Package 'betaclust'

June 8, 2022

Type Package

Title Family of mixture models for beta valued DNA methylation data for Clustering and Density Estimation

Version 1.0.0

Author Koyel Majumdar [aut] <koyel.majumdar@ucdconnect.ie>,
Isobel Claire Gormley [aut] <claire.gormley@ucd.ie>,
Thomas Brendan Murphy [aut] <brendan.murphy@ucd.ie>

Maintainer Koyel Majumdar <koyel.majumdar@ucdconnect.ie>

Description A family of novel beta mixture models (BMMs) is proposed to appositely model the innate beta valued data, objectively identify methylation state thresholds and identify the differentially methylated CpG (DMC) 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.

License GPL-3

Depends R (>= 3.5.0)

Imports foreach, doParallel, stats, utils, ggplot2, plotly

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

NeedsCompilation no

R topics documented:

`betaclust` *The betaclust wrapper function*

Description

A family of model based clustering techniques to identify the methylation profiles of the beta valued DNA methylation data

Usage

```
betaclust(
  data,
  K = 3,
  patients,
  samples,
  model_names = "C..",
  model_selection = "BIC",
  seed,
  register = NULL
)
```

Arguments

`K`	number of methylation groups to be identified (default=3)
`patients`	number of patients in the study
`samples`	number of samples collected from each patient for study
`model_names`	mixture model to run (Models= c(C..,CN.,C.R), default=C..)
`model_selection`	optimal model selection based on information criterion. (Methods=AIC,BIC,ICL,default=BIC)
`seed`	seed for reproducible work
`register`	setting for parallelization
`X`	methylation values for CpG sites frpm R samples collected from N patients

Details

This is a wrapper function which can be used to run all three models (C.., CN., C.R) together. The C.. and CN. models are used to analyse a single DNA sample and cluster the CpG sites into the 3 methylation profiles (hypomethylation, hemimethylation, hypermethylation). The thresholds can be objectively identified from the clustering solution. The C. R model is used to analyse R samples to the differentially methylated CpG sites between R DNA samples.

Value

The function returns an object of "betaclust" class. The class object contains following values:

- Information_criterion - The information criterion used to select the optimal model.
- ic_output - This stores the information criterion value calculated for each model.
- optimal_model - The model selected as optimal.
- function_call - The parameters passed as arguments to the function betaclust.

- CpG_sites - The number of CpG sites analysed using the beta mixture models.
- patients - The number of patients analysed using the beta mixture models.
- samples - The numder of samples analysed using the beta mixture models.
- best_model - This contains the final results for the optimal model selected. Thus this contains the following values:
  - cluster_count - The total number of CpG sites identified in each cluster.
  - llk - The vector containing log-likelihood values calculated for each step of parameter estimation.
  - data - This contains the methylation dataset along with the cluster label as determined by the mixture model.
  - alpha - This contains the shape parameter 1 for the beta mixtures for K^R groups.
  - beta - This contains the shape parameter 2 for the beta mixtures for K^R groups.
  - tau - The proportion of CpG sites in each cluster.
  - z - The matrix contains the probability calculated for each CpG site belonging to the K^R clusters.
  - uncertainty - The uncertainty of a CpG site belonging to the identified cluster.

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.

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 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.

See Also

beta_c

beta_cn

beta_cr

pca.methylation.data

plot.betaclust

summary.betaclust

Examples

```
## Not run:
data(pca.methylation.data)
my.seed=190
K=3
patients=4
samples=2
data_output=betaclust(pca.methylation.data[,2:9],K,patients,samples,
            model_names=c("C..","CN.","C.R"),model_selection="BIC",seed=my.seed)
```

```
## End(Not run)
```

beta_c *The C.. model*

Description

The C.. model from the family of beta mixture models for DNA methylation data. This model analyses a single DNA sample and identify the thresholds for the different methylation profiles.

