Package 'mSigHdp'

October 14, 2022

Title Mutational signature discovery using HDP (hierarchical Dirichlet process)

Version 2.1.2

Description Mutational signature discovery using hierarchical Dirichlet process mixture modeling. mSigHdp stands for 'mutational signature (discovery using) hierarchical Dirichlet processes'. This package uses https://github.com/steverozen/hdpx for the hierarchical Dirichlet process implementation. Most users will start with the function RunHdpxParallel. Please see the vignette for an example. Please also see our paper: Mo Liu, Yang Wu, Nanhai Jiang, Arnoud Boot, Steven G. Rozen, mSigHdp: hierarchical Dirichlet process mixture modeling for mutational signature discovery, https://www.biorxiv.org/content/10.1101/2022.01.31.478587v1. Only supported on Linux systems.

License GPL-3

Encoding UTF-8

Language en-US

BuildManual no

biocViews

Roxygen list(markdown = TRUE)

Depends R (>= 4.0)

RoxygenNote 7.2.1

Remotes github::steverozen/hdpx@master,
github::steverozen/ICAMS@*release

Imports data.table,
hdpx (>= 1.0.5),
ICAMS (>= 2.2.4),
mSigTools,
reshape2

Suggests cosmicsig,
knitr,
rmarkdown,
testthat,
utils

VignetteBuilder knitr

R topics documented:

`Burnin` *Run the Gibbs sampling burnin (one thread)*

Description

Run the Gibbs sampling burnin (one thread)

Usage

```
Burnin(
  hdp.state,
  seedNumber = 1,
  burnin = 5000,
  cpiter = 3,
  burnin.verbosity = 0,
  burnin.multiplier = 2,
  checkpoint = TRUE
)
```

Arguments

`hdp.state`	An `hdpState-class` object or a list representation of an `hdpState-class` object.
`seedNumber`	Set the random seed to this value.
`burnin`	The number of burn-in iterations in one batch. The total number of burnin iterations is `burnin * burnin.multiplier`.
`cpiter`	The number of iterations of concentration parameter sampling to perform after each main Gibbs-sample iteration. (See Teh et al., "Hierarchical Dirichlet Processes", Journal of the American Statistical Association 2006;101(476):1566-1581 (https://doi.org/10.1198/016214506000000302).)
`burnin.verbosity`	Verbosity of message statements.
`burnin.multiplier`	Run `burnin.multiplier` rounds of `burnin` iterations. If `checkpoint` is `TRUE`, save the burnin chain (see parameter `checkpoint`.) The diagnostic plot `diagnostics.likelihood.pdf` can help determine if the chain is stationary. The burnin can be continued from a checkpoint file with `ExtendBurnin` (see argument `checkpoint`).

`checkpoint` If `TRUE`, create a checkpoint file called mSigHdp.burnin.checkpoint.*seedNumber*.Rdata in the current working directory.

Value

A list with 2 elements:

`hdplist` A list representation of an `hdpState-class` object.

`likelihood` A numeric vector with the likelihood at each iteration. This is the same type as returned from `link[hdp]{hdp_burnin}` in package hdpx.

CombineChainsAndExtractSigs

Extract signatures etc. from multiple Gibbs sample chains

Description

Extract signatures etc. from multiple Gibbs sample chains

Usage

```
CombineChainsAndExtractSigs(
  clean.chlist,
  input.catalog,
  verbose = FALSE,
  high.confidence.prop = 0.9,
  merge.raw.cluster.args = hdpx::default_merge_raw_cluster_args()
)
```

Arguments

`clean.chlist` A list of `hdpSampleChain-class` S4 objects, each with the information from one Gibbs sampling chain. See `hdpSampleChain-class` in package hdpx.

`input.catalog` Input spectra catalog as a matrix.

`verbose` If `TRUE` then `message` progress information.

`high.confidence.prop` Raw clusters of mutations found in $>=$ `high.confidence.prop` proportion of posterior samples are signatures with high confidence.

`merge.raw.cluster.args` See `default_merge_raw_cluster_args` in package `hdpx`.

Value

Invisibly, a list with the following elements:

signature The extracted signature profiles as a matrix; rows are mutation types, columns are signatures with high confidence.

signature.post.samp.number A data frame with two columns. The first column corresponds to each signature in `signature` and the second columns contains the number of posterior samples that found the raw clusters contributing to the signature.

signature.cdc A numeric data frame. Columns correspond to signatures as in `signature`. Rows correspond to either biological samples or to parent and grandparent Dirichlet processes.

exposureProbs The inferred exposures as a matrix of mutation probabilities; rows are signatures, columns are samples (e.g. tumors). This is similar to `signature.cdc`, but every column was normalized to sum to 1.

low.confidence.signature The profiles of signatures extracted with low confidence as a matrix; rows are mutation types, columns are signatures with < `high.confidence.prop` of posterior samples.

low.confidence.post.samp.number Analogous to `signature.post.samp.number`, except that this one is for signatures in `low.confidence.signature`.

low.confidence.cdc Analogous to `signature.cdc`, except that columns in this matrix correspond to columns in `low.confidence.signature`.

extracted.retval A list object returned from `extract_components` in package hdpx.

