Case Study 2: Clustering the PBC Dataset

Bayesian mixture model using the mixAK 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))
```

Bayesian mixture model (mixAK package)

```
# not run to reduce compiling time
# determining the number of clusters
set.seed(22)
#PED <- NULL
#for (kk in 1:8){
#modK <- GLMM_MCMC(y = dnew910.before[,c("lbili_scale", "lalbumin_scale",
#                                        "lalk.phos_scale", "lsgot_scale", "lplatelet_scale")],
#           dist = c("gaussian","gaussian","gaussian","gaussian","gaussian"),
#           id = dnew910.before[, "id"],
#    z = list(lbili_scale = dnew910.before[, c("time")],
#           lalbumin_scale = dnew910.before[, c("time")],
#           lalk.phos_scale = dnew910.before[, c("time")],
#           lsgot_scale = dnew910.before[, c("time")],
#           lplatelet_scale = dnew910.before[, c("time")]),
#   random.intercept = c(TRUE,TRUE,TRUE,TRUE,TRUE),
#   prior.b = list(Kmax = kk), nMCMC = c(burn = 1000,
#                                        keep = 1000, thin = 1, info = 1000), parallel = TRUE)
#   PED <- c(PED,modK$PED[3])
#}
# print the best number of clusters with the smallest PED
# num.clust.mixAK <- which.min(PED); num.clust.mixAK
```

```
num.clust.mixAK <-   2  # optimal number of clusters based on PED
# note that even seed is used, each time running the model, the
# clustering results (e.g., cluster proportions and membership) are
# slightly different
set.seed(2022)
fit_mixAK <- GLMM_MCMC(y = dnew910.before[,c("lbili_scale", "lalbumin_scale",
                                             "lalk.phos_scale", "lsgot_scale", "lplatelet_scale")],
            dist = c("gaussian","gaussian","gaussian","gaussian","gaussian"),
            id = dnew910.before[, "id"],
     z = list(lbili_scale = dnew910.before[, c("time")],
            lalbumin_scale = dnew910.before[, c("time")],
            lalk.phos_scale = dnew910.before[, c("time")],
            lsgot_scale = dnew910.before[, c("time")],
            lplatelet_scale = dnew910.before[, c("time")]),
    random.intercept = c(TRUE,TRUE,TRUE,TRUE,TRUE),
    prior.b = list(Kmax = num.clust.mixAK),
    nMCMC = c(burn = 1000, keep = 1000, thin = 1, info = 1000), parallel = TRUE)
```