Usage

```
beta_c(data, K = 3, seed, register = NULL)
```

Arguments

`K`	number of methylation groups to be identified (default=3)
`seed`	seed for reproducible work
`register`	setting for parallelization
`X`	methylation values for CpG sites frpm R samples collected from N patients

Details

This model clusters each of the C CpG sites into one of K = M methylation states, based on data from N patients where R = 1. The default value for M = 3 as a CpG site can be either hypomethylated, hemimethylated or hypermethylated. Under the C.. model the shape parameters of each cluster are constrained to be equal for each patient.

Value

A list of clustering solution results.

- cluster_count - The total number of CpG sites identified in each cluster.
- llk - The vector containing log-likelihood values calculated for each step of parameter estimation.
- data - This contains the methylation dataset along with the cluster label as determined by the mixture model.
- alpha - This contains the shape parameter 1 for the beta mixtures for K^R groups.
- beta - This contains the shape parameter 2 for the beta mixtures for K^R groups.
- tau - The proportion of CpG sites in each cluster.
- z - The matrix contains the probability calculated for each CpG site belonging to the K^R clusters.
- uncertainty - The uncertainty of a CpG site belonging to the identified cluster.

See Also

`beta_cn`

`betaclust`

Examples

```
## Not run:
data(pca.methylation.data)
my.seed=190
K=3
data_output=beta_c(pca.methylation.data[,2:5],K,seed=my.seed)

## End(Not run)
```

`beta_cn` *The CN. model*

Description

The CN. model from the family of beta mixture models for DNA methylation data. This model analyses a single DNA sample and identify the thresholds for the different methylation profiles.

Usage

```
beta_cn(data, K = 3, seed, register = NULL)
```

Arguments

`K`	number of methylation groups to be identified (default=3)
`seed`	seed for reproducible work
`register`	setting for parallelization
`X`	methylation values for CpG sites frpm R samples collected from N patients

Details

This model clusters each of the C CpG sites into one of K = M methylation states, based on data from N patients where R = 1. The default value for M = 3 as a CpG site can be either hypomethylated, hemimethylated or hypermethylated. The CN. model differs from the C.. model as it is less parsimonious, allowing cluster and patient-specific shape parameters.

Value

A list of clustering solution results.

- cluster_count - The total number of CpG sites identified in each cluster.
- llk - The vector containing log-likelihood values calculated for each step of parameter estimation.
- data - This contains the methylation dataset along with the cluster label as determined by the mixture model.
- alpha - This contains the shape parameter 1 for the beta mixtures for K^R groups.
- beta - This contains the shape parameter 2 for the beta mixtures for K^R groups.
- tau - The proportion of CpG sites in each cluster.
- z - The matrix contains the probability calculated for each CpG site belonging to the K^R clusters.
- uncertainty - The uncertainty of a CpG site belonging to the identified cluster.

See Also

beta_c

betaclust

Examples

```
## Not run:
data(pca.methylation.data)
my.seed=190
K=3
data_output=beta_cn(pca.methylation.data[,2:5],K,seed=my.seed)

## End(Not run)
```

beta_cr *The C. R Model*

Description

Beta mixture model for identifying differentially methylated CpG sites between R DNA samples collected from N patients.

Usage

```
beta_cr(data, K = 3, patients, samples, seed, register = NULL)
```

Arguments

K	number of methylation groups to be identified (default=3)
patients	number of patients in the study
samples	number of samples collected from each patient for study
seed	seed for reproducible work
register	setting for parallelization
X	methylation values for CpG sites frpm R samples collected from N patients

Details

The C. R model allows identification of the differentially methylated CpG sites between the R DNA samples collected from each of the N patients. The model attempts to identify $K = M^R$ clusters identifying each possible combination of the M = 3 methylation profiles for R samples. The parameters vary for each sample type but are constrained to be equal for each patient.

An initial clustering using K-means is performed to identify K^R cluster. These values are provided as starting values to the Expectation-Maximisation algorithm. A digamma approximation is used to obtain the maximised parameters in the M-step instead of the computationally inefficient numerical optimisation step.

Value

A list of clustering solution results.

