Case Study 2: Clustering the PBC Dataset

Group-based trajectory modeling using the flexmix 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,]

# remove missing values
dnew910.before <- dnew910.before[is.na(dnew910.before$lbili)==FALSE,]
dnew910.before <- dnew910.before[is.na(dnew910.before$lalbumin)==FALSE,]
dnew910.before <- dnew910.before[is.na(dnew910.before$lalk.phos)==FALSE,]
dnew910.before <- dnew910.before[is.na(dnew910.before$lsgot)==FALSE,]
dnew910.before <- dnew910.before[is.na(dnew910.before$lplatelet)==FALSE,]
# standardize the variables
dnew910.before$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))
```

group-based trajectory modeling (flexmix package)

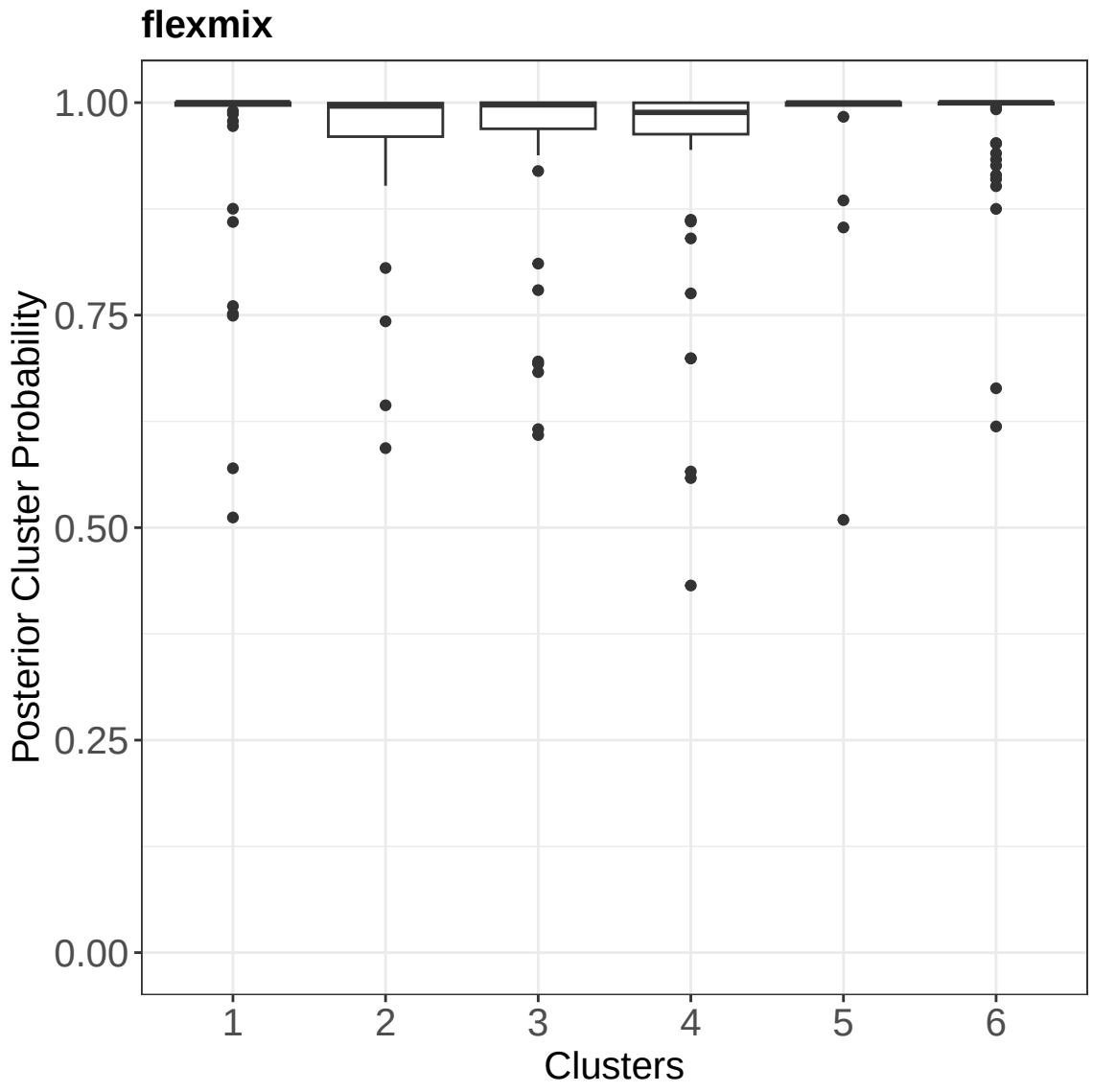
```
# install.packages("flexmix")
library(flexmix)
# not run to reduce compiling time
#set.seed(12421)
#bic <- NULL
#for (kk in 1:8){
#   fit.flexmix <- flexmix(  ~ time|id, data = dnew910.before, k = kk,
#                            model = list(FLXMRglm(lbili_scale ~ time),
#                        FLXMRglm(lalbumin_scale ~ time),
#                                     FLXMRglm(lalk.phos_scale ~ time),
#                                     FLXMRglm(lsgot_scale ~ time),
#                                     FLXMRglm(lplatelet_scale ~ time)))
```

