Cluster for whole data via Gower distance

Weijia Xiong

6/30/2020

```
load("data/DiGiulio.RData")
otu_data = as.data.frame(DiGiulio$OTU)  # 927 samples, 1271 OTU
taxonomy = DiGiulio$Taxonomy  # 1271
sampledata = DiGiulio$SampleData  #  927 samples, other covariates
```

```
otu_data_all=
  cbind(sampledata, otu_data) %>%
  mutate(
    Preg = as.factor(Preg),
    Subject = as.factor(Subject)
  ) %>%
  na.omit()
```

Term data

```
term =
  otu_data_all %>%
  filter(preterm == "Term")

term_data =
  term %>%
  dplyr::select(-SampleID,-Subject)
```

Gower distance for mixed variables

```
gower_dist <- daisy(term_data, metric = "gower")
gower_mat <- as.matrix(gower_dist)
```

```
#' Print most similar
term[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind = TRUE)[1, ], 1:10]
```

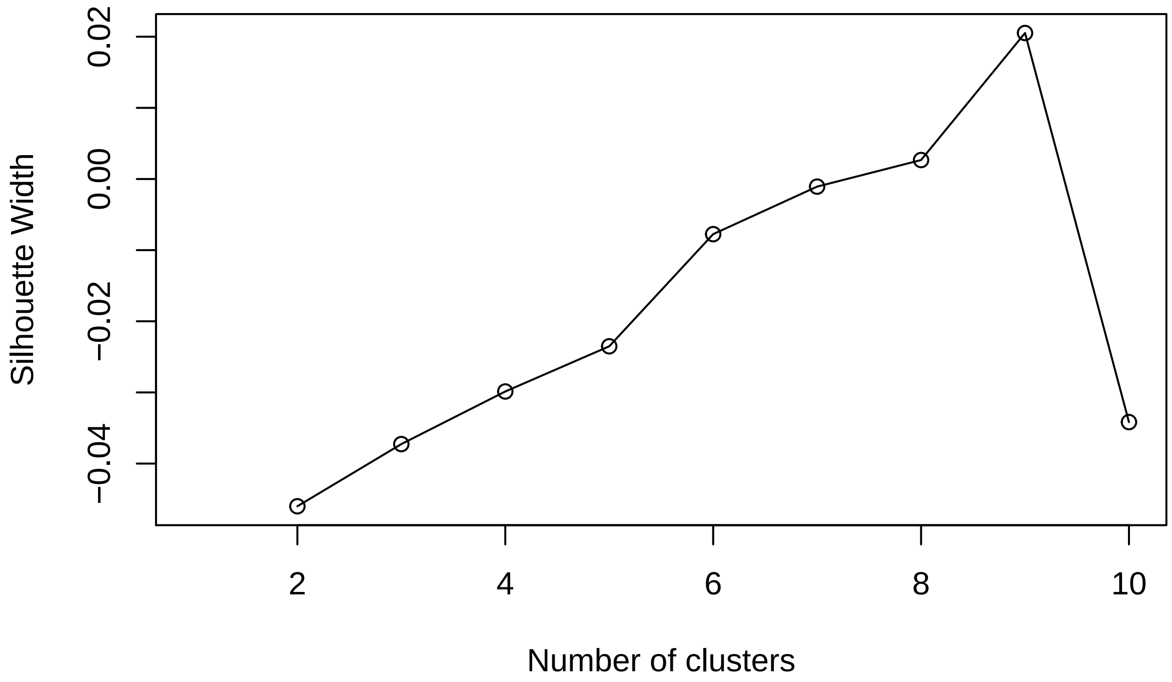
```
##      SampleID Subject weeks  Race NumReads Preg preterm CST 4330849 4400869
## 27 1000601208   10006    20 White     2193 TRUE    Term   0       0       0
## 26 1000601198   10006    19 White     2385 TRUE    Term   0       0       0
```

```
#' Print most dissimilar
term[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]), arr.ind = TRUE)[1, ], 1:10]
```

```
##          SampleID Subject weeks  Race NumReads  Preg preterm CST 4330849
## 458 1004301328.rs   10043    32 White     5708  TRUE    Term   1       0
## 51     1000601718   10006    71 White    10165 FALSE    Term   0       0
##     4400869
## 458       0
## 51        0
```

