Case Study 2: Clustering the PBC Dataset

Group-based trajectory modeling using the gbmt 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))
```

group-based trajectory modeling (gbmt package)

```
# install.packages("gbmt")
library(gbmt)
N <- length(unique(dnew910.before$id))
varNames <- c("lbili_scale", "lalbumin_scale",
              "lalk.phos_scale", "lsgot_scale", "lplatelet_scale")
# not run to reduce compiling time
#bic <- NULL
#for (kk in 1:8){
#   fit.gbmt <- gbmt(x.names=varNames, unit="id",
# time="time", d=1, ng=kk, data=dnew910.before, scaling=0)
#   bic <- c(bic,fit.gbmt$ic[2])
#   }
# print the best number of clusters with the smallest BIC
#num.clust.gbmt <- which.min(bic);num.clust.gbmt
```

```
num.clust.gbmt <- 5  # optimal number of clusters based on bic
fit_gbmt <- gbmt(x.names=varNames, unit="id", time="time", d=1,
                    ng=num.clust.gbmt, data=dnew910.before, scaling=0)
```

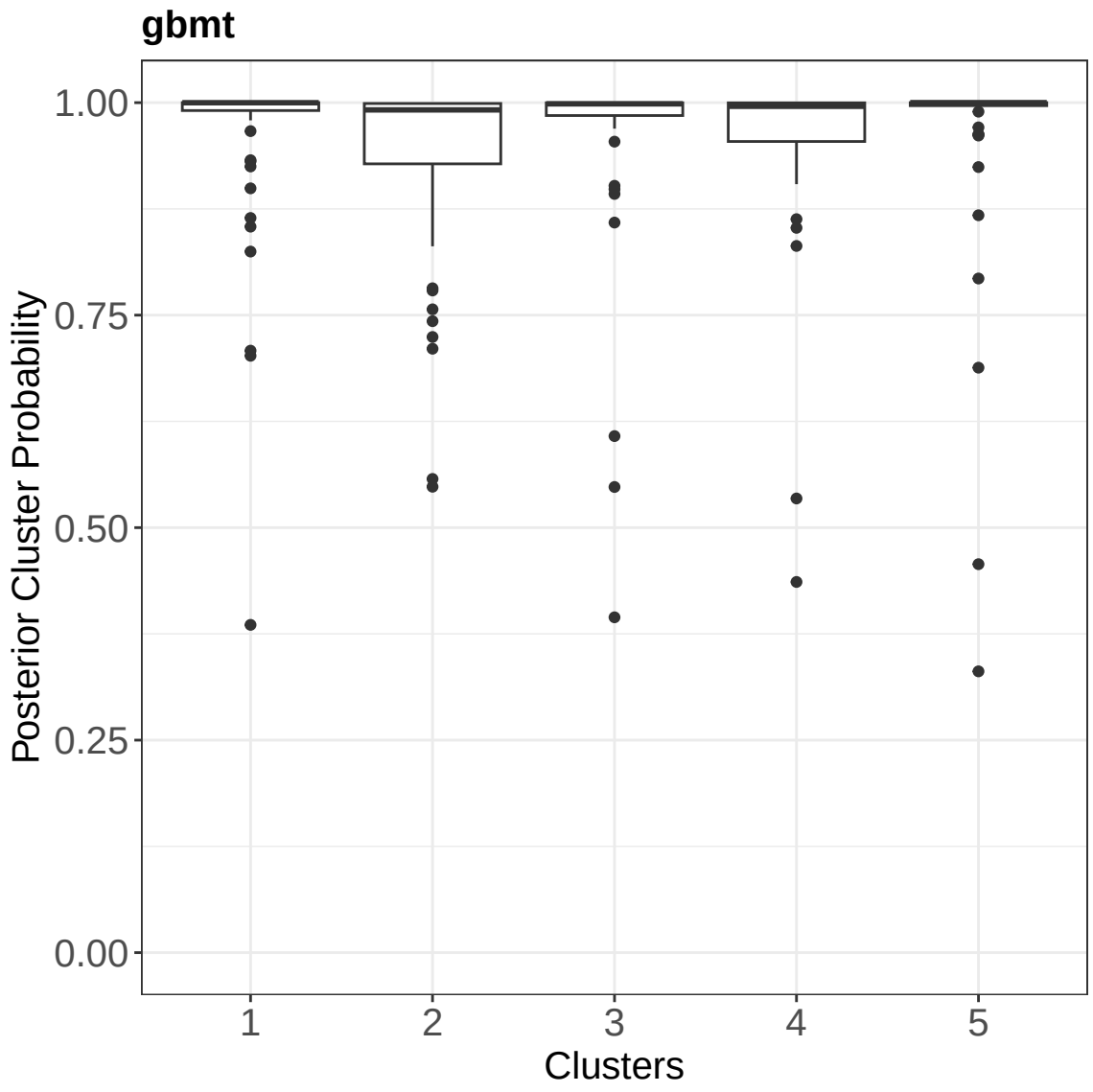
```
##  EM iteration 0. Log likelihood: -5859.2585  EM iteration 1. Log likelihood: -5555.1588  EM iteratio
```

```
# Posterior Cluster Probability of Assignment
postprob <- apply(posterior(fit_gbmt),1,max)

# relabeling the clusters to be consistent with other methods
cluster.re <- (fit_gbmt$assign ==2)*1 +
          (fit_gbmt$assign ==5)*2 +
          (fit_gbmt$assign ==1)*3 +
          (fit_gbmt$assign ==4)*4 +
          (fit_gbmt$assign ==3)*5
# Keep last observation per id
dnew_uq <- dnew910.before[!duplicated(dnew910.before$id, fromLast=TRUE),]
dnew_uq$cluster.gbmt <- cluster.gbmt <- cluster.re
dnew_uq$postprob <- postprob
