betaclust: a family of beta mixture models for beta valued DNA methylation data

Koyel Majumdar

2022-09-10

Introduction

A differentially methylated CpG (DMC) site is a CpG site which has a different methylation state between different samples. Identifying the DMCs between benign and malignant tissue samples can help understand disease and its treatment. The methylation values are known as beta values and can be modelled using beta distributions. Due to a lack of suitable methods for the beta values in their innate form, beta values are usually transformed to M-values, which can be modelled using Gaussian distribution. The methylation state of a CpG locus is measured as hypermethylated if both the alleles are methylated, hypomethylated if neither of the alleles is methylated and hemimethylated otherwise. The beta values are constrained between 0 and 1 and a beta value close to 0 suggests hypomethylation whereas a value close to 1 suggests hypermethylation. Typically, arbitrary thresholds are selected to identify the methylation states and based on these, differentially methylated regions are identified between different samples.

The package **betaclust** contains a family of novel beta mixture models (BMMs) to (i) appositely model the innate beta valued data, (ii) objectively identify methylation state thresholds and (iii) identify the differentially methylated CpG sites using a model-based clustering approach. The family of BMMs employs different parameter constraints applicable to different study settings. Parameter estimation proceeds via the EM algorithm, with a novel approximation during the M-step providing tractability and ensuring computational feasibility.

This document gives a quick tour of the functionalities in **betaclust**. See `help(package="betaclust")` for further details and references provided by `citation("betaclust")`.

Walk through

Prerequisites

Before starting the **betaclust** walk through, the user should have a working R software environment installed on their machine. The **betaclust** package has the following dependencies which, if not already installed on the machine will automatically be installed along with the package: **foreach, doParallel, stats, utils, ggplot2, plotly**.

Assuming that the user has the **devtools** and **betaclust** packages installed, the user first needs to load them:

```
library(devtools)
library(betaclust)
```

Loading the data

The **betaclust** package provides a preprocessed methylation dataframe which contains beta values of DNA samples collected from 4 patients suffering from high-grade prostate cancer. The samples are collected from benign and tumor prostate tissues. The methylation profiling of these samples is done using the Infinium MethylationEPIC Beadchip technology. The dataset comprises *R = 2* DNA samples collected from each of *N = 4* patients and each sample contains beta values at *C = 694,820* CpG sites. The data were collected for a study on prostate cancer methylomics (Silva et al. 2020).

The methylation array data was quality controlled and preprocessed using the **RnBeads** package (Mueller et al. 2019). The data were then normalized and probes located outside of CpG sites and on the sex chromosome were filtered out. The CpG sites with missing values were removed from the resulting dataset. The user can load the data available in the package and look at the first 6 rows present in the dataframe as follows:

```
data(pca.methylation.data)
head(pca.methylation.data)
#>       IlmnID Benign_Patient_1 Benign_Patient_2 Benign_Patient_3
#> 1 cg14817997        0.9289358        0.9371402        0.9406154
#> 2 cg26928153        0.8945523        0.9140979        0.8416376
#> 3 cg16269199        0.7499294        0.8568665        0.7106825
#> 4 cg12045430        0.1764357        0.2129593        0.1446858
#> 5 cg20826792        0.1353336        0.2385359        0.1102691
#> 6 cg20253340        0.7672509        0.7055869        0.5647254
#>   Benign_Patient_4 Tumour_Patient_1 Tumour_Patient_2 Tumour_Patient_3
#> 1        0.9268555        0.9208016        0.9563707        0.9288394
#> 2        0.9307889        0.8753521        0.9582674        0.9196868
#> 3        0.8832769        0.7528014        0.8781866        0.7939099
#> 4        0.1087709        0.1766444        0.2378470        0.1387028
#> 5        0.1361931        0.1599522        0.1956061        0.1356275
#> 6        0.6199342        0.5664521        0.5762739        0.6626614
#>   Tumour_Patient_4
#> 1        0.9281329
#> 2        0.9348243
#> 3        0.9123488
#> 4        0.2098278
#> 5        0.1480923
#> 6        0.5314872
```