Calculate silhouette width for many k using PAM

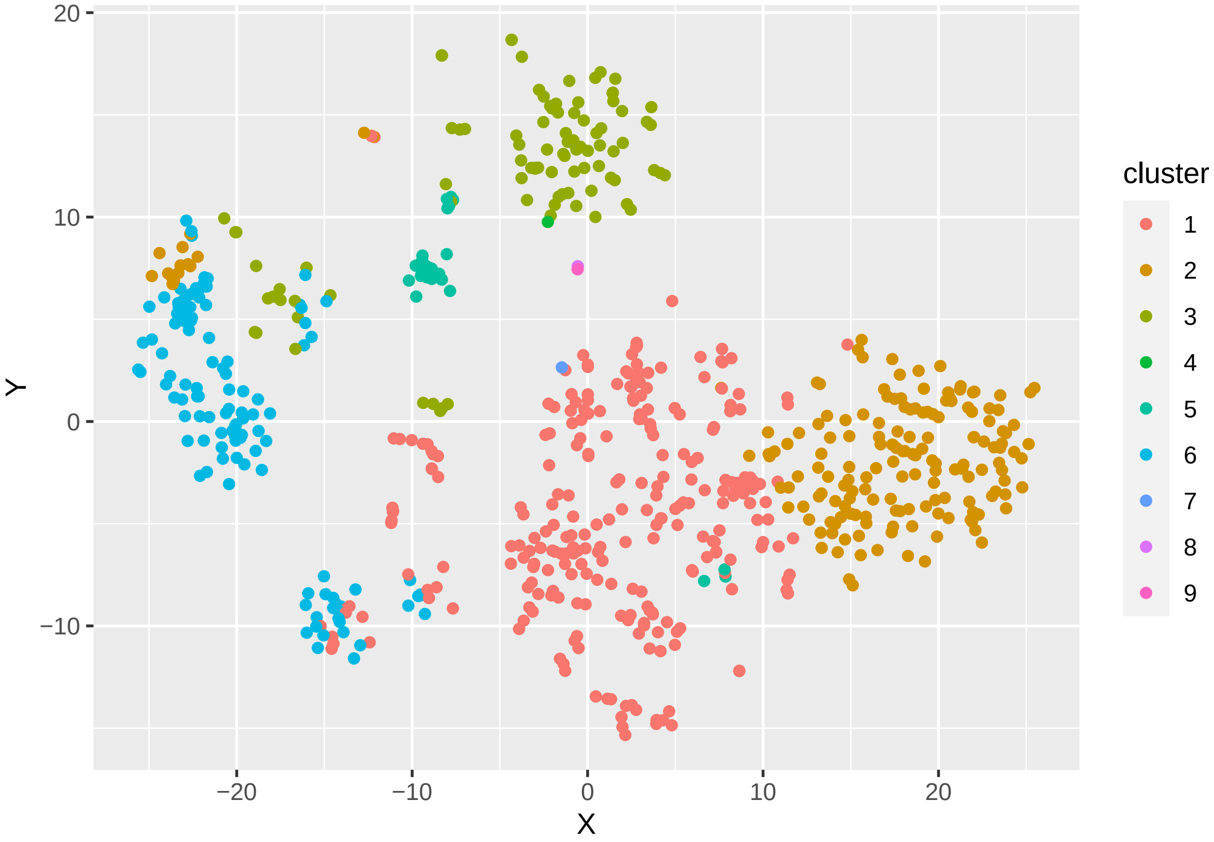
```
## Cluster
sil_width <- c(NA)
for(i in 2:10){
  pam_fit <- pam(gower_dist, diss = TRUE, k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:10, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:10, sil_width)
```



```
k <- 9
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- term_data %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
result = pam_results$the_summary
term[pam_fit$medoids, 1:10]
```

```
##       SampleID Subject weeks                  Race NumReads  Preg preterm CST
## 137 1002101308   10021    30                 White     3408  TRUE    Term   0
## 159 1002201268   10022    27                 White     5668  TRUE    Term   0
## 534 1004501618   10045    61                 White     3820 FALSE    Term   0
## 51  1000601718   10006    71                 White    10165 FALSE    Term   0
## 424 1004001338   10040    33                Indian     4335  TRUE    Term   0
## 630 1900501178   19005    18 Other (Specify below)     6134  TRUE    Term   0
## 389 1003901258   10039    26                 White     8045  TRUE    Term   0
## 404 1003901458    10039     46                White      2218 FALSE     Term    0
## 408 1003901608    10039     61                White      5415 FALSE     Term    0
##      4330849 4400869
## 137        0       0
## 159        0       0
## 534        0       0
## 51         0       0
## 424        0       0
## 630        0       0
## 389        0       0
## 404        0       0
## 408        0       0
```

```
tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```



Preterm data