# Posterior Cluster Probability of Assignment
library(ggplot2)
```

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

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




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

```
library(ggplot2)
library(cowplot)
p1.gbmt <- ggplot(data =dnew, aes(x =month, y = lbili,
                                  color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt))+
    ggtitle("gbmt")+
        geom_smooth(aes(x =month, y = lbili,
                            color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        theme_bw() + ylim(c(min(dnew$lbili,na.rm=TRUE),max(dnew$lbili,na.rm=TRUE)))+
        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") +
        scale_color_manual(values=c("green", "black","blue","red","purple"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple"))
p2.gbmt <- ggplot(data =dnew, aes(x =month, y = lalbumin,
                                  color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt))+
    ggtitle("gbmt")+
        geom_smooth(aes(x =month, y = lalbumin,
                        color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt),
                    method = "loess", linewidth= 3,se = FALSE,span=2)+
        theme_bw() + ylim(c(min(dnew$lalbumin,na.rm=TRUE),max(dnew$lalbumin,na.rm=TRUE)))+
        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") +
        scale_color_manual(values=c("green", "black","blue","red","purple"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple"))
p3.gbmt <- ggplot(data =dnew, aes(x =month, y = lalk.phos,
                                  color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt))+
    ggtitle("gbmt")+
        geom_smooth(aes(x =month, y = lalk.phos,
                        color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        theme_bw() + ylim(c(min(dnew$lalk.phos,na.rm=TRUE),max(dnew$lalk.phos,na.rm=TRUE)))+
        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") +
        scale_color_manual(values=c("green", "black","blue","red","purple"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple"))
p4.gbmt <- ggplot(data =dnew, aes(x =month, y = lsgot,
```

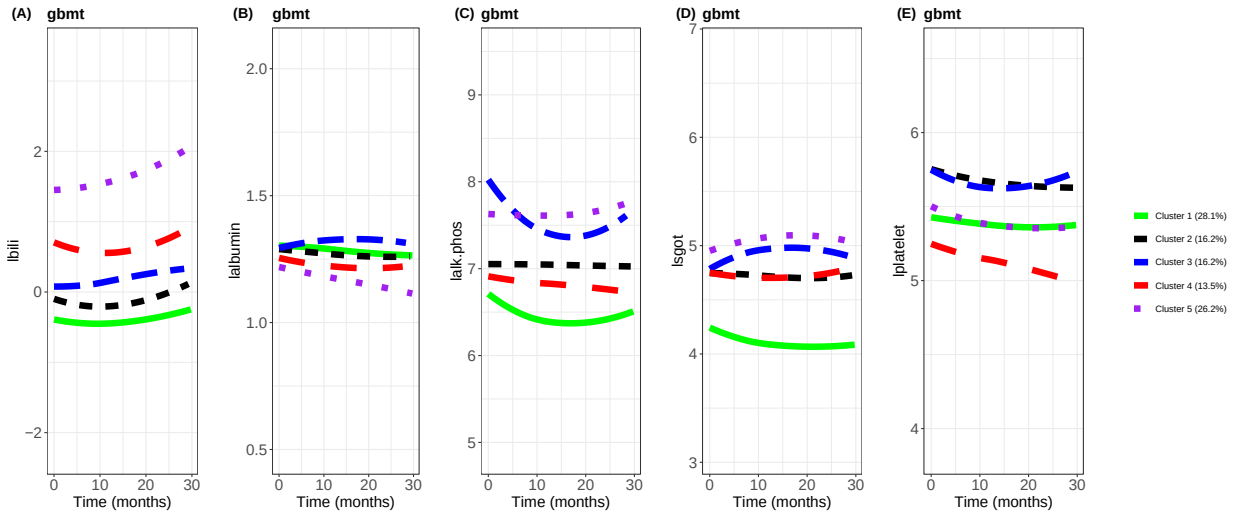
```
                                color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt))+
    ggtitle("gbmt")+
        geom_smooth(aes(x =month, y = lsgot,
                        color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        theme_bw() + ylim(c(min(dnew$lsgot,na.rm=TRUE),max(dnew$lsgot,na.rm=TRUE)))+
        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") +
        scale_color_manual(values=c("green", "black","blue","red","purple"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple"))
p5.gbmt <- ggplot(data =dnew, aes(x =month, y = lplatelet,
                                color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt))+
    ggtitle("gbmt")+
        geom_smooth(aes(x =month, y = lplatelet,
                        color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt),
                    method = "loess", linewidth = 3,se = FALSE,span=2)+
        theme_bw() + ylim(c(min(dnew$lplatelet,na.rm=TRUE),max(dnew$lplatelet,na.rm=TRUE)))+
