Figure 1B

Vi Dang

2022-09-10

```r
#Load library
library(tidyverse)
library(ggplot2)
library(FactoMineR)
library(factoextra)
library(MixtureInf)
```

```r
#Load data
setwd("D:/PhD/git_PhD/TSS-cluster-Classification/")
Expo_For<-read.table("241EXPO.d17.fwd.norm.txt",header=T)%>%
  mutate(strand="+",GrowthPhase="EXPO",temperature="30")
Expo_Rev<-read.table("241EXPO.d17.rev.norm.txt",header=T)%>%
  mutate(strand="-",GrowthPhase="EXPO",temperature="30")
Stat_For<-read.table("241STAT.d17.fwd.norm.txt",header=T)%>%
  mutate(strand="+",GrowthPhase="STAT",temperature="30")
Stat_Rev<-read.table("241STAT.d17.rev.norm.txt",header=T)%>%
  mutate(strand="-",GrowthPhase="STAT",temperature="30")
Expo37_For<-read.table("241EXPO-37.d17.fwd.norm.txt",header=T)%>%
  mutate(strand="+",GrowthPhase="EXPO",temperature="37")
Expo37_Rev<-read.table("241EXPO-37.d17.rev.norm.txt",header=T)%>%
  mutate(strand="-",GrowthPhase="EXPO", temperature="37")
Stat37_For<-read.table("241STAT-37.d17.fwd.norm.txt",header=T)%>%
  mutate(strand="+",GrowthPhase="STAT", temperature= "37")
Stat37_Rev<-read.table("241STAT-37.d17.rev.norm.txt",header=T)%>%
  mutate(strand="-",GrowthPhase="STAT", temperature="37")
Expo_Combined<-rbind(Expo_For,Expo_Rev)%>%
  mutate(PhaseTemp=interaction(GrowthPhase,temperature))
Stat_Combined<-rbind(Stat_For,Stat_Rev)%>%
  mutate(PhaseTemp=interaction(GrowthPhase,temperature))
Expo37_Combined<-rbind(Expo37_For,Expo37_Rev)%>%
  mutate(PhaseTemp=interaction(GrowthPhase,temperature))
Stat37_Combined<-rbind(Stat37_For,Stat37_Rev)%>%
  mutate(PhaseTemp=interaction(GrowthPhase,temperature))
```

```r
#Construct the SI and Size with a range of percentile using only high express genes (small sets)
Small_sets<-list(Expo_Combined,Stat_Combined,Expo37_Combined,Stat37_Combined)
Sum_function<-function(set){
Combined_long<-set%>%
  pivot_longer(cols=c("Size","SI"),
               names_to = "feature",
               values_to = "value")
Combined_long%>%group_by(feature)%>%
  summarise(fifty_per_range=quantile(value,probs=c(0.25,0.75)),
            sixty_per_range=quantile(value,probs=c(0.20,0.80)),
            seventy_per_range=quantile(value,probs=c(0.15,0.85)),
            eighty_per_range=quantile(value,probs=c(0.10,0.90)),
            ninety_per_range=quantile(value,probs=c(0.05,0.95)),
            ninetyfive_per_range=quantile(value,probs=c(0.025,0.975)))
}

Criteria<-map_df(Small_sets,Sum_function)%>%as.data.frame()
```