Identifying methylation thresholds in a DNA sample

The K.. and KN. models (Majumdar et al. 2022) are used to analyse a single DNA sample (*R* = *1*) and cluster the CpG sites into *K* = *M* groups, where *M* is the number of methylation states of a CpG site (typically CpG sites can be either hypomethylated, hemimethylated or hypermethylated). The thresholds are objectively inferred from the clustering solution. The optimal model can be selected using the AIC, BIC or ICL model selection criterion. The KN. model is selected as the optimal model using BIC to cluster the CpG sites in the benign sample into 3 methylation states and objectively infer the thresholds.

```
M <- 3 # No. of methylation states in a DNA sample
N <- 4 # No. of patients
R <- 1 # No. of DNA samples
my.seed <- 190 ## set seed for reproducibility

threshold_out <- betaclust(pca.methylation.data[,2:5],M,N,R,
                         model_names = c("K..","KN."),
                         model_selection = "BIC",
                         seed = my.seed)
#> fitting ...
#> 
  |
  |                                                                      |   0%
  |
  |===================================                                   |  50%
  |
  |======================================================================| 100%
```

Summary statistics of the clustering solution

Summary statistics can then be obtained using the `summary()` function (see `help("summary.betaclust")`). The output contains the clustering solution of the model selected as optimal.

```
summary(threshold_out)
#> ------------------------------------------------------
#> Multivariate beta mixture model fitted by EM algorithm
#> ------------------------------------------------------
#> betaclust KN. model with 3 components:
#>  log-likelihood Information-criterion IC-value CpG-sites Patients Samples
#>         2185315                   BIC -4370280    694820        4       1
#> 
#> Clustering table:
#>      1      2      3 
#> 273823 169903 251094 
#> 
#> Proportion of CpG sites in each cluster:
#> 0.394 0.245 0.361
```

Thresholds objectively inferred from the clustering solution

The thresholds calculated based on the estimated shape parameters are provided in the output:

```
threshold_points <- threshold_out$optimal_model_results$thresholds
threshold_points$threholds
#>      Patient 1 Patient 2 Patient 3 Patient 4
#> [1,]     0.258     0.252     0.189     0.198
#> [2,]     0.747     0.773     0.766     0.813
```

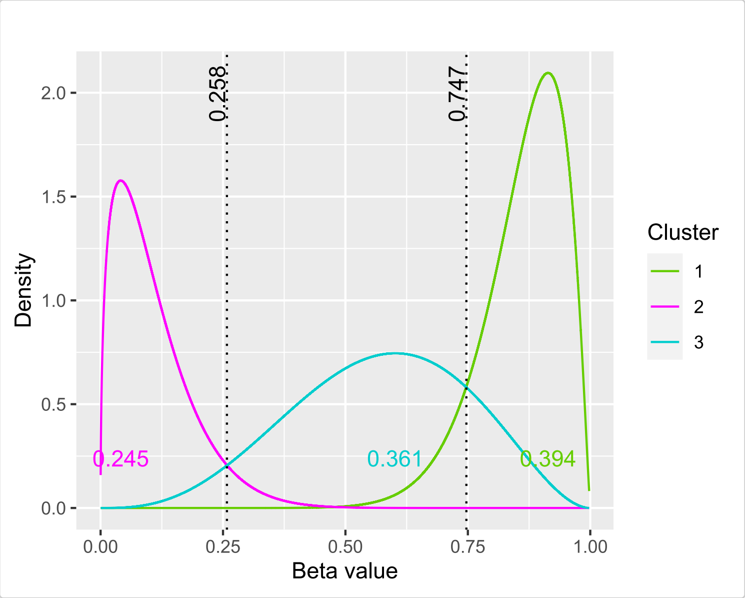
Plotting the density estimates of the clustering solution

The clustering solution can be visualized using the `plot()` function. The fitted and kernel density estimates, uncertainty and model selection plots can be obtained using the `what` parameter in this function. Apart from static plots using `plot_type = "ggplot"` interactive plots are also available using `plot_type = "plotly"`.

The fitted densities and the proportion of data in each cluster is displayed in the plots. The thresholds for the methylation states can be displayed using `threshold = TRUE`. As the K.. model constrains the shape parameters to be equal for each patient, a single pair of threshold points is calculated for all patients, and this results in the same fitted density for all patients. In the KN. model, a pair of thresholds is independently determined for each patient based on the estimated shape parameters since the shape parameters differ for each patient. The parameter `patient number` can be used to choose the patient for whom the fitted density and thresholds need to be visualized. For visualization, the option accepts the patient's column number as a value.