```
## Parallel MCMC sampling of two chains started on Tue Jun  6 21:58:51 2023.
## Parallel MCMC sampling finished on Tue Jun  6 21:58:59 2023.
##
## Computation of penalized expected deviance started on Tue Jun  6 21:58:59 2023.
## Computation of penalized expected deviance finished on Tue Jun  6 21:59:24 2023.
```

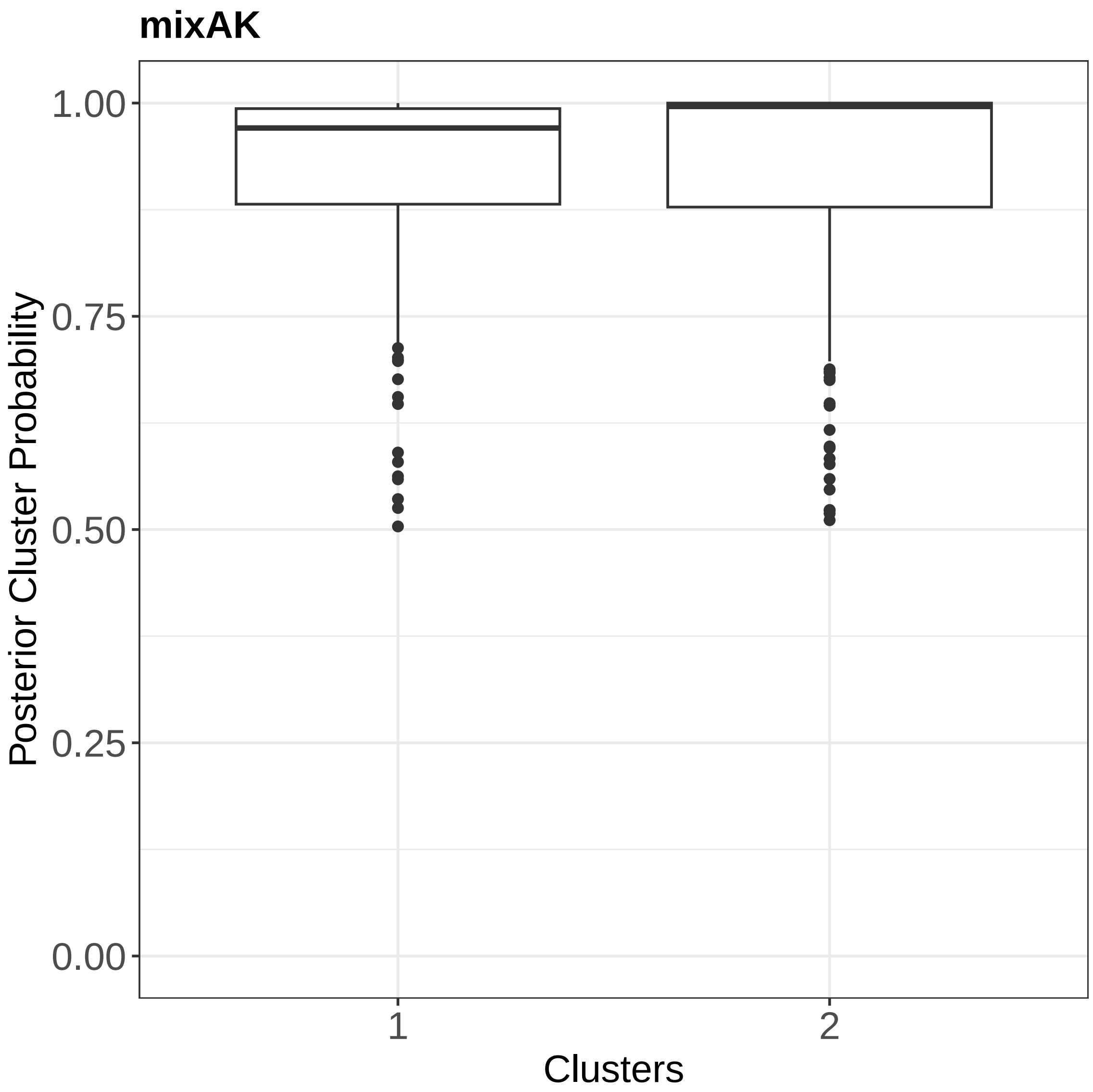
```
fit_mixAK <- NMixRelabel(fit_mixAK,type = "stephens",keep.comp.prob=TRUE)
```

```
##
## Re-labelling chain number 1
## ===========================
## MCMC Iteration (simple re-labelling) 1000
## Stephens' re-labelling iteration (number of labelling changes): 1 (0)
##
## Re-labelling chain number 2
## ===========================
## MCMC Iteration (simple re-labelling) 1000
## Stephens' re-labelling iteration (number of labelling changes): 1 (0)
```

```
cluster.mixAK <- apply(fit_mixAK[[1]]$poster.comp.prob,1,which.max);

