Package 'betaclust'

June 17, 2022

Type Package

Title A family of mixture models for clustering beta valued DNA methylation data.

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) to appositely model beta valued DNA methylation data, to objectively identify methylation state thresholds and to 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 methylation profiles in beta valued DNA methylation data.

Usage

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

Arguments

`data`	Methylation values for C CpG sites from R samples collected from N patients.
`K`	Number of methylation profiles to be identified.
`patients`	Number of patients in the study.
`samples`	Number of samples collected from each patient for study.
`model_names`	Models to run from the set of models, C.., CN. and C.R, default = C.. . See details.
`model_selection`	Information criterion used for model selection. Options are AIC/BIC/ICL/default=BIC.
`seed`	Seed to allow for reproducibility.
`register`	Setting for registering the parallel backend with the "foreach" package. To start parallel execution of R code on machine with multiple cores, "NULL" value needs to be assigned to this parameter.

Details

This is a wrapper function which can be used to fit all three models (C.., CN., C.R) together. The C.. and CN. models are used to analyse a single DNA sample ($R = 1$) and cluster the C CpG sites into the K methylation profiles. As each CpG site can belong to either of the $M = 3$ methylation profiles (hypomethylation, hemimethylation and hypermethylation), the default value for $K = M = 3$. The thresholds between methylation profiles can be objectively identified from the clustering solution. The C.R model is used to analyse R independent samples collected from N patients, where each sample contains C CpG sites, and cluster the dataset into $K = M^R$ clusters to identify the differentially methylated CpG sites between the R DNA samples.

Value

The function returns an object of "betaclust" class which contains the 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.
- C - the number of CpG sites analysed using the beta mixture models.
- N - the number of patients analysed using the beta mixture models.
- R - the number of samples analysed using the beta mixture models.
- optimal_model_results - this contains information from the optimal model. Specifically,
  - cluster_size - the total number of CpG sites identified in each cluster.
  - llk - a vector containing the log-likelihood value at each step of the EM algorithm.
  - data - this contains the methylation dataset along with the cluster label for each CpG site.
  - alpha - this contains the shape parameter 1 for the beta mixture model.
  - delta - this contains the shape parameter 2 for the beta mixture model.
  - tau - the proportion of CpG sites in each cluster.
  - z - a matrix containing the probability for each CpG site of belonging to each of the K clusters.
  - uncertainty - the uncertainty of each CpG site's clustering.

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.

Microsoft, Weston, S. (2022): foreach: Provides Foreach Looping Construct. R package version 1.5.2. https://CRAN.R-project.org/package=foreach.

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

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

Usage

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

Arguments

data	Methylation values for C CpG sites from $R = 1$ samples collected from N patients.
K	Number of methylation profiles to be identified.
seed	Seed to allow for reproducibility.
register	Setting for registering the parallel backend with the "foreach" package. To start parallel execution of R code on machine with multiple cores, "NULL" value needs to be assigned to this parameter.

Details

This model clusters each of the C CpG sites into one of K methylation profiles, based on data from N patients for one DNA sample (i.e. $R = 1$). As each CpG site can belong to either of the $M = 3$ methylation profiles (hypomethylated, hemimethylated or hypermethylated), the default value of $K = M = 3$. Under the C.. model the shape parameters of each cluster are constrained to be equal for each patient.

Value

A list containing:

- cluster_size - the total number of CpG sites identified in each cluster.
- llk - a vector containing the log-likelihood value at each step of the EM algorithm.
- data - this contains the methylation dataset along with the cluster label for each CpG site.
- alpha - this contains the shape parameter 1 for the beta mixture model.
- delta - this contains the shape parameter 2 for the beta mixture model.
- tau - the proportion of CpG sites in each cluster.
- z - a matrix containing the probability for each CpG site of belonging to each of the K clusters.
- uncertainty - the uncertainty of each CpG site's clustering.

References

Microsoft, Weston, S. (2022): foreach: Provides Foreach Looping Construct. R package version 1.5.2. https://CRAN.R-project.org/package=foreach.