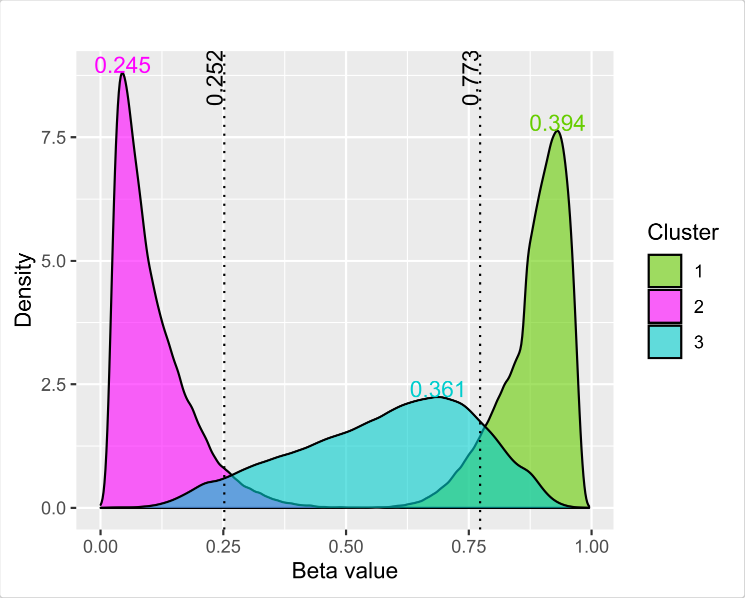
The fitted density and threshold points for the first patient in the dataset can be visualized as shown below:

```
plot(threshold_out,what = "fitted density",threshold = TRUE, patient_number = 1,plot_type =
         "ggplot")
```



The kernel density estimates for second patient is displayed as below:

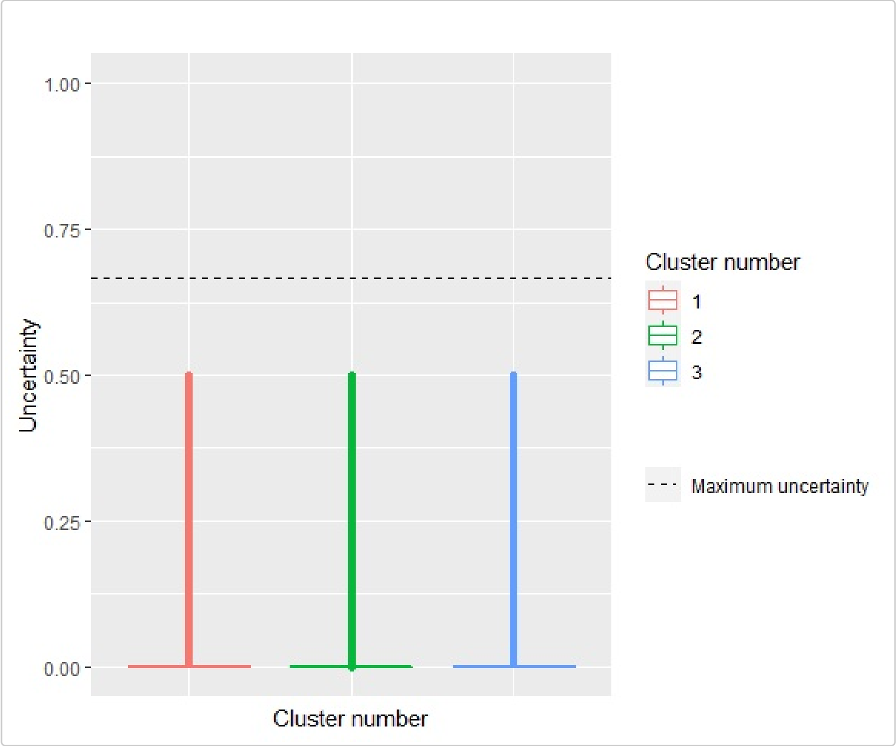
```
plot(threshold_out,what = "kernel density",threshold = TRUE, patient_number = 2,plot_type =
         "ggplot")
```



Plotting the uncertainties in the clustering solution

The uncertainties in clustering represent the probability of a CpG site not belonging to the corresponding cluster. A low uncertainty shows good decision in grouping the CpG site in the cluster it is highly likely to belong. A boxplot of the uncertainties in the clustering solution can be obtained.

```
plot(threshold_out,what = "uncertainty")
```



Identifying DMCs between *R = 2* samples

The methylation values in multiple samples can be analysed and the differentially methylated CpG sites between these samples can be identified. The K.R model (Majumdar et al. 2022) clusters the $C \times (NR)$ data into $K = M^R$ groups.

