing values)
# for data with missingness, the following codes can be used
# for imputation prior to the cluster analysis
#==========================================================================================#
set.seed(3342)
# use ?imputation to see available imputation methods
epileptic.qol.wide.imp <- imputation(as.matrix(epileptic.qol.wide[,-1]),
                                     method = "linearInterpol.bisector")
# convert the object to a data.frame
epileptic.qol.wide.imp <- as.data.frame(epileptic.qol.wide.imp)
epileptic.qol.wide.imp$id <- epileptic.qol.wide$id

# performing K-means clustering
cldPreg <- cld3d(epileptic.qol.wide.imp,
            idAll=epileptic.qol.wide.imp$id,
            time = c(0,3,12,24),
                 varNames = c("Anxiety","Depress","Liverpool Adverse Events Profile"),
                 timeInData = list(anxiety =c(1,4,7,10),    # specify the columns of variables
                                    depress= c(2,5,8,11),
                                    aep= c(3,6,9,12)))

kml3d(cldPreg, nbClusters = 2:8)
```

```
##  ~ Fast KmL3D ~
## ****************************************************************************************************S
## 100
## ****************************************
## S
```

```r
# extracting bic for models with K=2 to 8;
# other criteria are can also be extracted in a similar manner
bic <- rbind( BIC_Keq2 = cldPreg@c2[[1]]@criterionValues[6],
            BIC_Keq3 = cldPreg@c3[[1]]@criterionValues[6],
            BIC_Keq4 = cldPreg@c4[[1]]@criterionValues[6],
            BIC_Keq5 = cldPreg@c5[[1]]@criterionValues[6],
            BIC_Keq6 = cldPreg@c6[[1]]@criterionValues[6],
            BIC_Keq7 = cldPreg@c7[[1]]@criterionValues[6],
            BIC_Keq8 = cldPreg@c8[[1]]@criterionValues[6])
# a model with 2 clusters (K=2) has the lowest BIC
num.clust.kml3d <-  which.min(bic) + 1 ; num.clust.kml3d
```

```
## [1] 2
```

```r
# obtain/extract the clusters using the getClusters() function
cluster.kml3d  <- getClusters(cldPreg, num.clust.kml3d)
cluster.kml3d <- as.numeric(cluster.kml3d); sum(table(cluster.kml3d))
```

```
## [1] 544
```

```r
# process the cluster variable and merge it back to the original data
per <- paste(round(100*table(cluster.kml3d)/N,1),"%",sep="")
cluster.kml3d <- factor(cluster.kml3d,
                        labels=paste("Cluster ",1:num.clust.kml3d," (",per,")",sep=""))

 # Keep last observation per id
dnew_uq <- epileptic.qol[!duplicated(epileptic.qol$id, fromLast=TRUE),]
```

```r
dat.cluster <- data.frame(dnew_uq$id,cluster.kml3d)
colnames(dat.cluster) <- c("id","cluster.kml3d")

dnew_uq <- merge(dnew_uq,dat.cluster,by="id")
dnew <- merge(epileptic.qol,dat.cluster,by="id")
```

