Case Study 2: Clustering the PBC Dataset

K-means clustering using the kml3d package

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

```
## Warning: package 'lme4' was built under R version 4.2.2
```

```
data(PBCseq)
# patients known to be alive  and without liver transplantation at 910 days of follow-up
idx <- unique(PBCseq[PBCseq$alive>910,]$id)
dnew910 <- PBCseq[PBCseq$id %in% idx,]
dnew910_uq <- dnew910[!duplicated(dnew910$id, fromLast=TRUE),] # Keep last observation per ID

dnew910$time <-  dnew910$month
dnew910$time <-  dnew910$month - mean(dnew910$month,na.rm=TRUE)
dnew910$time2 <-  dnew910$time^2

# use only data before 910 days (2.5 years)
dnew910.before <- dnew910[dnew910$day<=910,];
subj <- unique(dnew910.before$id)
N <- length(subj)
dnew910.before$lbili_scale <- as.numeric(scale(dnew910.before$lbili))
dnew910.before$lalbumin_scale <- as.numeric(scale(dnew910.before$lalbumin))
dnew910.before$lalk.phos_scale <- as.numeric(scale(dnew910.before$lalk.phos))
dnew910.before$lsgot_scale <- as.numeric(scale(dnew910.before$lsgot))
dnew910.before$lplatelet_scale <- as.numeric(scale(dnew910.before$lplatelet))

library(kml3d)
```

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

```
n.obs <- table(dnew910.before$id)       # number of observations
visit <- NULL
for (i in 1:N){visit <- c(visit,1:n.obs[i])}
dnew910.before$visit <- visit

# change to wide format
dnew910.before.wide <- reshape(dnew910.before[dnew910.before$visit!=5,
                c("id","lbili_scale", "lalbumin_scale", "lalk.phos_scale",
                                "lsgot_scale", "lplatelet_scale","visit")],
                idvar = "id", timevar = "visit", direction = "wide", sep="_")

# imputing the missing data
dnew910.before.wide.imp <- imputation(as.matrix(dnew910.before.wide)[,-1],
                                      method = "linearInterpol")
dnew910.before.wide.imp <- as.data.frame(dnew910.before.wide.imp)   # need to convert to data.frame
dnew910.before.wide.imp$id <- dnew910.before.wide$id
```

K-means clustering (kml3d package)

```
cldPreg <- cld3d(dnew910.before.wide.imp,
            idAll=dnew910.before.wide.imp$id,
             time = c(1,2,3,4),
                 varNames = c("lbili", "lalbumin",
                              "lalk.phos", "lsgot", "lplatelet"),
                 timeInData = list(lbili =      c(1,6,11,16),   # specify the columns of variables
                                   lalbumin=    c(2,7,12,17),
                                   lalk.phos=   c(3,8,13,18),
                          lsgot=      c(4,9,14,19),
                          lplatelet=  c(5,10,15,20)))
kml3d(cldPreg,nbClusters = 2:8 )                    # kml3d does not calculate bic for cluster=1
```

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

```
# extracting the bic values
bic <- c(cldPreg@c2[[1]]@criterionValues[6],
       cldPreg@c3[[1]]@criterionValues[6],
       cldPreg@c4[[1]]@criterionValues[6],
       cldPreg@c5[[1]]@criterionValues[6],
       cldPreg@c6[[1]]@criterionValues[6],
       cldPreg@c7[[1]]@criterionValues[6],
       cldPreg@c8[[1]]@criterionValues[6])
num.clust.kml3d <-  which.min(bic) + 1
```

```
# obtain the clusters
cluster.kml3d  <- getClusters(cldPreg, num.clust.kml3d)
cluster.kml3d <- as.numeric(cluster.kml3d)

# relabel clusters
cluster.re <- (cluster.kml3d==2)*1 +  (cluster.kml3d==1)*2
per <- paste(round(100*table(cluster.re)/N,1),"%",sep="")
cluster.kml3d <- factor(cluster.re, labels=paste("Cluster ",1:num.clust.kml3d," (",per,")",sep=""))
# Keep last observation per id
dnew_uq <- dnew910.before[!duplicated(dnew910.before$id, fromLast=TRUE),]
dat.cluster <- data.frame(dnew_uq$id,cluster.kml3d)
colnames(dat.cluster) <- c("id","cluster.kml3d")

dnew_uq <- merge(dnew_uq,dat.cluster,by="id")
dnew <- merge(dnew910.before,dat.cluster,by="id")
```

```
library(ggplot2)
```