A total of *R* = *2* DNA samples were collected from each patient (Silva et al. 2020). Each DNA sample has CpG sites that belong to one of the *M* methylation states, typically *M* = *3*. As a result, the *C* CpG sites are clustered into $K = M^R = 9$ groups.

```
M <- 3  #No. of methylation states in a DNA sample
N <- 4  #No. of patients
R <- 2  #No. of DNA samples

dmc_output <- betaclust(pca.methylation.data[,2:9],M,N,R,
                      model_names = "K.R",
                      seed = my.seed)
#> fitting ...
#> 
  |                                                                            
  |                                                                      |   0%
  |                                                                            
  |======================================================================| 100%
```

Summary statistics of the clustering solution

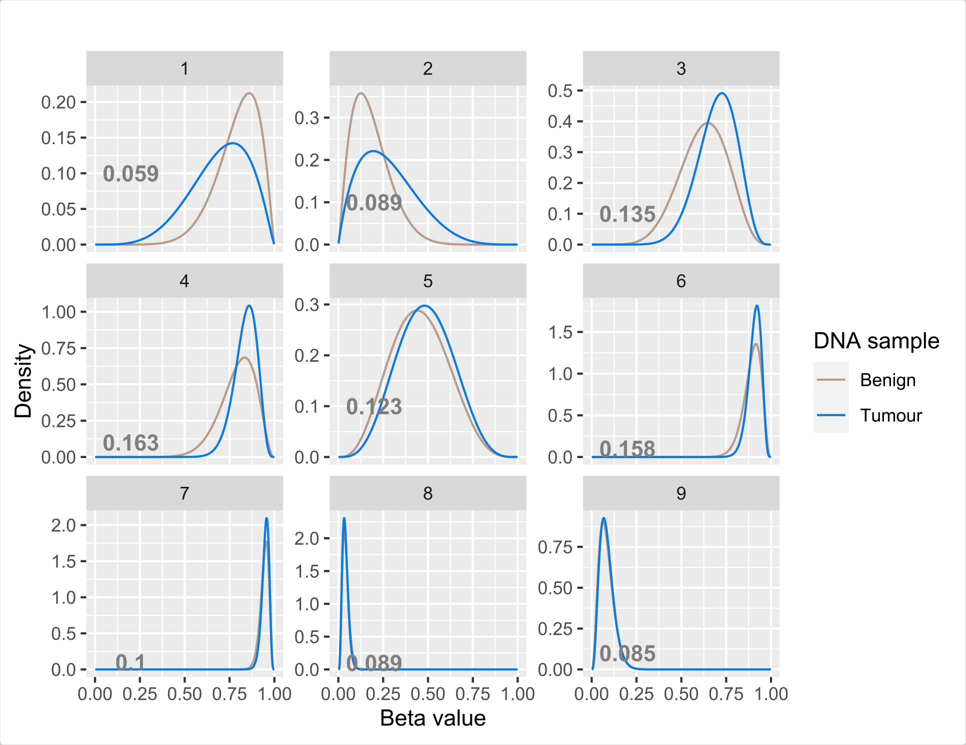
Summary statistics of the clustering solution can then be obtained as follows:

```
summary(dmc_output)
#> -------------------------------------------------------
#> Multivariate beta mixture model fitted by EM algorithm 
#> -------------------------------------------------------
#> betaclust K.R model with 9 components: 
#>  log-likelihood Information-criterion  IC-value CpG-sites Patients Samples
#>         6192125                   BIC -12384063    694820        4       2
#> 
#> Clustering table:
#>      1      2      3      4      5      6      7      8      9 
#>  40886  61871  93468 113158  85142 109736  69245  61972  59342 
#> 
#> Proportion of CpG sites in each cluster: 
#> 0.059 0.089 0.135 0.163 0.123 0.158 0.1 0.089 0.085
```