```r
# making trajectory plots by clusters to visualize the results
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```r
#============================================================================================#
# plotting the trajectory of the first feature (anxiety) by cluster
#============================================================================================#
p1.kml3d <- ggplot(data =dnew, aes(x =time_month, y = anxiety,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,
                                   fill=cluster.kml3d))+
              geom_smooth(aes(x =time_month, y = anxiety,
                                      color=cluster.kml3d,
                                      linetype=cluster.kml3d,
                                      fill=cluster.kml3d),
                          method = "loess", linewidth = 3,se = FALSE,span=2)+
            ggtitle("kml3d")+
              theme_bw() +
            theme(legend.position = "none",
                plot.title = element_text(size = 15, face = "bold"),
                axis.text=element_text(size=15),
                axis.title=element_text(size=15),
                axis.text.x = element_text(angle = 0 ),
                strip.text.x = element_text(size = 15, angle = 0),
                strip.text.y = element_text(size = 15,face="bold")) +
                guides(fill=guide_legend(title=NULL,nrow = 1,byrow=TRUE),
                       color=guide_legend(title=NULL,nrow = 1,byrow=TRUE),
                        linetype=guide_legend(title=NULL,nrow = 1,byrow=TRUE)) +
            xlab("Time (months)") +
            ylab("anxiety") +
                ylim(c(min(dnew$anxiety,na.rm=TRUE),max(dnew$anxiety,na.rm=TRUE)))+
                                 scale_color_manual(values=c("green", "black"))+
                                 scale_fill_manual(values=c("green", "black"))
#============================================================================================#
# plotting the trajectory of the second feature (depress) by cluster
#============================================================================================#
p2.kml3d <- ggplot(data =dnew, aes(x =time_month, y = depress,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,
                                   fill=cluster.kml3d))+
              geom_smooth(aes(x =time_month, y = depress,
                                    color=cluster.kml3d,
                                    linetype=cluster.kml3d,
                                    fill=cluster.kml3d),
                          method = "loess", linewidth= 3,se = FALSE,span=2)+
            ggtitle("kml3d")+
              theme_bw() +
                theme(legend.position = "none",
                plot.title = element_text(size = 15, face = "bold"),
                axis.text=element_text(size=15),
                axis.title=element_text(size=15),
                axis.text.x = element_text(angle = 0 ),
                strip.text.x = element_text(size = 15, angle = 0),
                strip.text.y = element_text(size = 15,face="bold")) +
                guides(fill=guide_legend(title=NULL,nrow = 1,byrow=TRUE),
                       color=guide_legend(title=NULL,nrow = 1,byrow=TRUE),
                        linetype=guide_legend(title=NULL,nrow = 1,byrow=TRUE)) +
              xlab("Time (months)") +
            ylab("depress") +
                ylim(c(min(dnew$depress,na.rm=TRUE),max(dnew$depress,na.rm=TRUE)))+
                                 scale_color_manual(values=c("green", "black"))+
                                 scale_fill_manual(values=c("green", "black"))
#===========================================================================================#