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

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

Usage

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

Arguments

data	Methylation values for C CpG sites from $R = 1$ samples collected from N patients.
K	Number of methylation profiles to be identified.
seed	Seed to allow for reproducibility.
register	Setting for registering the parallel backend with the 'foreach' package. To start parallel execution of R code on machine with multiple cores, 'NULL' value needs to be assigned to this parameter.

Details

This model clusters each of the C CpG sites into one of K methylation profiles, based on data from N patients for one DNA sample (i.e. $R = 1$). As each CpG site can belong to either of the $M = 3$ methylation profiles (hypomethylated, hemimethylated or hypermethylated), the default value of $K = M = 3$. The CN. model differs from the C.. model as it is less parsimonious, allowing cluster and patient-specific shape parameters.

Value

A list containing:

- cluster_size - the total number of CpG sites identified in each cluster.
- llk - a vector containing the log-likelihood value at each step of the EM algorithm.
- data - this contains the methylation dataset along with the cluster label for each CpG site.
- alpha - this contains the shape parameter 1 for the beta mixture model.
- delta - this contains the shape parameter 2 for the mixture model.
- tau - the proportion of CpG sites in each cluster.
- z - a matrix containing the probability for each CpG site of belonging to each of the K clusters.
- uncertainty - the uncertainty of each CpG site's clustering.

References

Microsoft, Weston, S. (2022): foreach: Provides Foreach Looping Construct. R package version 1.5.2. https://CRAN.R-project.org/package=foreach.

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

A 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

data	Methylation values for C CpG sites from R samples collected from N patients.
K	Number of methylation profiles to be identified.
patients	Number of patients in the study.
samples	Number of samples collected from each patient for study.
seed	Seed to allow for reproducibility.
register	Setting for registering the parallel backend with the "foreach" package. To start parallel execution of R code on machine with multiple cores, "NULL" value needs to be assigned to this parameter.

Details

The C.R model allows identification of the differentially methylated CpG sites between the R DNA samples collected from each of N patients. As each CpG site in a DNA sample can belong to either of M methylation profiles, there can be $K = M^R$ methylation profile changes between R DNA samples. The parameters vary for each DNA sample but are constrained to be equal for each patient. An initial clustering using K-means is performed to identify K clusters. The resulting clustering solution is 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 a computationally inefficient numerical optimisation step.

Value

A list containing:

- cluster_size - the total number of CpG sites identified in each cluster.
- llk - a vector containing the log-likelihood value at each step of the EM algorithm.
- data - this contains the methylation dataset along with the cluster label for each CpG site.
- alpha - this contains the shape parameter 1 for the beta mixture model.
- delta - this contains the shape parameter 2 for the beta mixture model.
- tau - the proportion of CpG sites in each cluster.
- z - a matrix containing the probability for each CpG site of belonging to each of the K clusters.
- uncertainty - the uncertainty of each CpG site's clustering.

References

Microsoft, Weston, S. (2022): foreach: Provides Foreach Looping Construct. R package version 1.5.2. https://CRAN.R-project.org/package=foreach.

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

An empirical cumulative distribution function (ECDF) plot for a betaclust object.

Usage

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

Arguments

`x`	A dataframe containing 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_Patient1, Sample1_Patient2, Sample2_Patient1, Sample2_Patient2.
`samples`	number of tissue samples from which DNA methylation data are collected (default samples = 2).
`sample_name`	The order in which the samples are grouped in the dataframe x (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 visualises the methylation profile changes between the different DNA samples for each patient.

Value

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

See Also

`betaclust`

`beta_cr`

em_aic *Akaike Information Criterion*

Description

Compute the AIC value for the optimal model.

Usage

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

Arguments

`llk`	log-likelihood value.
`C`	number of CpG sites.
`K`	number of methylation profiles identified.
`patients`	number of patients.
`samples`	number of DNA samples collected from each patient.
`model_name`	fitted mixture model (method=c("C..","CN.","C.R")).

Details

Computes the AIC for a specified model given the log-likelihood, the dimension of the data, and the model specification.

Value

The AIC value for the selected model.

