Case Study 2: Clustering the PBC Dataset

Latent class mixed effect model using the lcmm 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,]

# 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))
```

latent class mixed effect model (lcmm package)

```
library(lcmm)
```

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

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

```
mult1a <- multlcmm(lbili_scale + lalbumin_scale + lalk.phos_scale + lsgot_scale + lplatelet_scale ~ time
            random =~ time ,
            subject='id',
            data = dnew910.before,
            verbose = FALSE,
            randomY = TRUE,
            ng = 1)

# not run to reduce compiling time
#BIC <- NULL
#for (kk in 2:8){
#fit.multlcmm <- multlcmm(lbili_scale + lalbumin_scale + lalk.phos_scale + lsgot_scale + lplatelet_scal
#           mixture = ~ time,
```

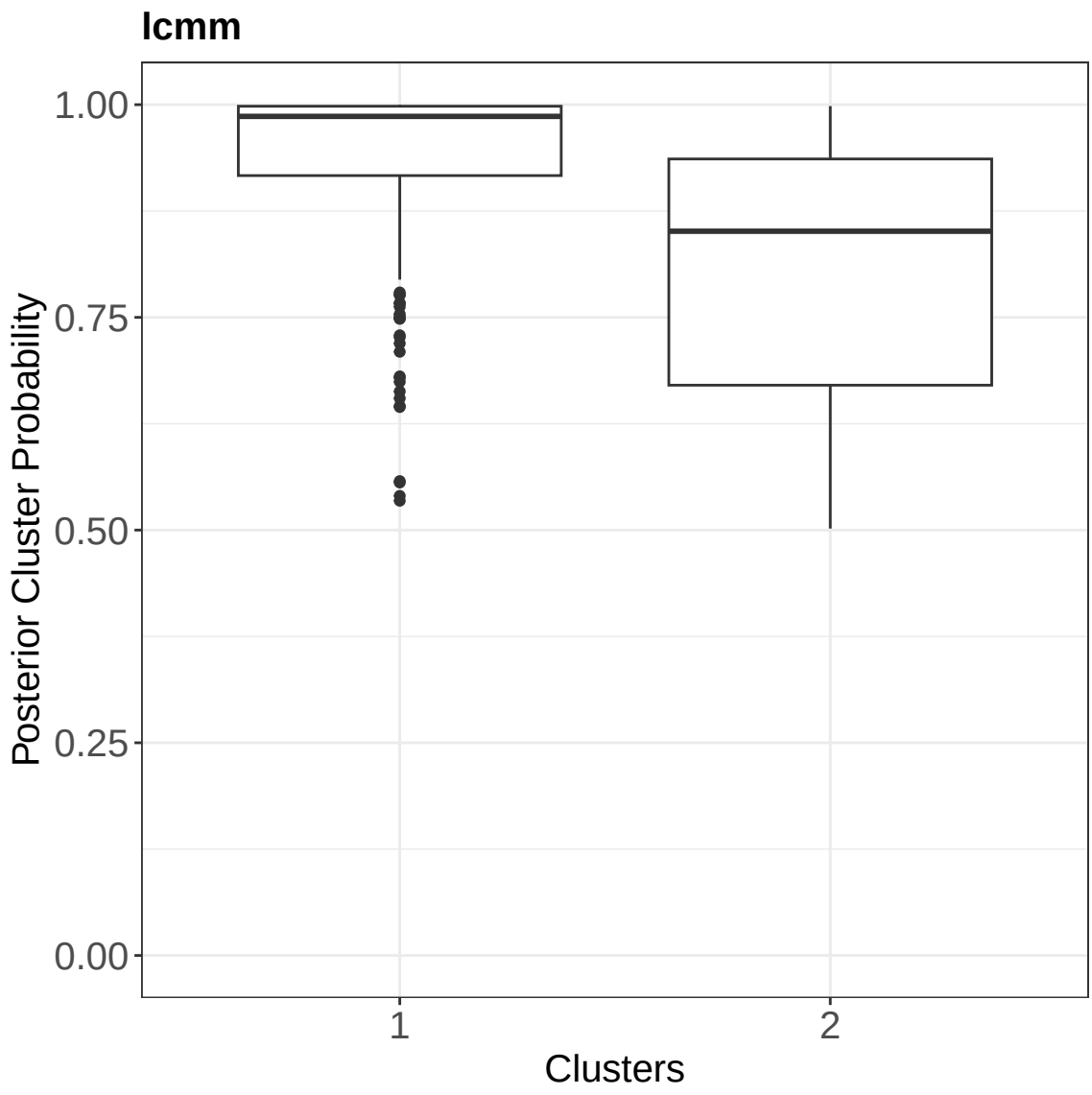
```
#            random =~ time ,
#            subject='id',
#            verbose = FALSE,
#            nwg = TRUE,
#            data = dnew910.before,
#            randomY = TRUE,ng = kk, B =mult1a  )
#BIC <- c(BIC,fit.multlcmm$BIC)
# }
# print the number of clusters with the smallest BIC
#num.clust.multlcmm <- which.min(BIC) + 1; num.clust.multlcmm
```