Plotting the density estimates of the clustering solution

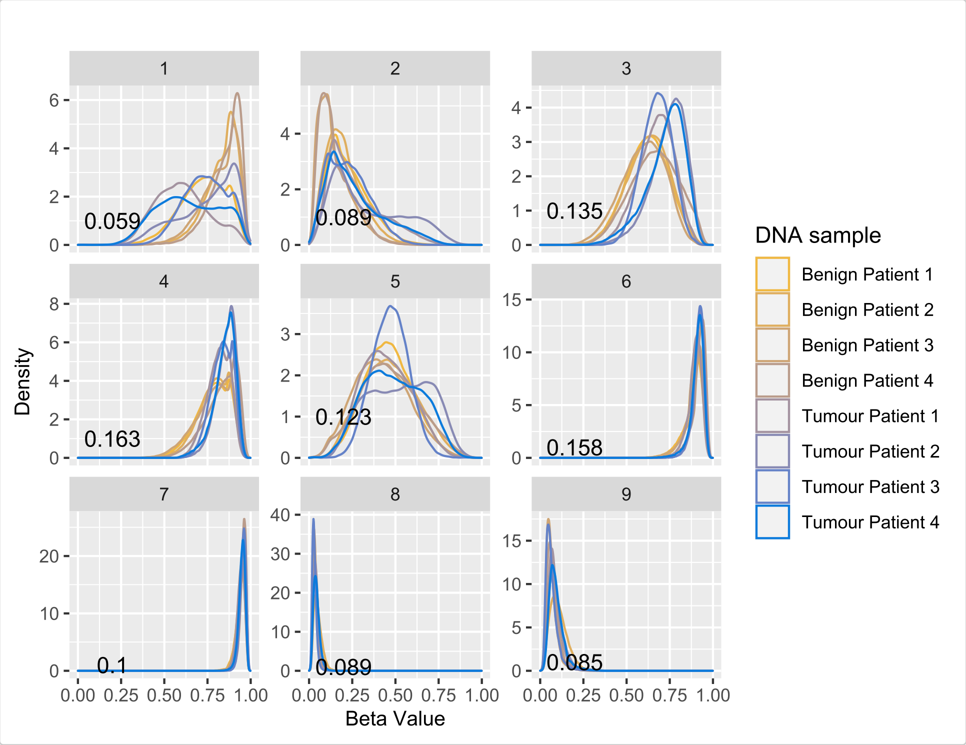
The fitted density estimates of the clustering solution can be visualised. The names of the DNA samples used in the analysis are passed to the function using `sample_name`. If no input is provided in `sample_name` then default values of sample names, for e.g. Sample 1, Sample 2, etc. are used. The proportion of CpG sites belonging to each cluster is also displayed in each panel. From the plot it can be observed that clusters 1-4 identify the CpG sites where the methylation state tends to change between the two samples whereas clusters 5-9 identify CpG sites which tend to have the same methylation states in both samples.

```
plot(dmc_output,what = "fitted density",plot_type = "ggplot",sample_name =
        c("Benign","Tumour"))
```



The kernel density estimates of the clustering solution are obtained as below:

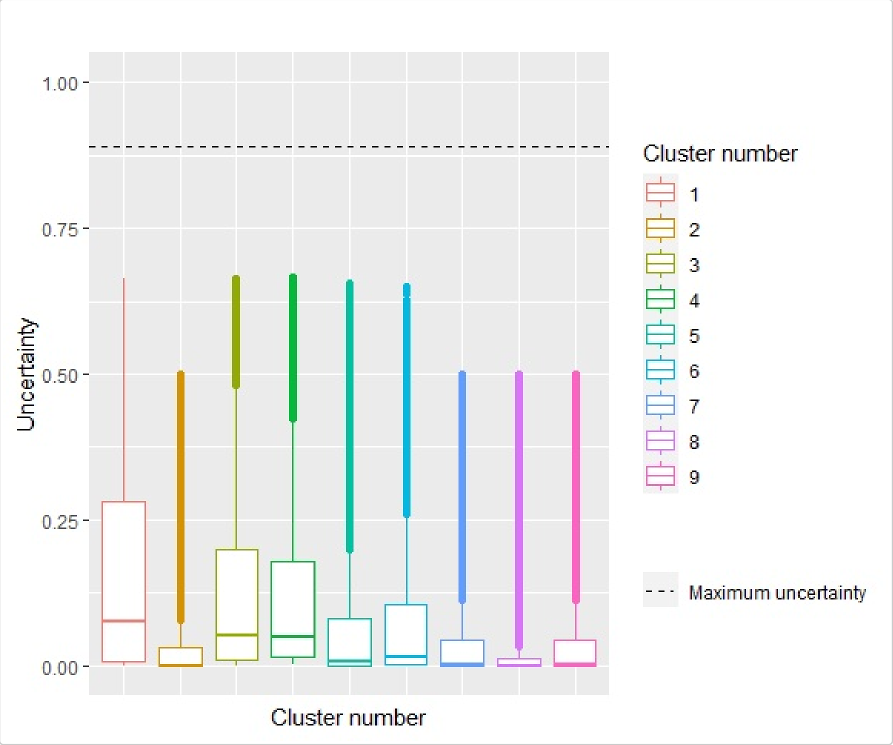
```
plot(dmc_output,what = "kernel density",plot_type = "ggplot")
```