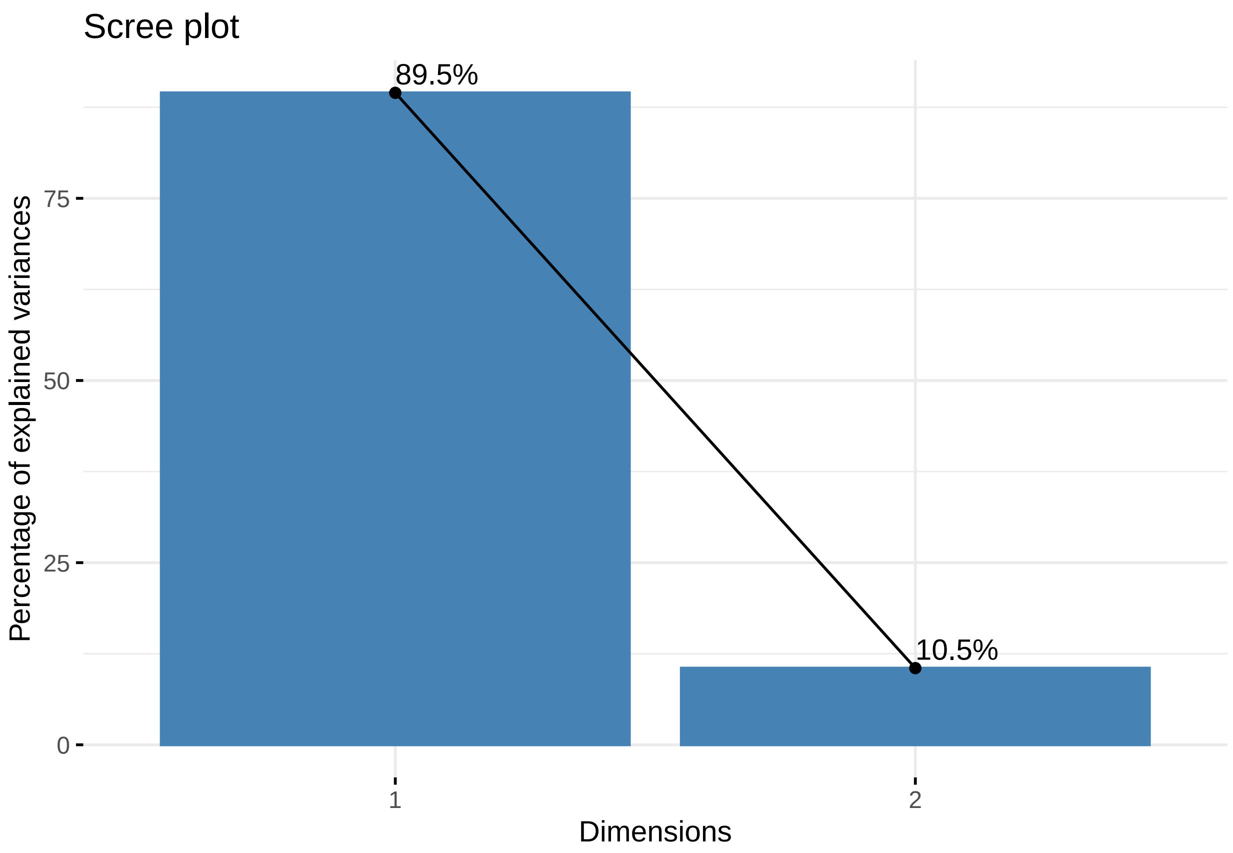
```
## ‘summarise()‘ has grouped output by ’feature’. You can override using the
## ‘.groups‘ argument.
## ‘summarise()‘ has grouped output by ’feature’. You can override using the
## ‘.groups‘ argument.
## ‘summarise()‘ has grouped output by ’feature’. You can override using the
## ‘.groups‘ argument.
## ‘summarise()‘ has grouped output by ’feature’. You can override using the
## ‘.groups‘ argument.
```

```
Criterias<-list(Criteria[1:4,],Criteria[5:8,],Criteria[9:12,],Criteria[13:16,])
Typical_cluster<-function(Set,cri=Criterias[[1]]){
Set%>%
  mutate(Typical_50=if_else(SI>= cri[1,2] & SI<= cri[2,2] & Size>=cri[3,2] & Size <=cri[4,2],"typical",
         Typical_60=if_else(SI>= cri[1,3] & SI<= cri[2,3] & Size>=cri[3,3] & Size <=cri[4,3],"typical",
         Typical_70=if_else(SI>= cri[1,4] & SI<= cri[2,4] & Size>=cri[3,4] & Size <=cri[4,4],"typical",
         Typical_80=if_else(SI>= cri[1,5] & SI<= cri[2,5] & Size>=cri[3,5] & Size <=cri[4,5],"typical",
