Getting started with DPBBM pcakge

Lin Zhang lin.zhang@cumt.edu.cn

September 21, 2016

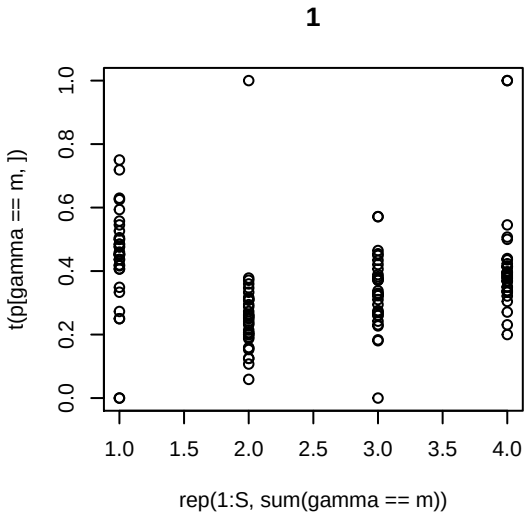
Load the package and generate a dataset.

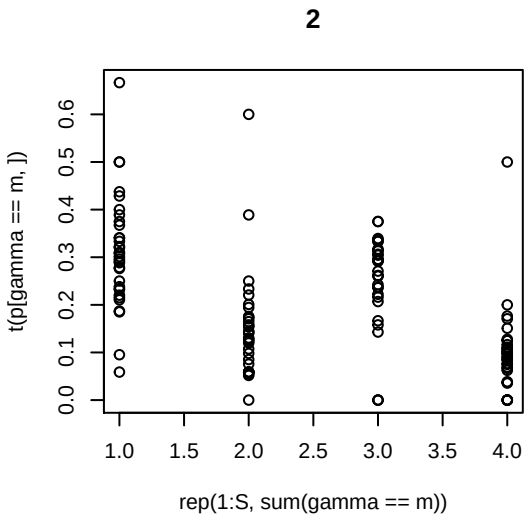
```
library(DPBBM)
```

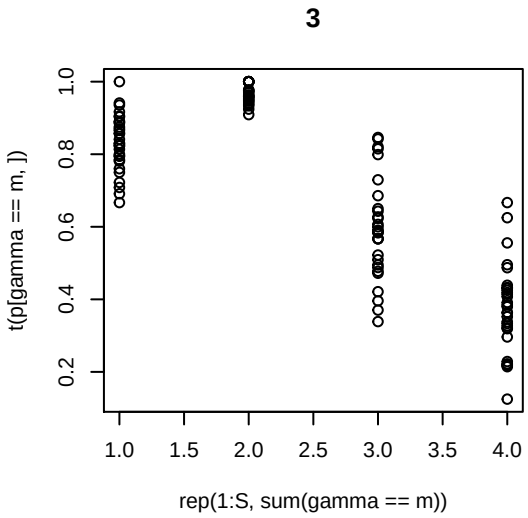
```
## Warning: replacing previous import by 'splines::splineDesign' when loading
## 'VGAM'
```

```
set.seed(123455)
S <- 4
G <- 100
K <- 3
nb_mu <- 100
nb_size <- 0.8
prob <- c(1,1,1)
mat <- bbm_data_generate(S=S,G=G,K=K,prob=prob,alpha_band=c(2,6),beta_band=c(2,6),
                         nb_mu=nb_mu,nb_size=nb_size, plotf = TRUE, max_cor=0.5)
```

```
## [1] "after tried 42 times"
## [1] "sample data generated with correlation smaller than: 0.5"
```

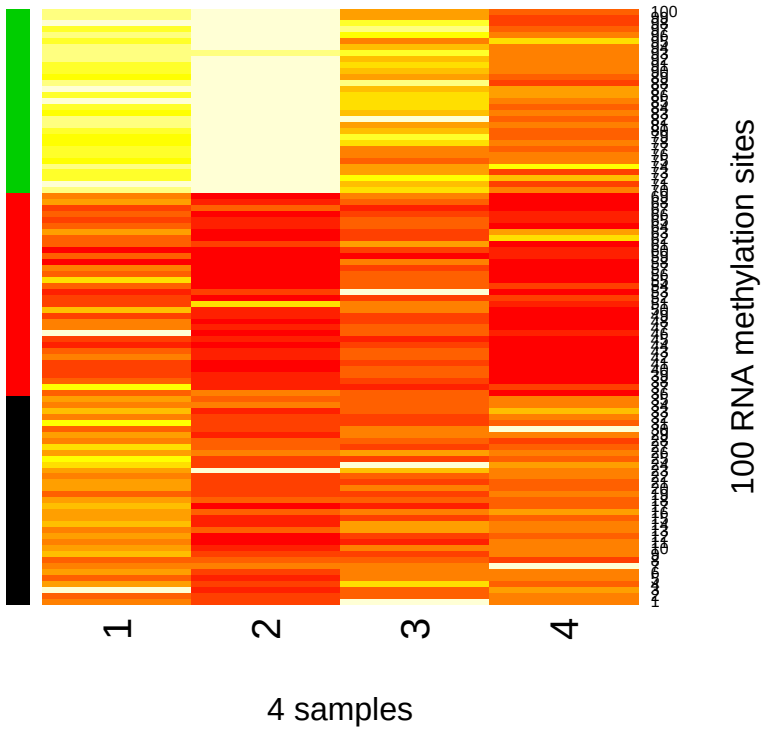

1
t(p[gamma == m,])
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
rep(1:S, sum(gamma == m))