Plotting the uncertainty in the clustering solution

The uncertainties in the clustering solution can be plotted as follows:

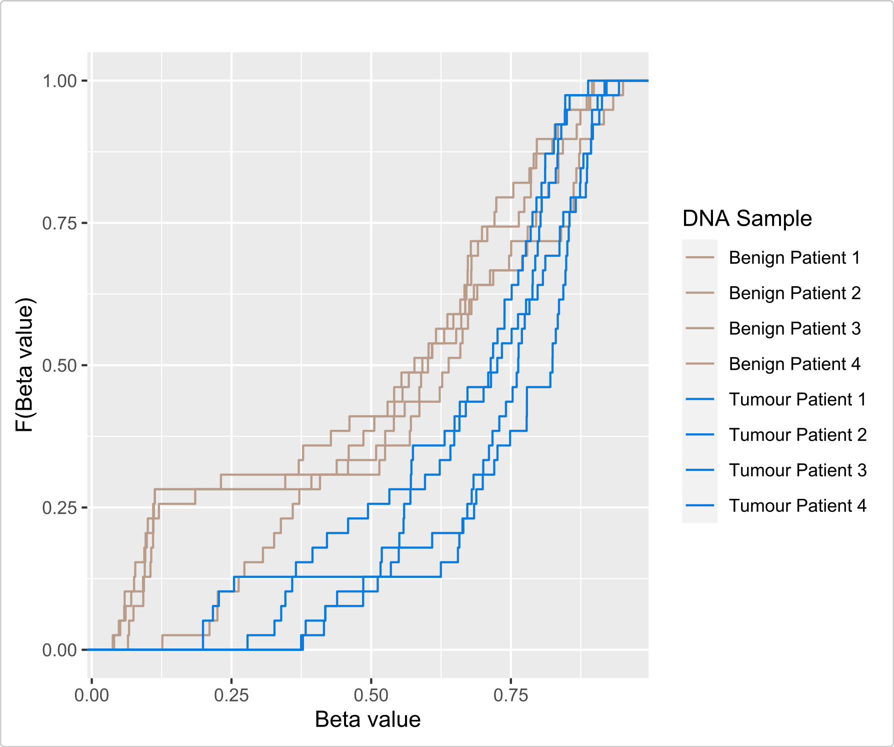
```
plot(dmc_output,what = "uncertainty",plot_type = "ggplot")
```



Plotting the empirical cumulative distribution function

Hypermethylation of RARB genes is an important biomarker in prostate cancer studies (Ameri et al. 2011). The MethylationEPIC annotated data (see `help("legacy.data")`) available in the package **betaclust** can be used to obtain information on the genes to which the selected DMCs are related. The differentially methylated CpG sites related to RARB genes are selected and passed to the `ecdf.betaclust()` function to visualise the empirical distribution functions (see `help("ecdf.betaclust")`). From the plot it is visible that the CpG sites in the tumour samples have increased beta values compared to those in the benign samples, suggesting hypermethylation of the CpG sites.

```
## save the clustering solution in a dataframe
dmc_df <- dmc_output$optimal_model_results$data

## merge the IlmnID column to the dataframe and change the column name to IlmnID
dmc_df <- as.data.frame(cbind(pca.methylation.data$IlmnID,dmc_df))
colnames(dmc_df)[1] <- "IlmnID"

## select the differentially methylated CpG sites identified using the K.R model
dmc_df <- dmc_df[dmc_df$mem_final == "1" |dmc_df$mem_final == "2" |dmc_df$mem_final == "3"|
                 dmc_df$mem_final == "4", ]

##read the legacy data
data(legacy.data)
head(legacy.data)
#>       IlmnID Genome_Build CHR   MAPINFO       UCSC_RefGene_Name
#> 1 cg07881041           37  19   5236016 PTPRS;PTPRS;PTPRS;PTPRS
#> 2 cg18478105           37  20  61847650                  YTHDF1
#> 3 cg23229610           37   1   6841125
#> 4 cg03513874           37   2 198303466
#> 5 cg09835024           37   X  24072640                  EIF2S3
#> 6 cg05451842           37  14  93581139       ITPK1;ITPK1;ITPK1
#>      UCSC_CpG_Islands_Name
#> 1    chr19:5237294-5237669
#> 2  chr20:61846843-61848103
#> 3     chr1:6844313-6846366
#> 4 chr2:198299244-198299972
#> 5   chrX:24072558-24073135
#> 6  chr14:93581083-93582797

## create empty dataframes and matrices to store the DMCs related to the RARB genes
ecdf_rarb <- data.frame()
cpg_rarb <- data.frame(matrix(NA,nrow = 0,ncol = 6))
colnames(cpg_rarb) <- colnames(legacy.data)

## split the UCSC_RefGene_name column to read the gene name related to that CpG site
## select the CpG sites related to the RARB genes
a <- 1
for(i in 1:nrow(legacy.data))
{
  str_vec = unlist(strsplit(as.character(legacy.data[i,"UCSC_RefGene_Name"]),"[;]"))
  if(length(str_vec) != 0)
  {
    if(str_vec[[1]] == "RARB")
    {
      cpg_rarb[a,] <- legacy.data[i,]
      a <- a+1
    }
  }
}

## Read the DMCs related to the RARB genes
ecdf_rarb <- dmc_df[dmc_df$IlmnID %in% cpg_rarb$IlmnID,]

## Plot the ecdf of the selected DMCs
ecdf.betaclust(ecdf_rarb[,2:9],R = 2,sample_name = c("Benign","Tumour"))
```