         Typical_90=if_else(SI>= cri[1,6] & SI<= cri[2,6] & Size>=cri[3,6] & Size <=cri[4,6],"typical",
         Typical_95=if_else(SI>= cri[1,7] & SI<= cri[2,7] & Size>=cri[3,7] & Size <=cri[4,7],"typical",
}

results_cri_expo30<-map(Small_sets,Typical_cluster)
```

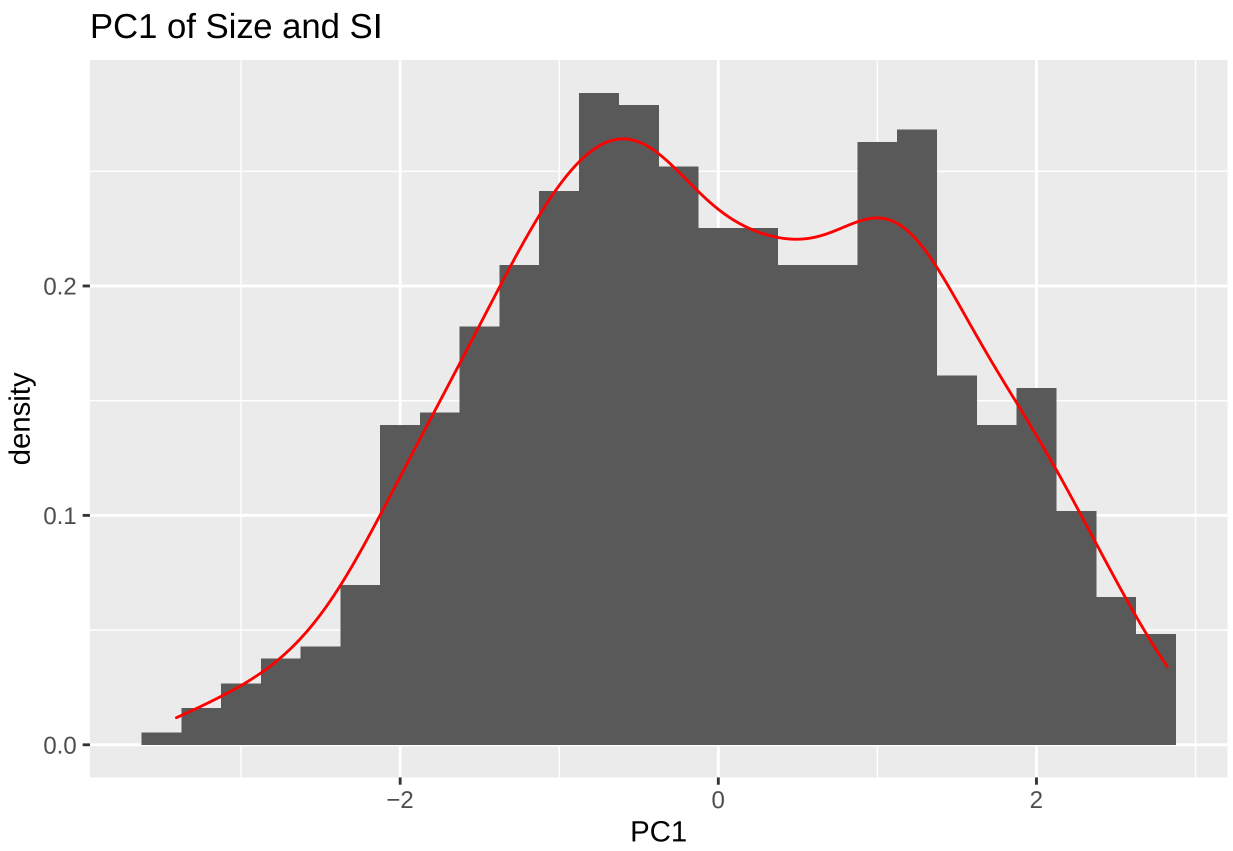
```
#The satisfied EXPO30 set
EXPO30_95<-results_cri_expo30[[1]]
EXPO30_95<-EXPO30_95%>%filter(Typical_95=="typical")
EXPO30_95<-EXPO30_95%>%filter(!str_detect(Id,"_\\D"))%>%select(1:11)
nrow(EXPO30_95)   #746
```

```
## [1] 746
```

```
#PCA
Expo_pca95<-PCA(EXPO30_95,scale.unit=TRUE, graph=F, ncp=5, quali.sup=c(1:6,9:11))
fviz_screeplot(Expo_pca95,addlabels=T)
```



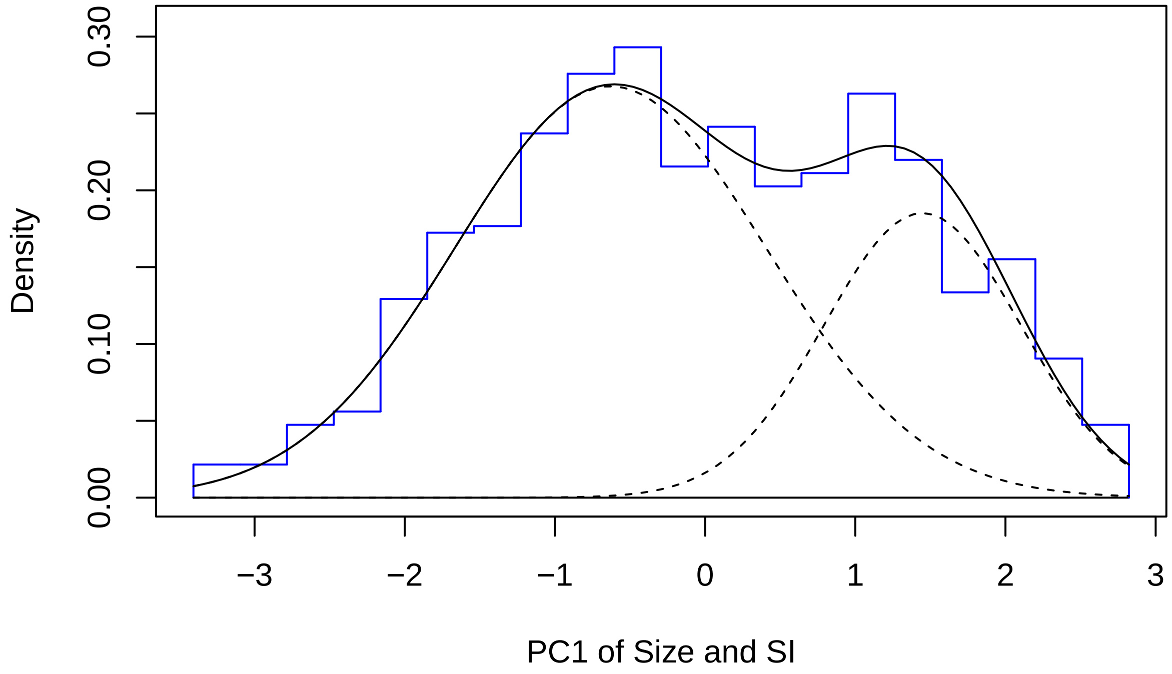
```
Expo_PC1_95<-cbind(EXPO30_95[,c(1,2,5,7:9)],Expo_pca95$ind[[1]][,1])
colnames(Expo_PC1_95)<-c("Id","Chr","pos_max","SI","Size","strand","PC1")
Expo_PC1_95<-as.data.frame(Expo_PC1_95)
Expo_PC1_95<-Expo_PC1_95%>%
  mutate(PC1=as.double(PC1))%>%
  arrange(PC1)

#density plot of PC1
Expo_PC1_95%>%ggplot(aes(x=PC1))+
  geom_histogram(aes(y=..density..),binwidth = 0.25)+
  geom_density(color="red")+
  labs(title="PC1 of Size and SI")
```



```
#Package MixtureInf detects two population of PC1
emtest.norm(Expo_PC1_95$PC1)
```

```
## $`MLE of Parameters under null hypothesis (order = m0)`
##                  [,1]
## alpha      1.000e+00
## mean      -1.680e-15
## variance   1.792e+00
##
## $`Parameter estimates under the order = 2m0`
##           [,1]   [,2]
## alpha    0.304  0.696
## mean     1.447 -0.630
## variance 0.429  1.076
##
## $`EM-test Statistics`
## [1] 40.806
##
## $`P-values`
## [1] 1.38e-09
##
## $`Level of penalty`
## [1] 1.00 0.25
```

```
plotmix.norm(Expo_PC1_95$PC1,theta=c(0.304,0.696,1.447,-0.630,0.429,1.076),
             hist = 1,
             xlab = "PC1 of Size and SI",
             ylab="Density")
```


Density
PC1 of Size and SI