        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") +
        scale_color_manual(values=c("green", "black","blue","red","purple"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple"))
#-------------------------------------------------------------------------------------------------------
# extract a legend that is laid out horizontally
legend.gbmt <- get_legend(ggplot(data =dnew, aes(x =month, y = lplatelet,
                          color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt))+
                           ggtitle("gbmt")+
                           geom_smooth(aes(x =month, y = lplatelet,
                         color=cluster.gbmt,linetype=cluster.gbmt,fill=cluster.gbmt),
                           method = "loess", linewidth = 3,se = FALSE,span=2)+
                           theme_bw() + ylim(c(min(dnew$lplatelet,na.rm=TRUE),
                                               max(dnew$lplatelet,na.rm=TRUE)))+
                           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"))+
        scale_fill_manual(values=c("green", "black","blue","red","purple"))
)
```

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

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

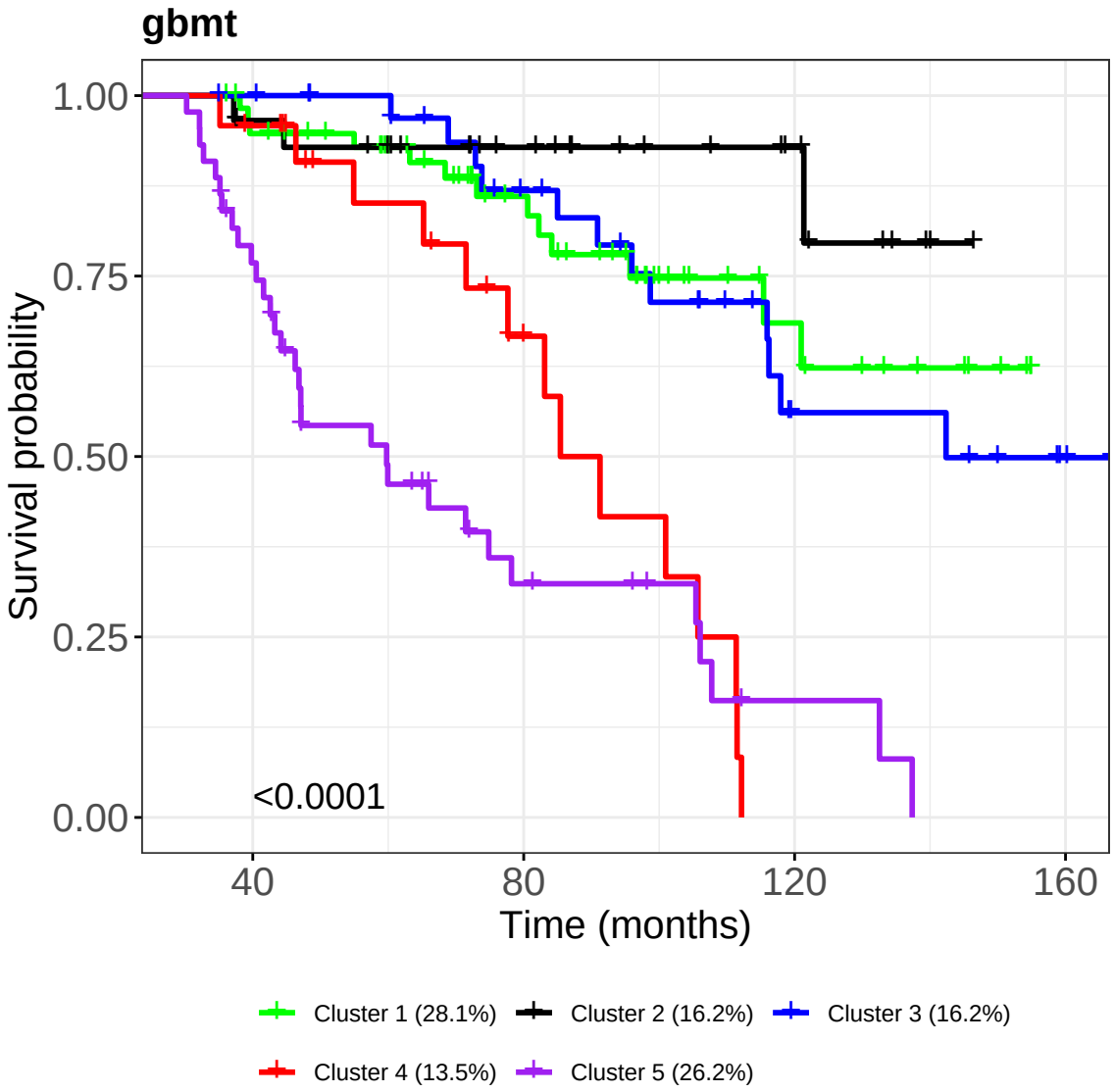
```
## Warning: package 'ggpubr' was built under R version 4.2.2
```

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

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


gbmt
1.00
0.75
0.50
0.25
0.00
Survival probability
<0.0001
40
80
120
160
Time (months)
Cluster 1 (28.1%)
Cluster 2 (16.2%)
Cluster 3 (16.2%)
Cluster 4 (13.5%)
Cluster 5 (26.2%)