Wrapper function to fit all models in the family of beta mixture models

All the BMM models or a selection of BMM models can be fit using the wrapper function `betaclust()`. All 3 BMM models can be fit to a dataset and the optimal model can be selected using AIC, BIC or ICL values.

```
wrapper_out <- betaclust(pca.methylation.data[,2:9],M,N,R,
                        model_names = c("K..","KN.","K.R"),
                        model_selection = "BIC",
                        seed = my.seed)
#> fitting ...
#>
  |
  |                                                                      |   0%
#> Warning: K.. model only considers a single sample and not multiple samples.
#>                    Model is fitted to 1st sample only.
#>
  |
  |=======================                                               |  33%
#> Warning: KN. model only considers a single sample and not multiple samples.
#>                    Model is fitted to 1st sample only.
#>
  |
  |===============================================                       |  67%
  |
  |======================================================================| 100%
```

Summary statistics of the clustering solution

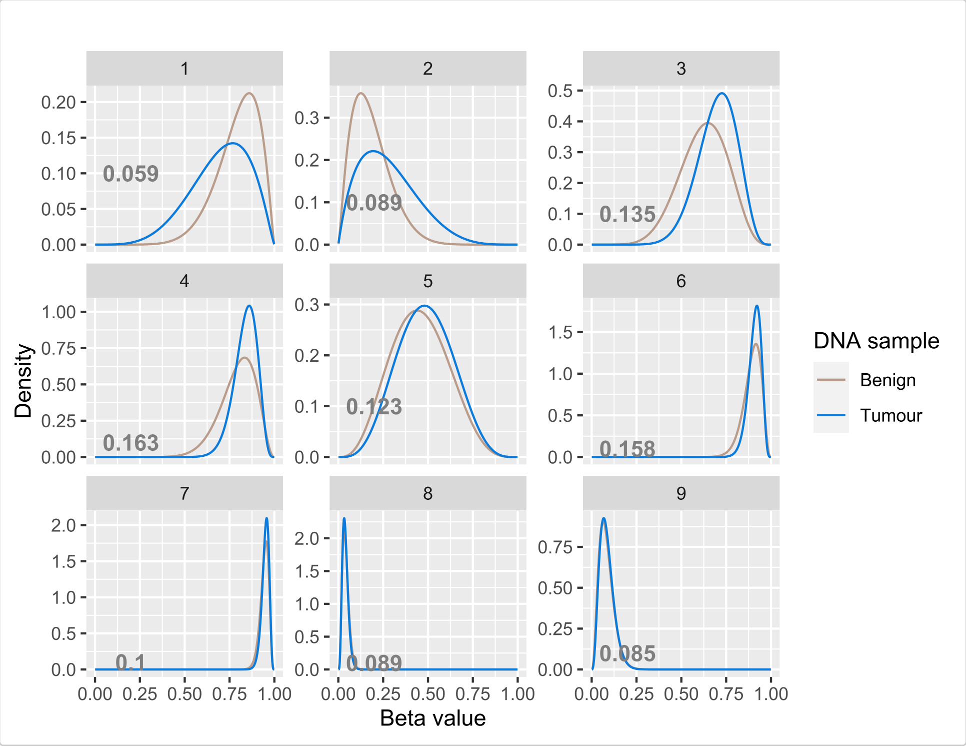
The summary statistics of the clustering solution of the optimal model can then be obtained. The K.R model has been selected as the optimal model to identify the DMCs between the benign and tumour samples. The proportion of data belonging to each group has been displayed below.

```
summary(wrapper_out)
#> ------------------------------------------------------
#> Multivariate beta mixture model fitted by EM algorithm
#> ------------------------------------------------------
#> betaclust K.R model with 9 components:
#>  log-likelihood Information-criterion  IC-value CpG-sites Patients Samples
#>         6192125                   BIC -12384063    694820        4       2
#>
#> Clustering table:
#>      1      2      3      4      5      6      7      8      9
#>  40886  61871  93468 113158  85142 109736  69245  61972  59342
#>
#> Proportion of CpG sites in each cluster:
#> 0.059 0.089 0.135 0.163 0.123 0.158 0.1 0.089 0.085
```