```
#   bic <- c(bic,summary(fit.flexmix)@BIC)
#}
# print the number of clusters with the smallest BIC
#num.clust.flexmix <- which.min(bic); num.clust.flexmix
```

```
set.seed(521)
num.clust.flexmix <- 6  # optimal number of clusters based on bic
fit_flexmix <- flexmix(  ~ time|id, data = dnew910.before, k = num.clust.flexmix,
                         model = list(FLXMRglm(lbili_scale ~ time),
                                      FLXMRglm(lalbumin_scale ~ time),
                                          FLXMRglm(lalk.phos_scale ~ time),
                                          FLXMRglm(lsgot_scale ~ time),
                                          FLXMRglm(lplatelet_scale ~ time)))
# compute and plot the posterior cluster probability
postprob <- apply(posterior(fit_flexmix),1,max)
cluster.tmp <-  apply(posterior(fit_flexmix),1,which.max);
cluster.flexmix <-  (cluster.tmp == 6)*1 +
             (cluster.tmp == 1)*2 +
             (cluster.tmp == 4)*3 +
             (cluster.tmp == 3)*4 +
             (cluster.tmp == 2)*5 +
             (cluster.tmp == 5)*6
df.new <- data.frame(id=dnew910.before$id,
              cluster.flexmix=cluster.flexmix,
              postprob=postprob)
df.new.uq <- df.new[!duplicated(df.new$id, fromLast=TRUE),]
cluster.flexmix <- df.new.uq$cluster.flexmix
library(ggplot2)
```

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

```
# Posterior cluster probability
bp.flexmix <- ggplot(df.new.uq, aes(x=factor(cluster.flexmix), y=postprob)) +
            geom_boxplot() + ggtitle("flexmix") +
            xlab("Clusters") + ylab("Posterior Cluster Probability") +
        ylim(c(0,1)) +
        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"))
bp.flexmix
```




```
per <- paste(round(100*table(cluster.flexmix)/length(cluster.flexmix)),"%",sep="")
df.new.uq$cluster.flexmix <-   factor(cluster.flexmix ,
                    labels=paste("Cluster ",1:num.clust.flexmix," (",per,")",sep=""))
dat.cluster <- data.frame(df.new.uq$id,df.new.uq$cluster.flexmix)
colnames(dat.cluster) <- c("id","cluster.flexmix")
dnew <- merge(dnew910.before,dat.cluster,by="id")
```