```
num.clust.multlcmm <- 2  # optimal number of clusters based on bic
fit.multlcmm <- multlcmm(lbili_scale + lalbumin_scale + lalk.phos_scale + lsgot_scale + lplatelet_scale
            mixture = ~ time,
            random =~ time ,
            subject='id',
            verbose = FALSE,
            nwg = TRUE,
            data = dnew910.before,
            randomY = TRUE,
            ng = num.clust.multlcmm, B =mult1a  )

# Posterior Cluster Probability of Assignment
postprob <- apply(fit.multlcmm$pprob[,-c(1,2)],1,max)
# relabel cluster
cluster.re <- (fit.multlcmm$pprob$class==1)*1 + (fit.multlcmm$pprob$class==2)*2
dnew_uq <- dnew910.before[!duplicated(dnew910.before$id, fromLast=TRUE),] # Keep last observation per i
dnew_uq$postprob <- postprob
dnew_uq$cluster.lcmm <-  cluster.re
# Posterior cluster probability
library(ggplot2)
```

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

```
bp.lcmm <- ggplot(dnew_uq, aes(x=factor(cluster.lcmm), y=postprob)) +
            geom_boxplot() + ggtitle("lcmm") +
            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.lcmm
```




```
N <- length(unique(dnew910.before$id))
per <- paste(round(100*table(cluster.re)/N,1),"%",sep="")
cluster.multlcmm <- factor(cluster.re,
                           label=paste("Cluster ",1:num.clust.multlcmm," (",per,")",sep=""))

dat.cluster <- data.frame(fit.multlcmm$pprob$id,cluster.multlcmm)
colnames(dat.cluster) <- c("id","cluster.multlcmm")
 # Keep last observation per id
dnew_uq <- dnew910.before[!duplicated(dnew910.before$id, fromLast=TRUE),]
dnew_uq <- merge(dnew_uq,dat.cluster,by="id")
dnew_uq$postprob <- postprob
dnew <- merge(dnew910.before,dat.cluster,by="id")
```

```
library(ggplot2)
library(cowplot)
p1.lcmm <- ggplot(data =dnew, aes(x = month, y = lbili,
                                  color=cluster.multlcmm,
```

```
                                linetype=cluster.multlcmm,fill=cluster.multlcmm))+
  ggtitle("lcmm") +
        geom_smooth(aes(x =month, y = lbili,
                        color=cluster.multlcmm,
                        linetype=cluster.multlcmm,fill=cluster.multlcmm),
                    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 = 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("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.lcmm <- ggplot(data =dnew, aes(x = month, y = lalbumin,
                                color=cluster.multlcmm,
                                linetype=cluster.multlcmm,fill=cluster.multlcmm))+
  ggtitle("lcmm") +
        geom_smooth(aes(x =month, y = lalbumin,
                        color=cluster.multlcmm,
                        linetype=cluster.multlcmm,fill=cluster.multlcmm),
                    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 = 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("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.lcmm <- ggplot(data =dnew, aes(x = month, y = lalk.phos,
                                color=cluster.multlcmm,
                                linetype=cluster.multlcmm,fill=cluster.multlcmm))+
  ggtitle("lcmm") +
        geom_smooth(aes(x = month, y = lalk.phos,
                        color=cluster.multlcmm,
                        linetype=cluster.multlcmm,fill=cluster.multlcmm),
```

