

Package ‘mSigHdp’

August 22, 2022

Title Mutational signature discovery using HDP (hierarchical Dirichlet process)

Version 2.1.0

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@v1.0.5-branch,
github::steverozen/ICAMS@v3.0.6-branch,
github::Rozen-Lab/mSigTools@v1.0.0-branch

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	2
CombineChainsAndExtractSigs	3
downsample	4
downsample_spectra	4
ExtendBurnin	5
GibbsSamplingAfterBurnin	6
RunHdpParallel	6

Index	10
--------------	-----------

Burnin	<i>Run the Gibbs sampling burnin (one thread)</i>
--------	---

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 <code>burnin * burnin.multiplier</code> .
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 <code>burnin.multiplier</code> rounds of <code>burnin</code> iterations. If <code>checkpoint</code> is TRUE, save the burnin chain (see parameter <code>checkpoint</code> .) The diagnostic plot <code>diagnostics.likelihood.pdf</code> can help determine if the chain is stationary. The burnin can be continued from a checkpoint file with ExtendBurnin (see argument <code>checkpoint</code>).
checkpoint	If TRUE, create a checkpoint file called <code>mSigHdp.burnin.checkpoint.seedNumber.Rdata</code> 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 `hdp`.

`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 = hdp::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 `hdp`.

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

`verbose` If TRUE then message progress information.

`high.confidence.prop` Raw clusters of mutations found in $\geq \text{high.confidence.prop}$ proportion of posterior samples are signatures with high confidence.

`merge.raw.cluster.args` See `default_merge_raw_cluster_args`.

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`.

<code>downsample</code>	<i>Down sample a vector of type numeric; nonsensical for negative values.</i>
-------------------------	---

Description

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

Usage

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

Arguments

<code>x</code>	A numeric vector.
<code>thres</code>	If <code>thres >= 3,000</code> , then in all input values <code>> thres</code> are downsampled. If <code>thres < 3,000</code> , some input values <code>> thres</code> are downsampled.

Value

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

<code>downsample_spectra</code>	<i>Down sample a set of mutational spectra</i>
---------------------------------	--

Description

Down sample a set of mutational spectra

Usage

```
downsample_spectra(spec, thres)
```

Arguments

spec	Input spectra as a numerical matrix or similar <code>data.frame</code> ; each column is a spectrum, each row is a mutation type (e.g. CAG -> CTG).
thres	See downsample .

Value

A numeric matrix with same shape as `spec`, with the entries each column reduced based on the ratio of `downsample(colSums(spec))` to `colSums(spec)`.

ExtendBurnin	<i>Extend burnin iterations generated from Burnin</i>
--------------	---

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

```
RunHdpParallel(
  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 = hdp::default_merge_raw_cluster_args(),
  overwrite = TRUE,
  out.dir = NULL,
  gamma.alpha = 1,
  gamma.beta = 20,
  checkpoint = TRUE,
  downsample_threshold = NULL
)
```

Arguments

<code>input.catalog</code>	Input spectra catalog as a matrix or in ICAMS format.
<code>seedNumber</code>	A random seed that ensures ensures reproducible results.
<code>K.guess</code>	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 <code>hdp::dp_activate</code> as argument <code>initcc</code> .
<code>multi.types</code>	<p>A logical scalar or a character vector.</p> <p>If <code>FALSE</code>, The HDP analysis will regard all input spectra as one tumor type, and the HDP structure will have one parent node for all tumors.</p> <p>If <code>TRUE</code>, Sample IDs in <code>input.catalog</code> must have the form <code>sample_type::sample_id</code>.</p> <p>If a character vector, then its length must be <code>ncol(input.catalog)</code>, and each value is the sample type of the corresponding column in <code>input.catalog</code>, e.g. <code>c(rep("Type-A", 23), rep("Type-B", 10))</code> for 23 Type-A samples and 10 Type-B samples.</p> <p>If not <code>FALSE</code>, HDP will have one parent node for each sample type and one grandparent node.</p>
<code>verbose</code>	If <code>TRUE</code> then message progress information.
<code>burnin</code>	The number of burn-in iterations in one batch. The total number of burnin iterations is <code>burnin * burnin.multiplier</code> .
<code>burnin.multiplier</code>	Run <code>burnin.multiplier</code> rounds of burnin iterations. If <code>checkpoint</code> is <code>TRUE</code> , save the burnin chain (see parameter <code>checkpoint</code> .) 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 \geq 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 .
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 <ul style="list-style-type: none"> 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 <ul style="list-style-type: none"> https://www.repository.cam.ac.uk/bitstream/handle/1810/275454/Roberts-2018-PhD.pdf <p>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</p>
checkpoint	If TRUE, then <ul style="list-style-type: none"> 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 .

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 column 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`.

Index

Burnin, [2](#), [5](#), [6](#)

CombineChainsAndExtractSigs, [3](#)

default_merge_raw_cluster_args,
[3](#), [8](#)

downsample, [4](#), [5](#)

downsample_spectra, [4](#), [8](#)

ExtendBurnin, [2](#), [5](#), [6](#), [8](#)

extract_components, [4](#), [9](#)

GibbsSamplingAfterBurnin, [6](#)

ICAMS, [7](#)

RunHdpxParallel, [6](#)