`downsample` *Downsample a vector of type* `numeric`*; nonsensical for negative values.*

Description

Downsample a vector of type `numeric`; nonsensical for negative values.

Usage

```
downsample(x, thres = NULL, downsample_threshold = 3000)
```

Arguments

`x`	A `numeric` vector.
`thres`	Deprecated alternative to `downsample_threshold`.
`downsample_threshold`	If `downsample_threshold` >= 3,000, then all input values > `downsample_threshold` are downsampled. If `downsample_threshold` < 3,000, only some input values > `downsample_threshold` are downsampled but all input values > 3000 are downsampled. See `show_downsample_curves`.

Value

A vector of integers (type `numeric`) of the same length as `x`, downsampled as described in the documentation for the `downsample_threshold` argument.

downsample_spectra *Downsample a set of mutational spectra*

Description

Downsample a set of mutational spectra

Usage

```
downsample_spectra(spec, downsample_threshold = NULL, thres = NULL)
```

Arguments

spec	Input spectra as a numerical matrix or similar data.frame; each column is a spectrum, each row is a mutation type (e.g. CAG -> CTG).
downsample_threshold	See downsample and show_downsample_curves.
thres	Deprecated alternative to downsample_threshold.

Value

A list with elements:

- down_spec: A numeric matrix with same shape as spec, with the mutation counts in each column reduced based on the corresponding ratio in the second element of this list, down_factor.
- down_factor Numeric vector of the ratios of downsample(colSums(spec)) to colSums(spec).

ExtendBurnin *Extend burnin iterations generated from* Burnin

Description

Extend burnin iterations generated from Burnin

Usage

```
ExtendBurnin(
  previous.burnin.output,
  burnin = 5000,
  cpiter = 3,
  burnin.verbosity = 0,
  seedNumber = NULL
)
```

Arguments

previous.burnin.output	Output from Burnin or the file path of a checkpoint file written by Burnin.
burnin	The number of burnin iterations to perform.
cpiter	The number of iterations of concentration parameter sampling to perform after each main Gibbs-sample iteration. (See Teh et al., "Hierarchical Dirichlet Processes", Journal of the American Statistical Association 2006;101(476):1566-1581 (https://doi.org/10.1198/016214506000000302).)
burnin.verbosity	Number that controls whether progress messages are printed.
seedNumber	A random seed for reproducible results.

Value

The same type of object as returned from Burnin.

The envisioned application is extending burnins from burnin checkpoints.

GibbsSamplingAfterBurnin

Start Gibbs sampling on one chain after burnin

Description

This function might be used to start Gibbs sampling after ExtendBurnin.

Usage

```
GibbsSamplingAfterBurnin(
  burnin.output,
  post.n,
  post.space,
  post.cpiter = 3,
  post.verbosity = 0,
  seedNumber = NULL
)
```

Arguments

burnin.output	A path to burnin checkpoint Rdata or to an S4 object from Burnin.
post.n	The number of posterior samples to collect.
post.space	The number of iterations between collected samples.
post.cpiter	The number of iterations of concentration parameter sampling to perform after each main Gibbs-sample iteration. (See Teh et al., "Hierarchical Dirichlet Processes", Journal of the American Statistical Association 2006;101(476):1566-1581 (https://doi.org/10.1198/016214506000000302).)
post.verbosity	Verbosity of debugging statements. No need to change unless for testing or debugging.
seedNumber	A random seed that ensures reproducible results.

Value

An hdpSampleChain S4 object with the salient information from each posterior sample. See hdpSampleChain-class in package hdpx.

RunHdpxParallel *Extract (discover) mutational signatures from a matrix of mutational spectra*

Description

Please see the vignette for an example.

Usage

```
RunHdpxParallel(
  input.catalog,
  seedNumber = 123,
  K.guess,
  multi.types = FALSE,
  verbose = FALSE,
  burnin = 1000,
  burnin.multiplier = 10,
  post.n = 200,
  post.space = 100,
  post.cpiter = 3,
  post.verbosity = 0,
  CPU.cores = 20,
  num.child.process = 20,
  high.confidence.prop = 0.9,
  hc.cutoff = NULL,
  merge.raw.cluster.args = hdpx::default_merge_raw_cluster_args(),
  overwrite = TRUE,
  out.dir = paste0("./RunHdpxParallel_out_", as.numeric(Sys.time())),
  gamma.alpha = 1,
  gamma.beta = 20,
  checkpoint = TRUE,
  downsample_threshold = NULL
)
```

Arguments

input.catalog	Input spectra catalog as a matrix or in ICAMS format.
seedNumber	A random seed that ensures ensures reproducible results.
K.guess	Suggested initial value of the number of raw clusters. Usually, the number of raw clusters is roughly twice the number of extracted signatures. Passed to hdpx::dp_activate as argument initcc.