```
library(ggplot2)
library(cowplot)
p1.flexmix <- ggplot(data =dnew, aes(x = month, y = lbili,
                                     color=cluster.flexmix,
                                     linetype=cluster.flexmix,fill=cluster.flexmix))+
  ggtitle("flexmix") +
        geom_smooth(aes(x =month, y = lbili,
                        color=cluster.flexmix,linetype=cluster.flexmix,fill=cluster.flexmix),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        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 = 3,byrow=TRUE),
               color=guide_legend(title=NULL,nrow = 3,byrow=TRUE),
                linetype=guide_legend(title=NULL,nrow = 3,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","blue","red","purple","goldenrod3"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple","goldenrod3"))

p2.flexmix <- ggplot(data =dnew, aes(x = month, y = lalbumin,
                                     color=cluster.flexmix,
                                     linetype=cluster.flexmix,fill=cluster.flexmix))+
  ggtitle("flexmix") +
        geom_smooth(aes(x =month, y = lalbumin,
                        color=cluster.flexmix,linetype=cluster.flexmix,fill=cluster.flexmix),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        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 = 3,byrow=TRUE),
               color=guide_legend(title=NULL,nrow = 3,byrow=TRUE),
                linetype=guide_legend(title=NULL,nrow = 3,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","blue","red","purple","goldenrod3"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple","goldenrod3"))

p3.flexmix <- ggplot(data =dnew, aes(x = month, y = lalk.phos,
                                     color=cluster.flexmix,
                                     linetype=cluster.flexmix,fill=cluster.flexmix))+
  ggtitle("flexmix") +
        geom_smooth(aes(x =month, y = lalk.phos,
                        color=cluster.flexmix,linetype=cluster.flexmix,fill=cluster.flexmix),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        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 ),
```

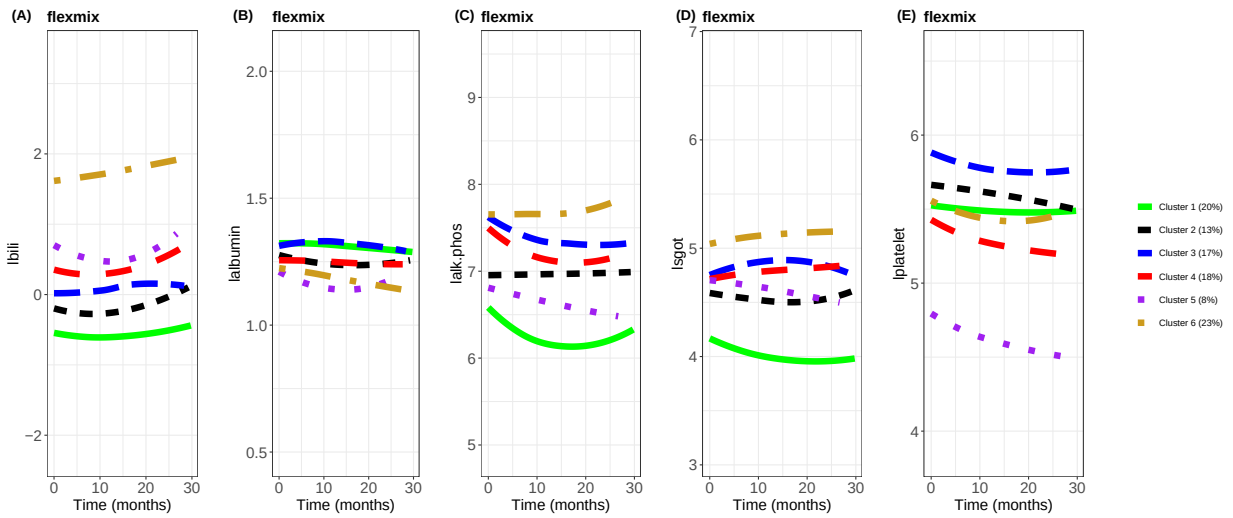
```
            strip.text.x = element_text(size = 15, angle = 0),
            strip.text.y = element_text(size = 15,face="bold")) +
        guides(fill=guide_legend(title=NULL,nrow = 3,byrow=TRUE),
               color=guide_legend(title=NULL,nrow = 3,byrow=TRUE),
                linetype=guide_legend(title=NULL,nrow = 3,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","blue","red","purple","goldenrod3"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple","goldenrod3"))
p4.flexmix <- ggplot(data =dnew, aes(x = month, y = lsgot,
                                     color=cluster.flexmix,
                                     linetype=cluster.flexmix,fill=cluster.flexmix))+
  ggtitle("flexmix") +
        geom_smooth(aes(x =month, y = lsgot,
                        color=cluster.flexmix,linetype=cluster.flexmix,fill=cluster.flexmix),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        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 = 3,byrow=TRUE),
               color=guide_legend(title=NULL,nrow = 3,byrow=TRUE),
                linetype=guide_legend(title=NULL,nrow = 3,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","blue","red","purple","goldenrod3"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple","goldenrod3"))