- cluster_count - The total number of CpG sites identified in each cluster.
- llk - The vector containing log-likelihood values calculated for each step of parameter estimation.
- data - This contains the methylation dataset along with the cluster label as determined by the mixture model.
- alpha - This contains the shape parameter 1 for the beta mixtures for K^R groups.
- beta - This contains the shape parameter 2 for the beta mixtures for K^R groups.
- tau - The proportion of CpG sites in each cluster.
- z - The matrix contains the probability calculated for each CpG site belonging to the K^R clusters.
- uncertainty - The uncertainty of a CpG site belonging to the identified cluster.

See Also

betaclust

Examples

```
## Not run:
data(pca.methylation.data)
my.seed=190
K=3
patients=4
samples=2
data_output=beta_cr(pca.methylation.data[,2:5],K,patients,samples,seed=my.seed)

## End(Not run)
```

---

ecdf.betaclust *The empirical cumulative distribution function*

Description

Empirical Cumulative Distribution Function (ECDF) plot for betaclust object

Usage

```
ecdf.betaclust(x, samples = 2, sample_name = c("Sample 1", "Sample 2"))
```

Arguments

x Methylation values of Identified Differentially methylated regions related to a gene. Group each sample together in the dataframe such that the columns are ordered as –> Sample1_P1, Sample1_P2, Sample2_P1, Sample2_P2

samples number of tissue samples from where DNA methylation data is collected (default samples=2)

sample_name The order in which the samples are grouped in the dataframe (default = c("Sample 1","Sample 2"))

Details

This function plots the ECDF graphs of the differentially methylated CpG sites identified using the C.R model for all patient samples. The graph can help visualise the methylation state changes between the different patient samples.

Value

The ECDF plot for the selected CpG sites for all patients and samples.

See Also

betaclust

beta_cr

---

em_aic *Akaike Information Criterion*

Description

The AIC value used to select the optimal model.

Usage

```
em_aic(llk, C, K, patients = 4, samples = 1, model_names = "C..")
```

Arguments

llk	log-likelihood value
C	number of CpG sites
K	number of clusters
patients	number of patients
samples	no. of samples
model_names	mixture model (method=c("C..","CN.","C.R"))

Details

Computes the AIC for the beta mixture models given the loglikelihood, the dimension of the data, and the mixture model names.

Value

The AIC value for the selected model.

See Also

em_bic

em_icl

em_bic *Bayesian Information Criterion*

Description

The BIC value used to select the optimal model.

Usage

```
em_bic(llk, C, K, patients = 4, samples = 1, model_names = "C..")
```

Arguments

`llk`	log-likelihood value
`C`	number of CpG sites
`K`	number of clusters
`patients`	number of patients
`samples`	no. of samples
`model_names`	mixture model (method=c("C..","CN.","C.R"))

Details

Computes the BIC for the beta mixture models given the loglikelihood, the dimension of the data, and the mixture model names.

Value

The BIC value for the selected model.

See Also

`em_aic`

`em_icl`

em_icl *Integrated Complete-data Likelihood (ICL) Criterion*

Description

The ICL value used to select the optimal model.

Usage

```
em_icl(llk, C, K, patients = 4, samples = 1, model_names = "C..", z)
```

Arguments

`llk`	log-likelihood value
`C`	number of CpG sites
`K`	number of clusters
`patients`	number of patients
`samples`	no. of samples
`model_names`	mixture model (method=c("C..","CN.","C.R"))
`z`	z matrix for each output

Details

Computes the ICL for the beta mixture models given the loglikelihood, the dimension of the data, and the mixture model names. This criterion penalises the BIC by including the entropy term favouring the well separated clusters.

Value

The ICL value for the selected model.

See Also

`em_aic`

`em_bic`

`legacy.data` *MethylationEPIC manifest data.*

Description

The dataset contains the manifest data from the Illumina MethylationEPIC beadchip array

Usage

```
data(legacy.data)
```

Format

A data frame with 867525 rows and 6 columns.