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

```
library(cowplot)
p1.kml3d <- ggplot(data =dnew, aes(x =month, y = lbili,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,fill=cluster.kml3d))+
        geom_smooth(aes(x =month, y = lbili,
                        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("lbili") +
        ylim(c(min(dnew$lbili,na.rm=TRUE),
               max(dnew$lbili,na.rm=TRUE)))+
        scale_color_manual(values=c("green", "black"))+
        scale_fill_manual(values=c("green", "black"))

p2.kml3d <- ggplot(data =dnew, aes(x =month, y = lalbumin,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,fill=cluster.kml3d))+
        geom_smooth(aes(x =month, y = lalbumin,
                        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("lalbumin") +
        ylim(c(min(dnew$lalbumin,na.rm=TRUE),
               max(dnew$lalbumin,na.rm=TRUE)))+
        scale_color_manual(values=c("green", "black"))+
        scale_fill_manual(values=c("green", "black"))

p3.kml3d <- ggplot(data =dnew, aes(x =month, y =  lalk.phos,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,
                                   fill=cluster.kml3d))+
        geom_smooth(aes(x =month, y =  lalk.phos,
                        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(" lalk.phos") +
        ylim(c(min(dnew$lalk.phos,na.rm=TRUE),
               max(dnew$lalk.phos,na.rm=TRUE)))+
        scale_color_manual(values=c("green", "black"))+
        scale_fill_manual(values=c("green", "black"))


p4.kml3d <- ggplot(data =dnew, aes(x =month, y = lsgot,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,fill=cluster.kml3d))+
        geom_smooth(aes(x =month, y = lsgot,
                        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("lsgot") +
        ylim(c(min(dnew$lsgot,na.rm=TRUE),
               max(dnew$lsgot,na.rm=TRUE)))+
        scale_color_manual(values=c("green", "black"))+
        scale_fill_manual(values=c("green", "black"))
p5.kml3d <- ggplot(data =dnew, aes(x =month, y = lplatelet,
                                   color=cluster.kml3d,
                                   linetype=cluster.kml3d,
                                   fill=cluster.kml3d))+
        geom_smooth(aes(x =month, y = lplatelet, 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",
```