```
preterm =
  otu_data_all %>%
  filter(preterm != "Term")
```

```r
preterm_data =
  preterm %>%
  dplyr::select(-SampleID,-Subject)
```

Gower distance for mixed variables

```r
gower_dist <- daisy(preterm_data, metric = "gower")
gower_mat <- as.matrix(gower_dist)
```

```r
#' Print most similar
preterm[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind = TRUE)[1, ], 1:10]
```

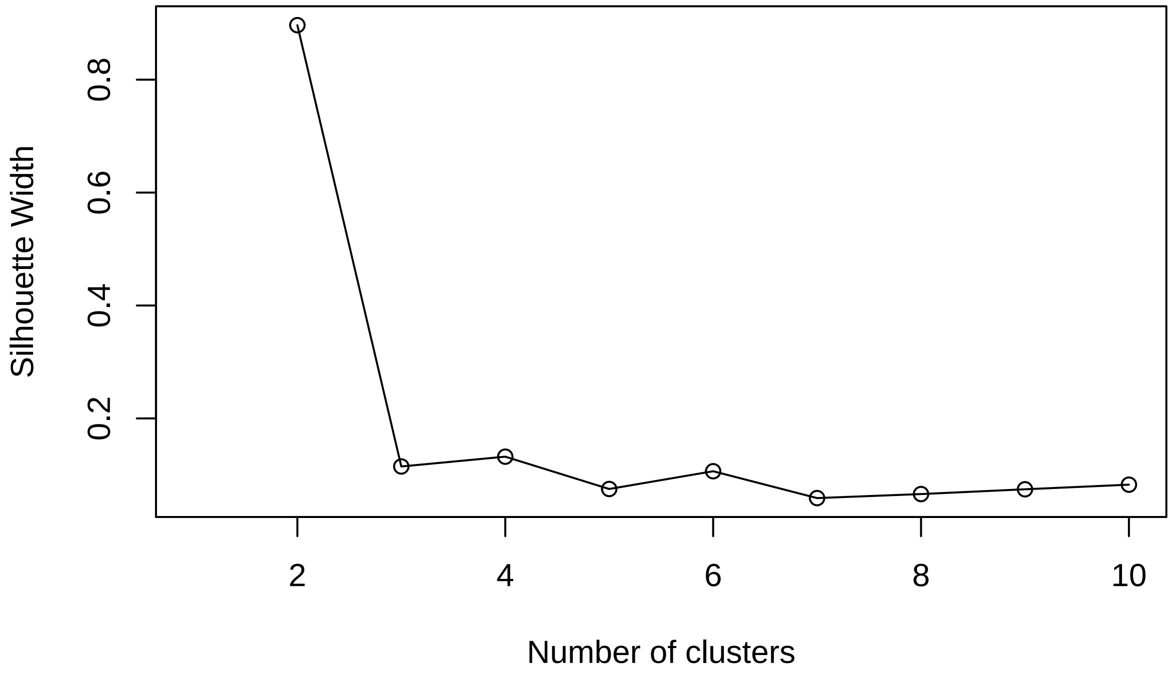
```
##       SampleID Subject weeks  Race NumReads Preg  preterm CST 4330849 4400869
## 195 1010101248   10101    14 White     8382 TRUE Marginal   0       0       0
## 194 1010101238   10101    14 White     8348 TRUE Marginal   0       0       0
```

```r
#' Print most dissimilar
preterm[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]), arr.ind = TRUE)[1, ], 1:10]
```