IlmnID This contains the Unique identifier from the Illumina CG database. (The probe ID).

Genome_Build Genome Build referenced by the manifest.

CHR Chromosome containing the CpG (Build 37).

MAPINFO This contains the methylation values from benign prostate tissue collected from patient 3.

UCSC_RefGene_Name Target gene name(s), from the UCSC database. *Note: multiple listings of the same gene name indicate splice variants

UCSC_CpG_Islands_Name Chromosomal coordinates of the CpG Island from UCSC.

See Also

`pca.methylation.data`

pca.methylation.data *DNA methylation dataset of patients suffering from prostate cancer disease.*

Description

The dataset contains pre-processed beta methylation values of R=2 samples which are collected from N=4 patients suffering from prostate cancer disease.

Usage

```
data(pca.methylation.data)
```

Format

A data frame with 694820 rows and 9 columns. The data contains no missing values.

IlmnID This contains the Unique identifier from the Illumina CG database. (The probe ID).

Patient_benign_1 This contains the methylation values from benign prostate tissue collected from patient 1.

Patient_benign_2 This contains the methylation values from benign prostate tissue collected from patient 2.

Patient_benign_3 This contains the methylation values from benign prostate tissue collected from patient 3.

Patient_benign_4 This contains the methylation values from benign prostate tissue collected from patient 4.

Patient_benign_1 This contains the methylation values from tumor prostate tissue collected from patient 1.

Patient_benign_2 This contains the methylation values from tumor prostate tissue collected from patient 2.

Patient_benign_3 This contains the methylation values from tumor prostate tissue collected from patient 3.

Patient_benign_4 This contains the methylation values from tumor prostate tissue collected from patient 4.

Details

The raw methylation array data was first quality controlled and preprocessed using RnBeads package. The array data was then normalized and 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.

See Also

`legacy.data`

`plot.betaclust` *Plots for visualizing the betaclust class object*

Description

This function helps visualise the clustering solution by plotting the density estimates, the uncertainty and the information criterion.

Usage

```
## S3 method for class 'betaclust'
plot(object, what = "density", plot_type = "ggplot", scale_param = "free_y")
```

Arguments

`object`	betaclust object
`what`	The different plots that can be obtained from the object (default="density") (what=c("density","uncertainty","InformationCriterion"))
`plot_type`	The plot type to be displayed (default="ggplot")(plot_type="ggplot" or"plotly")
`scale_param`	The axis that needs to be fixed or not for facet plot (scales=c("free_y","free_x","free"), default is "free_y")

Details

The density estimates of the clustering solution of the optimal model can be plotted by passing the parameter what="density" in the function. Apart from static plots interactive plots can also be plotted using the parameter plot_type = "plotly". The uncertainty in the clustering soluting can be plotted using what="uncertainty". The information criterion values for all models can be plotted using what="InformationCriterion" for selecting the optimal model.

See Also

`betaclust`

`summary.betaclust` *Summarizing the Beta Mixture Model Fits*

Description

Summary method for class "betaclust" object containing the results of the optimal model selected.

Usage

```
## S3 method for class 'betaclust'
summary(object)
```

Arguments

`x`	betaclust object

Value

An object of class "summary.betaclust". The object returns the following list of values:

- CpG_sites - The number of CpG sites analysed using the beta mixture models.
- patients - The number of patients analysed using the beta mixture models.
- samples - The numder of samples analysed using the beta mixture models.
- cluster_count - The number of groups, the data is clustered into.
- modelName - The optimal model selected.
- loglik - The log-likelihood value for the selected optimal model.
- Information_criterion - The information criterion used to select the optimal model.
- ic_output - This stores the information criterion value calculated for each model.
- classification - The total number of CpG sites identified in each cluster.

See Also

`betaclust`

Examples

```
## Not run:
data_output=betaclust(pca.methylation.data[,2:9],K,patients,samples,
            model_names=c("C..","CN.","C.R"),model_selection="BIC",seed=my.seed)
summary(data_output)
## End(Not run)
```