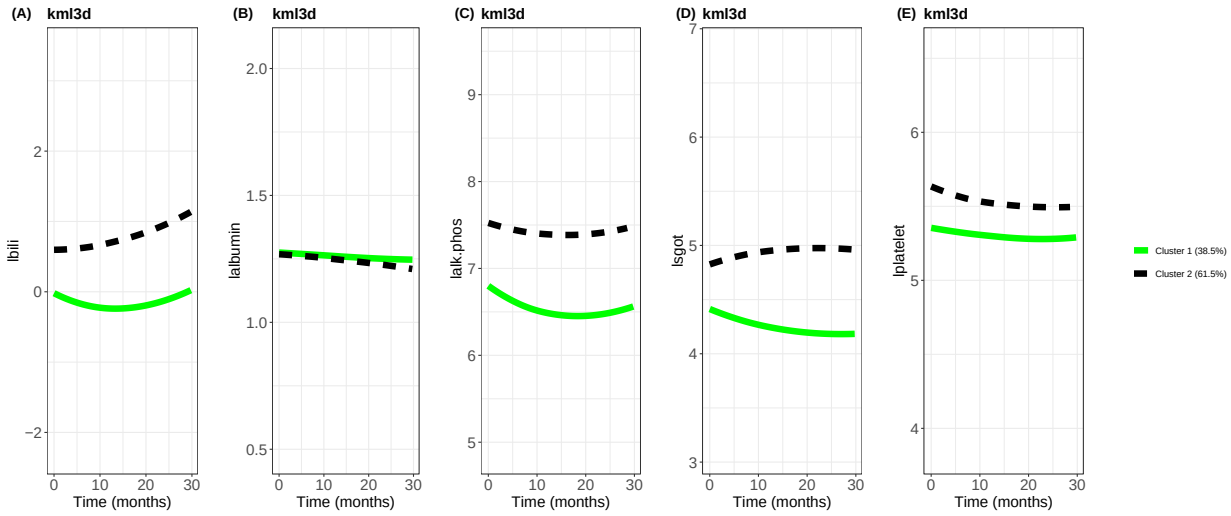
```
            plot.title = element_text(size = 15, face = "bold"),
            axis.text=element_text(size=15),
            axis.title=element_text(size=15),
            axis.text.x = element_text(angle = 0 ),
            strip.text.x = element_text(size = 15, angle = 0),
            strip.text.y = 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("lplatelet") +
        ylim(c(min(dnew$lplatelet,na.rm=TRUE),
               max(dnew$lplatelet,na.rm=TRUE)))+
        scale_color_manual(values=c("green", "black"))+
        scale_fill_manual(values=c("green", "black"))

#-------------------------------------------------------------------------------------------------------------
# extract a legend that is laid out horizontally
legend.kml3d <- get_legend(ggplot(data =dnew, aes(x =month, y = lplatelet,
                                                  color=cluster.kml3d,
                                                  linetype=cluster.kml3d,
                                                  fill=cluster.kml3d))+
                           geom_smooth(aes(x =month, y = lplatelet,
                                           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),
                                 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,ncol = 1,byrow=TRUE),
                                  color=guide_legend(title=NULL,ncol = 1,byrow=TRUE),
                                  linetype=guide_legend(title=NULL,ncol = 1,byrow=TRUE)) +
                           xlab("Time (months)") + ylab("lplatelet") +
                           ylim(c(min(dnew$lplatelet,na.rm=TRUE),
                                  max(dnew$lplatelet,na.rm=TRUE)))+
                           scale_color_manual(values=c("green", "black"))+
                           scale_fill_manual(values=c("green", "black"))

)
```

```
## Warning: Removed 15 rows containing non-finite values (`stat_smooth()`).
```

```
plot_grid(p1.kml3d,NULL,p2.kml3d,NULL,p3.kml3d,
          NULL,p4.kml3d,NULL,p5.kml3d,NULL,legend.kml3d,
          labels=c("(A)","", "(B)","","(C)","","(D)","","(E)","",""), nrow = 1,
          rel_widths = c(1,0.1,1,0.1,1,0.1,1,0.1,1,0.1,0.7))
```

```
## Warning: Removed 5 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 15 rows containing non-finite values (`stat_smooth()`).
```




```
library(survminer)
```

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

```
library(survival)
# use only data after 910 days (2.5 years)
dnew910.after <- dnew910[dnew910$day > 910,]; length(unique(dnew910.after$id))
```

```
## [1] 193
```

```
dnew910_uq <- merge(dnew910.after[!duplicated(dnew910.after$id, fromLast=TRUE),],
                    dnew_uq[,c("id","cluster.kml3d")], by="id")
fit <- survfit(Surv(month, delta.death) ~  cluster.kml3d,data = dnew910_uq, start.time=30.08)
res.cox <- coxph(Surv(month, delta.death) ~ cluster.kml3d, data =  dnew910_uq )
pvalue <-  ifelse(summary(res.cox)$sctest[3] >= 0.0001,
                  summary(res.cox)$sctest[3],'<0.0001')

names(fit$strata) <-  paste("Cluster ",1:num.clust.kml3d," (",per,")",sep="")
gp_survival.kml3d <-   ggsurvplot(fit, data = dnew910_uq, title="kml3d",
                          risk.table = FALSE,
                 risk.table.y.text.col = FALSE,
                 pval = TRUE,
                 pval.coord = c(40, 0.03),
                          legend = "bottom", # conf.int = TRUE,
                          xlab = "Time (months)",
                 legend.title="Clusters",
                          ggtheme = theme_bw() +
                   theme(legend.position ="none",legend.title=element_blank(),
                                        plot.title = element_text(size = 15, face = "bold"),
                                        axis.text=element_text(size=15),
                                        axis.title=element_text(size=15),
                                        strip.text.x = element_text(size=15),
                                        strip.text.y = element_text(size=15)))
gp_survival.kml3d$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
```