# Keep last observation per id
dnew_uq <- dnew910.before[!duplicated(dnew910.before$id, fromLast=TRUE),]
dnew_uq$postprob <- apply(fit_mixAK[[1]]$poster.comp.prob,1,max);
dnew_uq$cluster.mixAK <- cluster.mixAK
library(ggplot2)
```

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

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



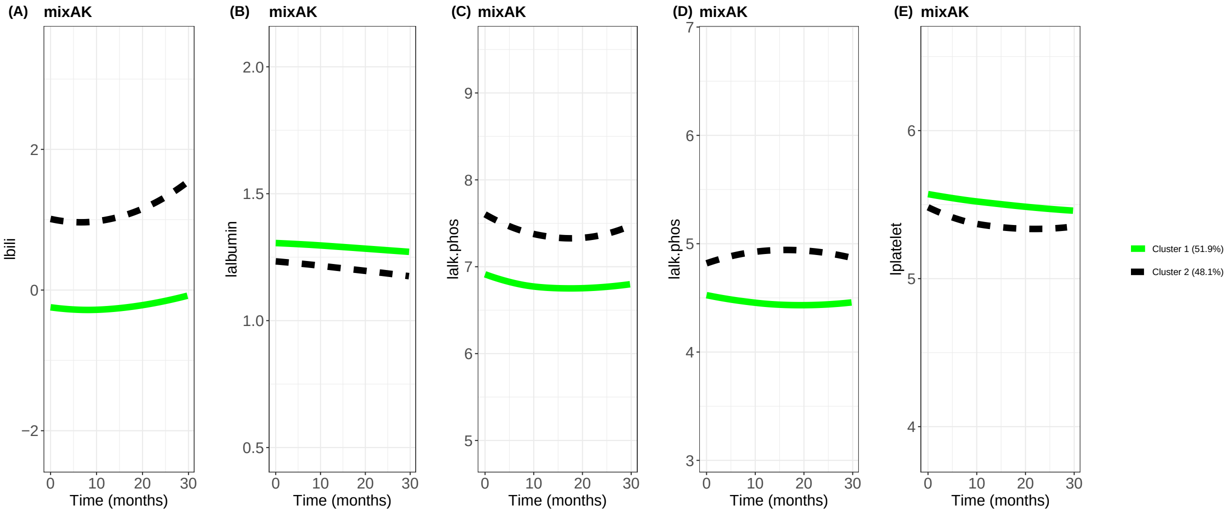
```
N <- length(unique(dnew910.before$id))
per <- paste(round(100*table(cluster.mixAK)/N,1),"%",sep="")
dnew_uq$cluster.mixAK <- factor(cluster.mixAK,
                                labels=paste("Cluster ",1:num.clust.mixAK," (",per,")",sep=""))
dat.cluster <- data.frame(dnew_uq$id,dnew_uq$cluster.mixAK)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

```
library(ggplot2)
library(cowplot)
p1.mixAK <- ggplot(data =dnew, aes(x = month, y = lbili,
                                   color=cluster.mixAK,linetype=cluster.mixAK,fill=cluster.mixAK))+
  ggtitle("mixAK") +
        geom_smooth(aes(x =month, y = lbili,
                        color=cluster.mixAK,linetype=cluster.mixAK,fill=cluster.mixAK),
                    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 = 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.mixAK <- ggplot(data =dnew, aes(x = month, y = lalbumin,
                                   color=cluster.mixAK,
                                   linetype=cluster.mixAK,fill=cluster.mixAK))+
  ggtitle("mixAK") +
        geom_smooth(aes(x = month, y = lalbumin,
                        color=cluster.mixAK,linetype=cluster.mixAK,fill=cluster.mixAK),
                    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 = 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.mixAK <- ggplot(data =dnew, aes(x =month, y = lalk.phos,
                                   color=cluster.mixAK,
                                   linetype=cluster.mixAK,fill=cluster.mixAK))+
  ggtitle("mixAK") +
        geom_smooth(aes(x =month, y = lalk.phos,
                        color=cluster.mixAK,linetype=cluster.mixAK,fill=cluster.mixAK),
                    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 = 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.mixAK <- ggplot(data =dnew, aes(x =month, y = lsgot,
                                   color=cluster.mixAK,linetype=cluster.mixAK,fill=cluster.mixAK))+
  ggtitle("mixAK") +
        geom_smooth(aes(x =month, y = lsgot,
                        color=cluster.mixAK,linetype=cluster.mixAK,fill=cluster.mixAK),
                    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 = 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$lsgot,na.rm=TRUE),max(dnew$lsgot,na.rm=TRUE)))+
        scale_color_manual(values=c("green", "black"))+
        scale_fill_manual(values=c("green", "black"))

p5.mixAK <- ggplot(data =dnew, aes(x =month, y = lplatelet,
                                   color=cluster.mixAK,
                                   linetype=cluster.mixAK,fill=cluster.mixAK))+
  ggtitle("mixAK") +
        geom_smooth(aes(x =month, y = lplatelet,
                        color=cluster.mixAK,
                        linetype=cluster.mixAK,fill=cluster.mixAK),
                    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 = 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.mixAK <- get_legend( ggplot(data =dnew, aes(x =month, y = lplatelet,
                                                   color=cluster.mixAK,
                                                   linetype=cluster.mixAK,fill=cluster.mixAK))+
                             ggtitle("mixAK") +
                             geom_smooth(aes(x =month, y = lplatelet,
                                             color=cluster.mixAK,
                                             linetype=cluster.mixAK,fill=cluster.mixAK),
                                         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"))+
                             scale_fill_manual(values=c("green", "black"))
)
```