# plotting the trajectory of the third feature (aep) by cluster
#===========================================================================================#
p3.kml3d <- ggplot(data =dnew, aes(x =time_month, y = aep,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,
                                   fill=cluster.kml3d))+
                geom_smooth(aes(x =time_month, y = aep,
                                color=cluster.kml3d,
                                linetype=cluster.kml3d,
                                fill=cluster.kml3d),
                          method = "loess", linewidth = 3,se = FALSE,span=2)+
            ggtitle("kml3d")+
                theme_bw() +
                theme(legend.position = "none",
                plot.title = element_text(size = 15, face = "bold"),
                axis.text=element_text(size=15),
                axis.title=element_text(size=15),
                axis.text.x = element_text(angle = 0 ),
                strip.text.x = element_text(size = 15, angle = 0),
                strip.text.y = element_text(size = 15,face="bold")) +
                guides(fill=guide_legend(title=NULL,nrow = 1,byrow=TRUE),
                       color=guide_legend(title=NULL,nrow = 1,byrow=TRUE),
                        linetype=guide_legend(title=NULL,nrow = 1,byrow=TRUE)) +
                xlab("Time (months)") +
            ylab("aep") +
                ylim(c(min(dnew$aep,na.rm=TRUE),max(dnew$aep,na.rm=TRUE)))+
                                 scale_color_manual(values=c("green", "black"))+
                                 scale_fill_manual(values=c("green", "black"))
#===========================================================================================#
# extract the figure legend
#===========================================================================================#
library(cowplot)
legend.kml3d <- get_legend(ggplot(data =dnew, aes(x =time_month, y = depress,
                                            color=cluster.kml3d,
                                            linetype=cluster.kml3d,
                                            fill=cluster.kml3d))+
                         geom_smooth(aes(x =time_month, y = depress, color=cluster.kml3d,
                                         linetype=cluster.kml3d,fill=cluster.kml3d),
                                     method = "loess", linewidth= 3,se = FALSE,span=2)+
                         ggtitle("kml3d")+
                         theme_bw() +
                         theme(legend.position = c(0.5,0.5),
                               legend.text = element_text(size = 12),
                               plot.title = element_text(size = 15, face = "bold"),
                               axis.text=element_text(size=15),
                               axis.title=element_text(size=15),
                               axis.text.x = element_text(angle = 0 ),
                               strip.text.x = element_text(size = 15, angle = 0),
                               strip.text.y = element_text(size = 15,face="bold")) +
                         guides(fill=guide_legend(title=NULL,nrow = 2,byrow=TRUE),
                                color=guide_legend(title=NULL,nrow = 2,byrow=TRUE),
                                linetype=guide_legend(title=NULL,nrow = 2,byrow=TRUE)) +
                         xlab("Time (months)") + ylab("depress") +
                         ylim(c(min(dnew$depress,na.rm=TRUE),
                                max(dnew$depress,na.rm=TRUE)))+
                         scale_color_manual(values=c("green", "black"))+
                         scale_fill_manual(values=c("green", "black"))
)
```

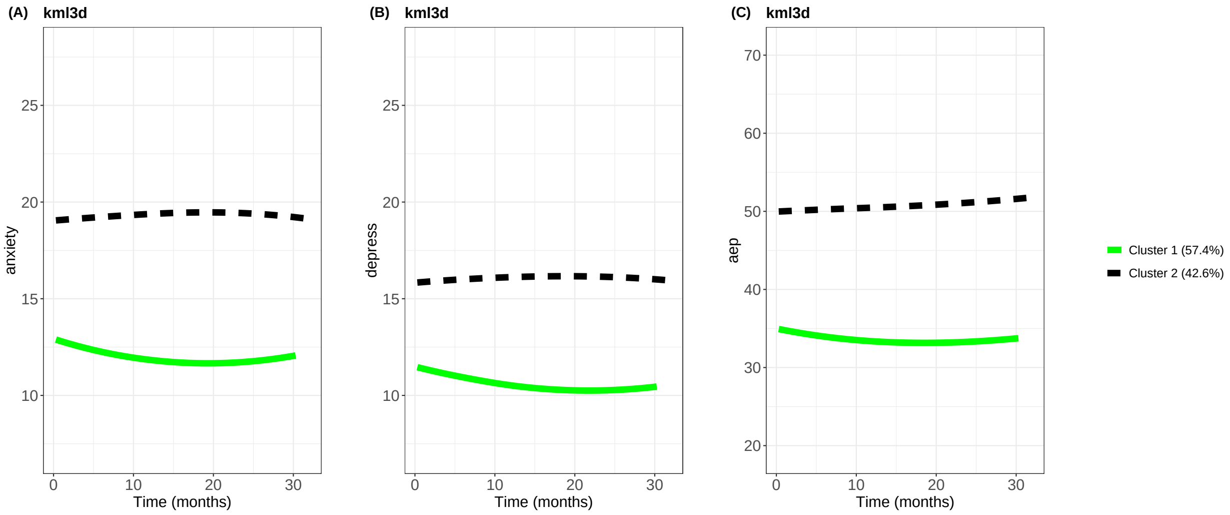
```
## Warning: Removed 53 rows containing non-finite values (`stat_smooth()`).
```