2
t(p[gamma == m,])
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
rep(1:S, sum(gamma == m))


3
t(p[gamma == m,])
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
rep(1:S, sum(gamma == m))

check the generated data. The color on the left shows the true clustering IDs of the site.

```
id <- order(mat$gamma);
c <- mat$gamma[id]
mat_ratio <- (mat$k+1)/(mat$n+1);
heatmap(mat_ratio[id,], Rowv = NA, Colv = NA, scale="none", RowSideColors=as.character(c),
        xlab = "4 samples", ylab="100 RNA methylation sites")
```



Run the DPBBM result. This step takes a really long time.

```
cluster_label <- dpbbm_mc_iterations(mat$k, mat$n)
```

```
## [1] "Gibbs sampling started. It will take a long time."
## [1] "Shown only the clustering information in the first 20 iterations."
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [71] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [1]  2  1  1  1 13  1  1  2  4 11  2  1  1  1  4  2  1  1  1  1  1  1  3
## [24]  6  1  4  1  3  1  1  1  1  2  1  1  1 15  1  1  4
##  [1]  3  1  2  1 28  1  2  2  1 12  1  1  2  1  1  1  6  2 29  1  3  1  1
##  [1]  4  1  4 31 33  4  6  2  4  9  1  1  3  1
##  [1]  5  1  9 31 37  4  2  2  1  2  9  2
```

```
## [1]  6 14  4 31 42  7  1  1
## [1]  7 16  1 30 30 17  1  2  3
## [1]  8 13  1 30 26 27  2  1
## [1]  9 14  1 31 34 20
## [1] 10 23  1 30 25 15  4  2
##  [1] 11 12  3 32 24 25  1  1  1  1
## [1] 12  9  1 30 26 30  2  1  1
##  [1] 13  4  2 31 12 34  1 12  2  2
##  [1] 14  2  1 30 10 40  3 11  2  1
## [1] 15  1  1 30  9 35  1 22  1
## [1] 16 20  1 32  1 45  1
## [1] 17 21  1 32  1 45
## [1] 18 20  1 31  1 46  1
## [1] 19 17  3 30  1 47  2
## [1] 20 25  1 32  1 41
```

Show the cluster sizes.

```
table(cluster_label)
```

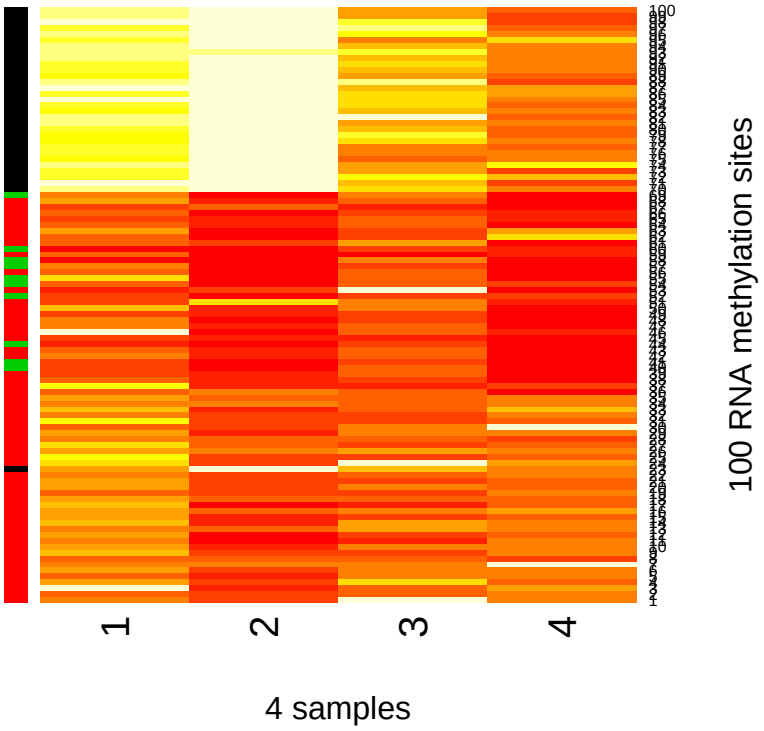
```
## cluster_label
##  1  2  3
## 32 58 10
```

```
table(mat$gamma)
```

```
##
##  1  2  3
## 35 34 31
```

Compare the clustering result with the true clustering IDs.

```
id <- order(mat$gamma);
c <- cluster_label;
mat_ratio <- (mat$k+1)/(mat$n+1);
heatmap(mat_ratio[id,], Rowv = NA, Colv = NA, scale="none",
          RowSideColors = as.character(cluster_label[id]),
          xlab = "4 samples", ylab="100 RNA methylation sites")
```



As is shown, clustering results are consistent for most of the sites, but there exist a few misclassied sites as well.