```
##       SampleID Subject weeks            Race NumReads  Preg  preterm CST
## 220 1010101618   10101    58           White     9103 FALSE Marginal   0
## 45  1001801118   10018    12 American Indian     3599  TRUE  Preterm   1
##     4330849 4400869
## 220       0       0
## 45        0       0
```

Calculate silhouette width for many k using PAM

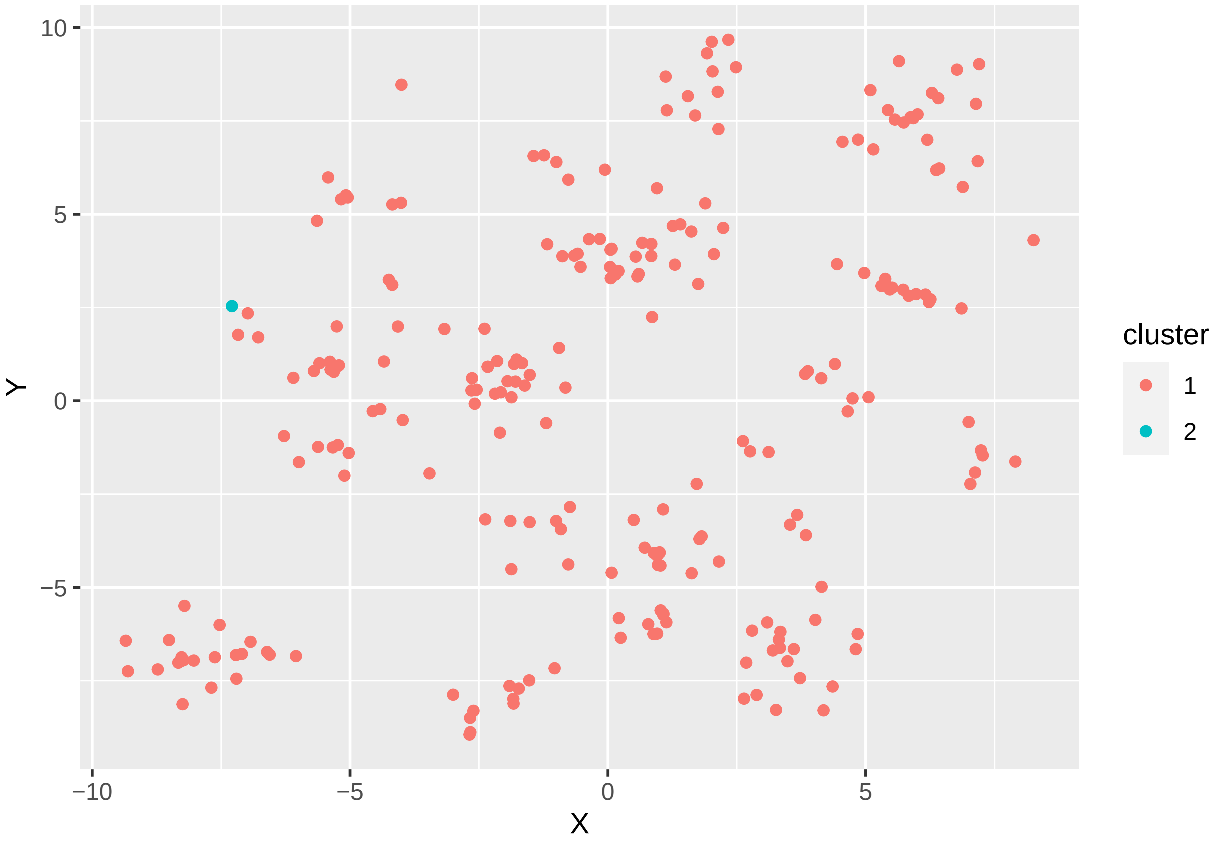
```r
## Cluster
sil_width <- c(NA)
for(i in 2:10){
  pam_fit <- pam(gower_dist, diss = TRUE, k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:10, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:10, sil_width)
```



```
k <- 2
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- preterm_data %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
result = pam_results$the_summary
term[pam_fit$medoids, 1:10]
```

```
##        SampleID Subject weeks  Race NumReads  Preg preterm CST 4330849 4400869
## 212 1002301618   10023    62 White     7341 FALSE    Term   0       0       0
## 220 1002401138   10024    14 White     5934  TRUE    Term   0       0       0
```

```
tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```


10
5
0
−5
Y
−10
−5
0
5
X
cluster
● 1
● 2