Plotting the density estimates of the clustering solution

The density estimates of the clustering solution of the optimal model can be visualised.

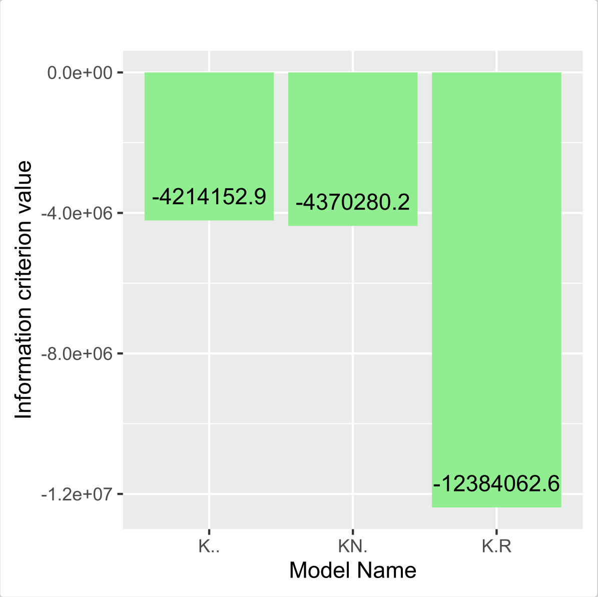
```
plot(wrapper_out,what = "fitted density",plot_type = "ggplot",sample_name =
        c("Benign","Tumour"))
```



Plotting the information criterion values of all models

The information criterion specified in the wrapper function provides values for all models fitted and can be visualised to support the selection of the optimal model.

```
plot(wrapper_out,what = "information criterion",plot_type = "ggplot")
```



References

Silva, R., Moran, B., Russell, N.M., Fahey, C., Vlajnic, T., Manecksha, R.P., Finn, S.P., Brennan, D.J., Gallagher, W.M., Perry, A.S.: Evaluating liquid biopsies for methylomic profiling of prostate cancer. Epigenetics 15(6-7), 715-727 (2020). doi:10.1080/15592294.2020.1712876.

Majumdar, K., Silva, R., Perry, A.S., Watson, R.W., Murphy, T.B., Gormley, I.C.: betaclust: a family of mixture models for beta valued DNA methylation data.

Mueller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, Bock C (2019): RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biology, 20(55). doi:10.1186/s13059-019-1664-9.

Ameri A, Alidoosti A, Hosseini SY, et al. Prognostic Value of Promoter Hypermethylation of Retinoic Acid Receptor Beta (RARB) and CDKN2 (p16/MTS1) in Prostate Cancer. Chinese journal of cancer research. 2011;23(4):306-311. doi:10.1007/s11670-011-0306-x.

Fraley, C., Raftery, A.E.: How many clusters? which clustering method? answers via model-based cluster analysis. The Computer Journal 41, 578-588 (1998). doi:10.1093/comjnl/41.8.578.

Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1-38 (1977). doi:10.1111/j.2517-6161.1977.tb01600.x.

Diamond, H.G., Straub, A.: Bounds for the logarithm of the euler gamma function and its derivatives. Journal of Mathematical Analysis and Applications 433(2), 1072-1083 (2016). doi:10.1016/j.jmaa.2015.08.034.