```
                    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 = 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("lalbumin")  +
        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.lcmm <- ggplot(data =dnew, aes(x = month, y = lsgot,
                                  color=cluster.multlcmm,
                                  linetype=cluster.multlcmm,fill=cluster.multlcmm))+
  ggtitle("lcmm") +
        geom_smooth(aes(x = month, y = lsgot,
                        color=cluster.multlcmm,
                        linetype=cluster.multlcmm,fill=cluster.multlcmm),
                    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 = 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("lalbumin")  +
        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.lcmm <- ggplot(data =dnew, aes(x = month, y = lplatelet, color=cluster.multlcmm,
                                  linetype=cluster.multlcmm,
                                  fill=cluster.multlcmm))+
  ggtitle("lcmm") +
        geom_smooth(aes(x = month, y = lplatelet,
                        color=cluster.multlcmm,
                        linetype=cluster.multlcmm,fill=cluster.multlcmm),
                    method = "loess", linewidth= 3,se = FALSE,span=2)+
        theme_bw() +
        theme(legend.position = "none",
```

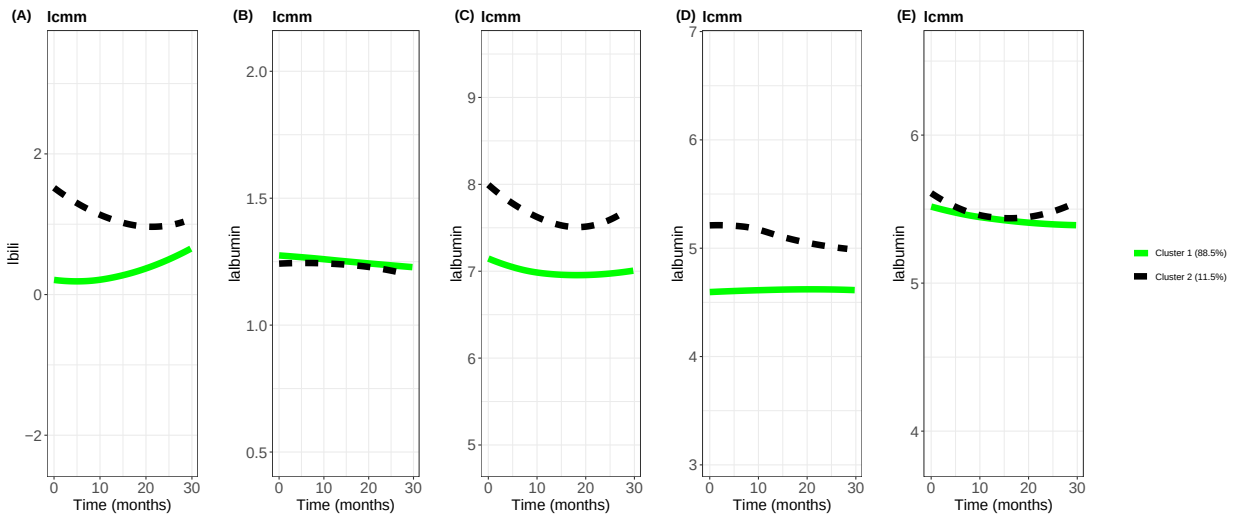
```
            plot.title = element_text(size = 15, face = "bold"),
            axis.text=element_text(size=15),
            axis.title=element_text(size=15),
            axis.text.x = element_text(angle = 0 ),
            strip.text.x = element_text(size = 15, angle = 0),
            strip.text.y = element_text(size = 15,face="bold")) +
        guides(fill=guide_legend(title=NULL,nrow = 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("lalbumin")  +
        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.lcmm <- get_legend(ggplot(data =dnew, aes(x = month, y = lplatelet,
                                                 color=cluster.multlcmm,
                                                 linetype=cluster.multlcmm,fill=cluster.multlcmm))+
                            ggtitle("lcmm") +
                            geom_smooth(aes(x = month, y = lplatelet,
                                            color=cluster.multlcmm,
                                            linetype=cluster.multlcmm,fill=cluster.multlcmm),
                                        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,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("lalbumin")  +
                            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.lcmm,NULL,p2.lcmm,NULL,p3.lcmm,NULL,
          p4.lcmm,NULL,p5.lcmm,NULL,legend.lcmm,
          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()`).
## 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$id, fromLast=TRUE),],
                dnew_uq[,c("id","cluster.multlcmm","postprob")], by="id")
fit <- survfit(Surv(month, delta.death) ~  cluster.multlcmm,data = dnew910_uq, start.time=30.08)
# weighted cox model
res.cox <- coxph(Surv(month, delta.death) ~ cluster.multlcmm, weights=postprob, data =  dnew910_uq )
pvalue <-  ifelse(summary(res.cox)$sctest[3] >= 0.0001,summary(res.cox)$sctest[3],'<0.0001')
#pvalue <-  0.17
names(fit$strata) <-  paste("Cluster ",1:num.clust.multlcmm," (",per,")",sep="")
gp_survival.lcmm <-   ggsurvplot(fit, data = dnew910_uq, title="lcmm",
                         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"),
                                       #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.lcmm$plot <- gp_survival.lcmm$plot +
        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))+
```

```
        scale_color_manual(values=c("green", "black"))+
        scale_fill_manual(values=c("green", "black"))
gp_survival.lcmm
```