p5.flexmix <- ggplot(data =dnew, aes(x = month, y = lplatelet,
                                     color=cluster.flexmix,
                                     linetype=cluster.flexmix,fill=cluster.flexmix))+
  ggtitle("flexmix") +
        geom_smooth(aes(x =month, y = lplatelet,
                        color=cluster.flexmix,linetype=cluster.flexmix,fill=cluster.flexmix),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        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 = 3,byrow=TRUE),
               color=guide_legend(title=NULL,nrow = 3,byrow=TRUE),
                linetype=guide_legend(title=NULL,nrow = 3,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","blue","red","purple","goldenrod3"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple","goldenrod3"))

#-------------------------------------------------------------------------------------------------------
# extract a legend that is laid out horizontally
legend.flexmix <- get_legend(ggplot(data =dnew, aes(x = month, y = lplatelet,
                                                    color=cluster.flexmix,
                                                    linetype=cluster.flexmix,fill=cluster.flexmix))+
                             ggtitle("flexmix") +
                             geom_smooth(aes(x =month, y = lplatelet,
                                             color=cluster.flexmix,
                                             linetype=cluster.flexmix,fill=cluster.flexmix),
                                         method = "loess", linewidth = 3,se = FALSE,span=2)+
                             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","blue","red","purple","goldenrod3"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple","goldenrod3"))
)

plot_grid(p1.flexmix,NULL,p2.flexmix,NULL,p3.flexmix,NULL,
          p4.flexmix,NULL,p5.flexmix,NULL,legend.flexmix,
          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))
```




```
#-------------------------------------------------------------------------------------------------------------
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$id, fromLast=TRUE),],df.new, by="id")
fit <- survfit(Surv(month, delta.death) ~  cluster.flexmix,data = dnew910_uq, start.time=30.08)
# weighted cox model
res.cox <- coxph(Surv(month, delta.death) ~ cluster.flexmix,
                 weights=postprob, data =  dnew910_uq )
pvalue <-  ifelse(summary(res.cox)$sctest[3] >= 0.0001,summary(res.cox)$sctest[3],'<0.0001')

names(fit$strata) <- paste("Cluster ",1:num.clust.flexmix," (",per,")",sep="")
gp_survival.flexmix <-   ggsurvplot(fit, data = dnew910_uq, title="flexmix",
                         risk.table = FALSE,
                 risk.table.y.text.col = FALSE,
                 pval = pvalue,
                 pval.coord = c(40, 0.03),
                         legend = "bottom", # conf.int = TRUE,
                         xlab = "Time (months)",
                 legend.title="Clusters",
                         ggtheme = theme_bw() +
                                          theme(legend.position ="none",legend.title=element_blank(),
                                       plot.title = element_text(size = 15, face = "bold"),
                                          axis.text=element_text(size=15),
                                          axis.title=element_text(size=15),
                                          strip.text.x = element_text(size=15),
                                          strip.text.y = element_text(size=15)))
gp_survival.flexmix$plot <- gp_survival.flexmix$plot +
        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))+
        scale_color_manual(values=c("green", "black","blue","red","purple","goldenrod3"))+
```

```
        scale_fill_manual(values=c("green", "black","blue","red","purple","goldenrod3"))
gp_survival.flexmix
```