```
#==============================================================================================#
# use plot_grid from the cowplot package to arrange the figure panels
#==============================================================================================#
plot_grid(p1.kml3d,NULL,p2.kml3d,NULL,p3.kml3d,NULL,legend.kml3d,
          labels=c("(A)","", "(B)","","(C)","",""),
          ncol = 7,
          rel_heights = c(1,0.1),
          rel_widths = c(1,0.1,1,0.1,1,0.1,0.5))
```

```
## Warning: Removed 57 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 53 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 93 rows containing non-finite values (`stat_smooth()`).
```

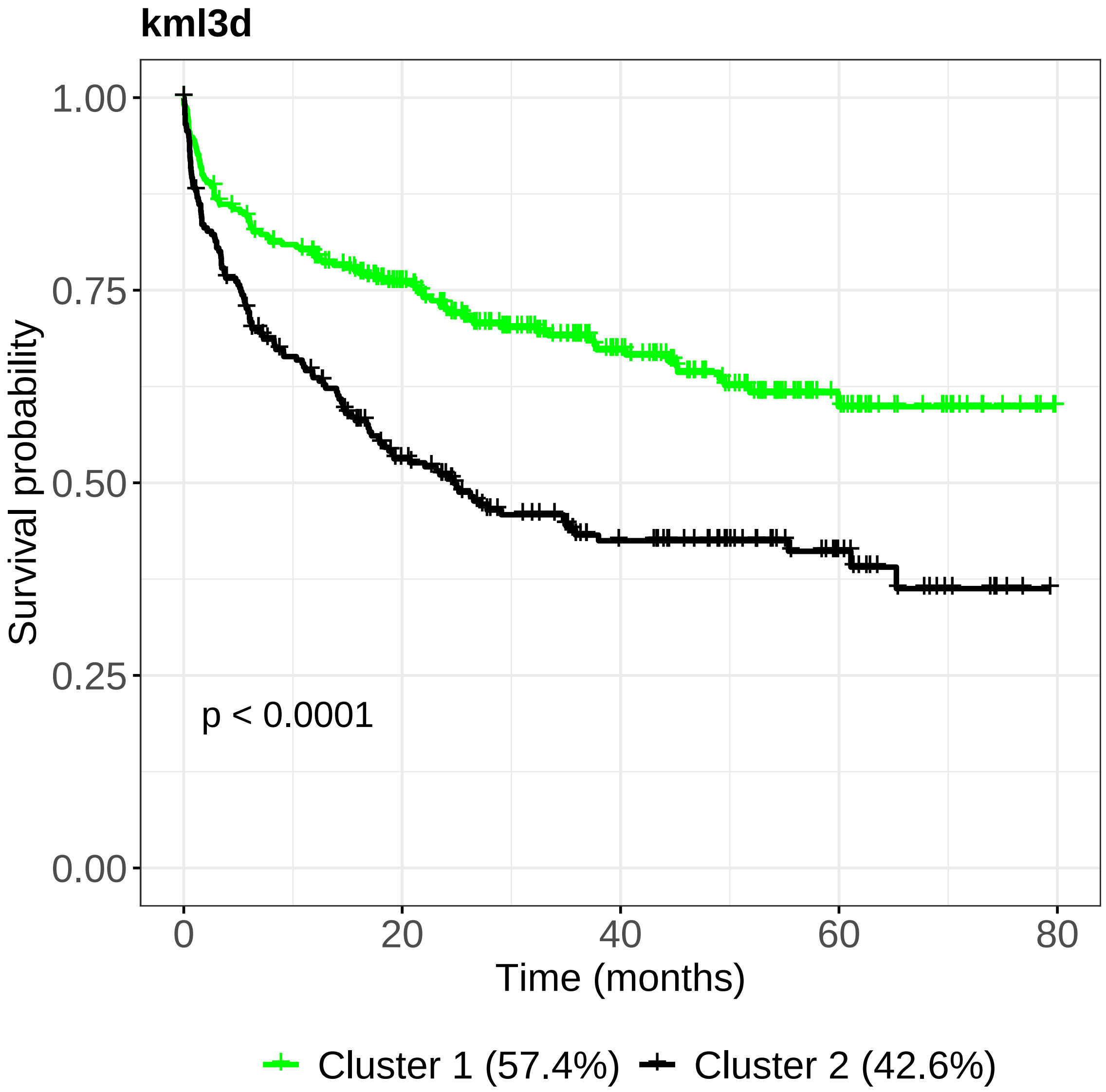


```r
library(survminer)
```

```
## Warning: package 'ggpubr' was built under R version 4.2.2
```

```r
library(survival)
# evaluate the Association between Clusters and Time to Treatment Failure
dnew_uq$with.time.month <- dnew_uq$with.time/30.25
fit <- survfit(Surv(with.time.month, with.status2) ~ cluster.kml3d, data = dnew_uq)
names(fit$strata) <-  paste("Cluster ",1:num.clust.kml3d," (",per,")",sep="")
gp_survival.kml3d <- ggsurvplot(fit, data = dnew_uq,  title = "kml3d",
                          risk.table = TRUE,
                                  risk.table.y.text.col = TRUE,
                                  pval = TRUE,
                          legend = "bottom", # conf.int = TRUE,
                          xlab = "Time (months)",
                                  legend.title="Clusters",
                          ggtheme =  theme_bw() + theme(legend.position ="none",
                                                        legend.title=element_blank(),
                                           plot.title = element_text(size = 15, face = "bold"),
                                           legend.text=element_text(size=15),
                                           axis.text=element_text(size=15),
                                           axis.title=element_text(size=15),
                                           strip.text.x = element_text(size=15),
                                           strip.text.y = element_text(size=15)))

gp_survival.kml3d$plot <- gp_survival.kml3d$plot +
                                guides(fill=guide_legend(title=NULL,nrow = 1),
                                          color=guide_legend(title=NULL,nrow = 1),
                                          linetype=guide_legend(title=NULL,nrow = 1))+
                                scale_color_manual(values=c("green", "black"))+
                                scale_fill_manual(values=c("green", "black"))
gp_survival.kml3d$plot
```


kml3d
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
20
40
60
80
Time (months)
Cluster 1 (57.4%)
Cluster 2 (42.6%)