See Also

`em_bic`

`em_icl`

em_bic *Bayesian Information Criterion*

Description

Compute the BIC value for the optimal model.

Usage

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

Arguments

`llk`	log-likelihood value.
`C`	number of CpG sites.
`K`	number of methylation profiles identified.
`patients`	number of patients.
`samples`	number of DNA samples collected from each patient.
`model_name`	fitted mixture model (method=c("C..","CN.","C.R")).

Details

Computes the BIC for a specified model given the log-likelihood, the dimension of the data, and the model specification.

Value

The BIC value for the selected model.

See Also

`em_aic`

`em_icl`

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

Description

Compute the ICL value for the optimal model.

Usage

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

Arguments

`llk`	log-likelihood value.
`C`	number of CpG sites.
`K`	number of methylation profiles identified.
`patients`	number of patients.
`samples`	number of DNA samples collected from each patient.
`model_name`	fitted mixture model (method=c("C..","CN.","C.R")).
`z`	z matrix used for computing the complete-data log-likelihood function.

Details

Computes the ICL for a specified model given the log-likelihood, the dimension of the data, and the model specification. 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: the unique identifier from the Illumina CG database, i.e. the probe ID.
- Genome_Build: the genome build referenced by the Infinium MethylationEPIC manifest.
- CHR: the chromosome containing the CpG (Genome_Build = 37).
- MAPINFO: the methylation values from benign prostate tissue collected from patient 3.
- UCSC_RefGene_Name: the target gene name(s), from the UCSC database. Note: multiple listings of the same gene name indicate splice variants.
- UCSC_CpG_Islands_Name: the 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 from $R = 2$ sample, 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: the unique identifier from the Illumina CG database, i.e. the probe ID.
- Patient_benign_1: the methylation values from benign prostate tissue collected from patient 1.
- Patient_benign_2: the methylation values from benign prostate tissue collected from patient 2.
- Patient_benign_3: the methylation values from benign prostate tissue collected from patient 3.
- Patient_benign_4: the methylation values from benign prostate tissue collected from patient 4.
- Patient_benign_1: the methylation values from tumor prostate tissue collected from patient 1.
- Patient_benign_2: the methylation values from tumor prostate tissue collected from patient 2.
- Patient_benign_3: the methylation values from tumor prostate tissue collected from patient 3.
- Patient_benign_4: the methylation values from tumor prostate tissue collected from patient 4.

Details

The raw methylation array data was first quality controlled and preprocessed using the 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.

References

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, https://rnbeads.org.

Assenov Y, Mueller F, Lutsik P, Walter J, Lengauer T, Bock C (2014). "Compehensive Analysis of DNA Methylation Data with RnBeads." Nature Methods, 11(11), 1138–1140. doi: 10.1038/nmeth.3115, https://rnbeads.org.

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",
  title = NULL,
  scale_param = "free_y"
)
```

Arguments

object	A betaclust object.
what	The different plots that can be obtained are either "density","uncertainty" or "InformationCriterion". (default="density").
plot_type	The plot type to be displayed are either "ggplot" or "plotly". (default="ggplot").
title	The title that the user wants to display on the graph. If no title is to be displayed the default is "NULL" value.
scale_param	The axis that needs to be fixed for density estimates plot for visualizing the C.R clustering solution are either "free_y","free_x" or "free". (default = "free_y").

Details

The density estimates under the optimal clustering solution by specifying what = "density" in the function. Interactive plots can also be produced using plot_type = "plotly". The uncertainty in the clustering solution can be plotted using what="uncertainty". The information criterion values for all fitted models can be plotted using what = "InformationCriterion".

See Also

betaclust

---

summary.betaclust *Summarizing the beta mixture model fits*

Description

Summary method for a "betaclust" object containing the results under the optimal model selected.

Usage

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

Arguments

object	A betaclust object.

Value

An object of class "summary.betaclust which contains the following list of values:

- C - the number of CpG sites analysed using the beta mixture models.
- N - the number of patients analysed using the beta mixture models.
- R - the number of samples analysed using the beta mixture models.
- K - the number of methylation profiles identified.
- 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.
- prop_data - the proportion 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)
```