multi.types
: A logical scalar or a character vector.
If FALSE, The HDP analysis will regard all input spectra as one tumor type, and the HDP structure will have one parent node for all tumors.
If TRUE, Sample IDs in input.catalog must have the form *sample_type*::*sample_id*.
If a character vector, then its length must be ncol(input.catalog), and each value is the sample type of the corresponding column in input.catalog, e.g. c(rep("Type-A", 23), rep("Type-B", 10)) for 23 Type-A samples and 10 Type-B samples.
If not FALSE, HDP will have one parent node for each sample type and one grandparent node.

verbose
: If TRUE then message progress information.

burnin
: The number of burn-in iterations in one batch. The total number of burnin iterations is burnin * burnin.multiplier.

burnin.multiplier
: Run burnin.multiplier rounds of burnin iterations. If checkpoint is TRUE, save the burnin chain (see parameter checkpoint.) The diagnostic plot diagnostics.likelihood.pdf can help determine if the chain is stationary. The burnin can be continued from a checkpoint file with ExtendBurnin (see argument checkpoint).

post.n
: The number of posterior samples to collect.

post.space
: The number of iterations between collected samples.

post.cpiter
: The number of iterations of concentration parameter samplings to perform after each iteration.

post.verbosity
: Verbosity of debugging statements. No need to change except for development purposes.

CPU.cores
: Number of CPUs to use; this should be no more than num.child.process.

num.child.process
: Number of posterior sampling chains; can set to 1 for testing. We recommend 20 for real data analysis

high.confidence.prop
: Raw clusters of mutations found in $>=$ high.confidence.prop proportion of posterior samples are signatures with high confidence.

hc.cutoff
: Deprecated, use merge.raw.cluster.args.

merge.raw.cluster.args
: See default_merge_raw_cluster_args in package hdpx.

overwrite
: If TRUE overwrite out.dir if it exists, otherwise raise an error.

out.dir
: If not NULL then a character string specifying a directory that will be created for the output, including csv files and plots (pdfs) of extracted signatures and their exposures. If NULL no directory will be created and no files will be generated.

gamma.alpha
: Shape parameter of the gamma distribution prior for the Dirichlet process concentration parameters α_0 and all α_j in Figure B.1 of

  - https://www.repository.cam.ac.uk/bitstream/handle/1810/275454/Roberts-2018-PhD.pdf

gamma.beta
: Inverse scale parameter (rate parameter) of the gamma distribution prior for the Dirichlet process concentration parameters: β_0 and all β_j in Figure B.1 of

  - https://www.repository.cam.ac.uk/bitstream/handle/1810/275454/Roberts-2018-PhD.pdf

We recommend gamma.alpha = 1 and gamma.beta = 20 for single-base-substitution signature extraction; gamma.alpha = 1 and gamma.beta = 50 for doublet-base-substitution and indel signature extraction

`checkpoint` If `TRUE`, then

- Checkpoint each final Gibbs sample chain to the current working directory, in a file called mSigHdp.sample.checkpoint.*x*.Rdata, where *x* depends on `seedNumber`.
- Periodically checkpoint the burnin state to the current working directory, in files called mSigHdp.burnin.checkpoint.*x*.Rdata, where *x* depends on the `seedNumber`.

`downsample_threshold`
See `downsample_spectra` and `link{show_downsample_curves}`.

Details

Please see our paper at https://www.biorxiv.org/content/10.1101/2022.01.31.478587v1 for suggestions on argument values to use.

Value

Invisibly, a list with the following elements:

signature The extracted signature profiles as a matrix; rows are mutation types, columns are signatures with high confidence.

signature.post.samp.number A data frame with two columns. The first column corresponds to each signature in `signature` and the second columns contains the number of posterior samples that found the raw clusters contributing to the signature.

signature.cdc A numeric data frame. Columns correspond to signatures as in `signature`. Rows correspond to either biological samples or to parent and grandparent Dirichlet processes.

exposureProbs The inferred exposures as a matrix of mutation probabilities; rows are signatures, columns are samples (e.g. tumors). This is similar to `signature.cdc`, but every column was normalized to sum to 1.

low.confidence.signature The profiles of signatures extracted with low confidence as a matrix; rows are mutation types, columns are signatures with < `high.confidence.prop` of posterior samples.

low.confidence.post.samp.number Analogous to `signature.post.samp.number`, except that this one is for signatures in `low.confidence.signature`.

low.confidence.cdc Analogous to `signature.cdc`, except that columns in this matrix correspond to columns in `low.confidence.signature`.

extracted.retval A list object returned from `extract_components` in package hdpx.

`show_downsample_curves` *Show the effects of the downsample_threshold argument to* `RunHdpxParallel`.

Description

Show the effects of the downsample_threshold argument to `RunHdpxParallel`.

Usage

```
show_downsample_curves(...)
```

Arguments

`...`	One or more positive numbers to use as possible values for `downsample_threshold`.

Value

Called for side effects.

Examples

```
show_downsample_curves(3000, 5000, 10000)
```

Index