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

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

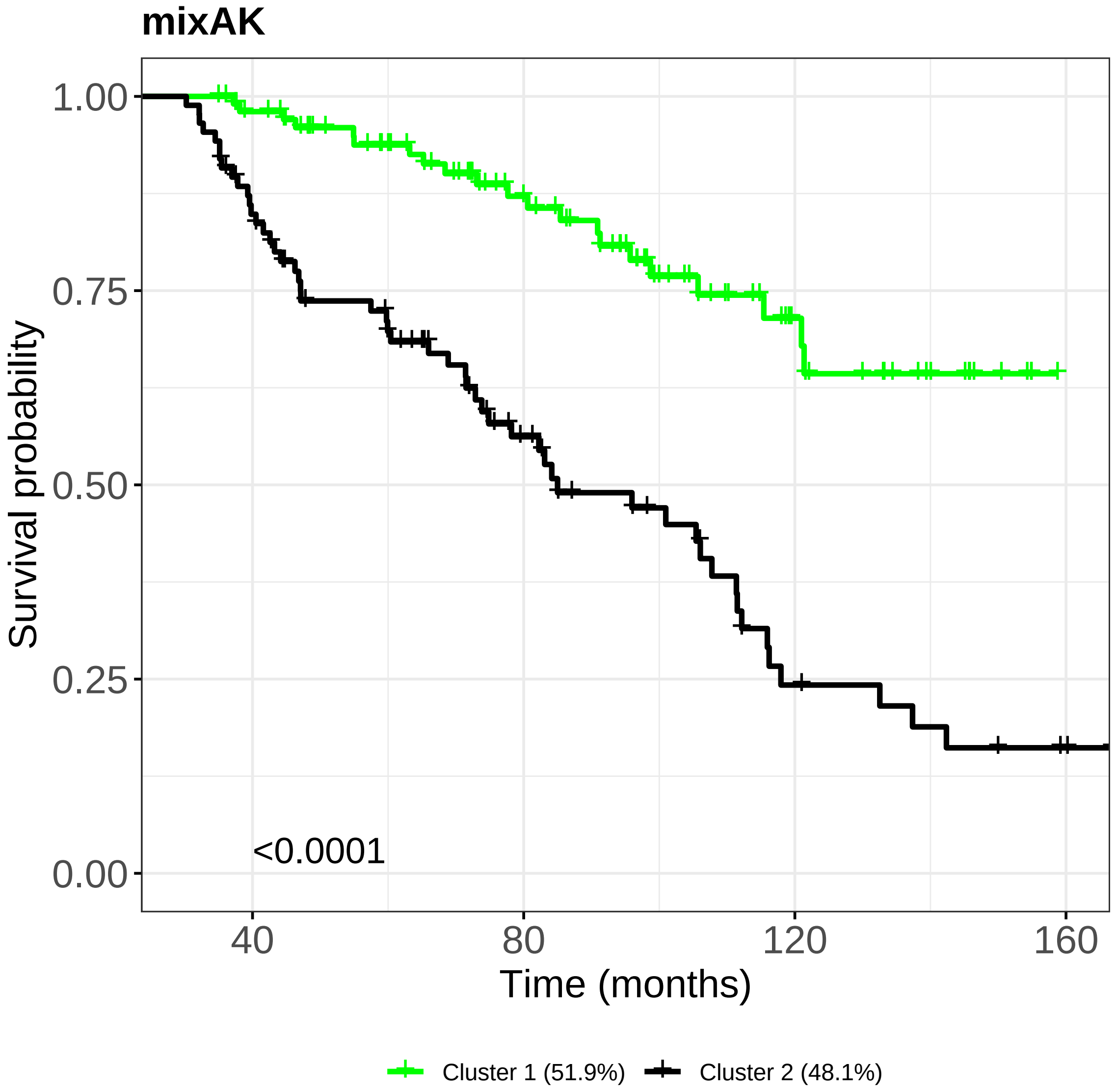
```
## Warning: Removed 5 rows containing non-finite values (`stat_smooth()`).
## Removed 15 rows containing non-finite values (`stat_smooth()`).
```



```
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$day > 910,];
dnew910_uq <- merge(dnew910.after[!duplicated(dnew910.after$id, fromLast=TRUE),],
            dnew_uq[,c("id","cluster.mixAK","postprob")], by="id")
fit <- survfit(Surv(month, delta.death) ~  cluster.mixAK,
               data = dnew910_uq, start.time=30.08)
res.cox <- coxph(Surv(month, delta.death) ~ cluster.mixAK,
                 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.mixAK," (",per,")",sep="")
gp_survival.mixAK <-   ggsurvplot(fit, data = dnew910_uq, title="mixAK",
                          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.mixAK$plot <- gp_survival.mixAK$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.mixAK
```


mixAK
Survival probability
<0.0001
Time (months)
Cluster 1 (51.9%)
Cluster 2 (48.1%)