Package 'BayRepulsive'

October 7, 2018

Type Package

Title BayRepulsive: A Bayesian Repulsive Deconvolution Model

Version 0.1.0

Date 2018-10-07

Depends mvtnorm, alabama, psych, optimx

Author Yuliang Li and Yanxun Xu

Maintainer Yuliang Li <yli193@jhu.edu>

Description A Non-negative Matrix Factorization (NMF) with repulsiveness introduced by determinantal point process (DPP).

License JHU

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

R topics documented:

`BayRepulsive_known` *BayRepulsive_known is a deconvolution function designed for inferring tumor heterogeneity, used when the number of subclones is known.*

Description

Takes in the observed data matrix, the number of subclones, the number of features and samples, gives the estiamted NMF results.

Usage

```
BayRepulsive_known(Datause, K, Nobs, Nfeature,
                   Niter = 100, epsilon = 0.0001, tau = 100, seed = 1 )
```

Arguments

Datause	The observed data matrix. Each row is a sample.
K	The number of subclones
Nobs	The number of samples, i.e., the number of rows of the Datause
Nfeature	The number of features, i.e., the number of columns of the Datause
Niter	The number of maximum iterations
epsilon	break if the L2 distance of the two estiamted proportion matrix in row is less than epsilon
tau	The hyperparameter for DPP, a large number is prefered, default value is 100
seed	The random seed, default as 1

Details

Given an observed matrix, whose rows are mixed samples of unknown number of subclones, returns the results of NMF.

This function will create a bunch of globel variables, named `Datause`, `Nobs`, `Nfeature`, `sigma0`, `mu0`, `K`, `Theta`, `W.star`, `Z.star`, `W_temp`, `sigma.square`, `data.now`, `Z.now`, `i`. Thus, users should avoid these variable names when using BayRepulsive_unknown, if they don't want the variables to be overwritten. Especially `i`, which is commonly used in loops.

Value

`W` the estiamted signature matrix.

`Z` the estiamted proportion matrix.

`C` sum of estimated square error used as measure of performance for deconvolution.

Source

BayRepulsive: A Bayesian Repulsive Deconvolution Model for Inferring Tumor Heterogeneity

Examples

```
rm(list=ls())
library(BayRepulsive)
data(CCLE)
set.seed(1)
Nobs     <- dim(CCLE$DATA)[1]
Nfeature <- dim(CCLE$DATA)[2]
error    <- matrix(rnorm(Nobs * Nfeature, mean = 0, sd = 0.1), nrow = Nobs)
DATA     <- CCLE$DATA + error
DATA     <- pmax(DATA,0)
result1  <- BayRepulsive_known(Datause = DATA, K = 3, Nobs = Nobs,
                               Nfeature = Nfeature)
cor(as.vector(result1$W), as.vector(CCLE$W))

#----------------
rm(list=ls())
library(BayRepulsive)
data(Inhouse)
Nobs      <- dim(Inhouse$DATA)[1]
Nfeature  <- dim(Inhouse$DATA)[2]
```

```
result1   <- BayRepulsive_known(Datause = Inhouse$DATA, K=3, Nobs = Nobs,
                                Nfeature = Nfeature, seed = 12)
# handle the label swithing issue
W_est     <- result1$W
W_est[,1] <- result1$W[,2]
W_est[,2] <- result1$W[,1]
cor(as.vector(W_est), as.vector(Inhouse$W))
```

`BayRepulsive_unknown` *BayRepulsive_unknown is a deconvolution function designed for inferring tumor heterogeneity, used when the number of subclones is unknown.*

Description

Takes in the observed data matrix, the range of number of subclones, the number of features and samples, gives the estimation of the NMF, including the estimated number of subclones.

Usage

```
BayRepulsive_unknown(Datause, K_min, K_max, Nobs, Nfeature,
                     Niter = 100, epsilon = 0.0001,tau = 100, seed = 1 )
```

Arguments

`Datause`	The observed data matrix. Each row is a sample.
`K_min`	The minimum number of subclones
`K_max`	The Maximum number of subclones
`Nobs`	The number of samples, i.e., the number of rows of the Datause
`Nfeature`	The number of features, i.e., the number of columns of the Datause
`Niter`	The number of maximum iterations
`epsilon`	Break if the L2 distance of the two estiamted proportion matrix in row is less than epsilon
`tau`	The hyperparameter for DPP, a large number is prefered, default value is 100
`seed`	The random seed, default as 1

Details

Given an observed matrix, whose rows are mixed samples of unknown number of subclones, we give an estimation of number of subclones along with NMF results.

This function will create a bunch of globel variables, named `Datause`, `Nobs`, `Nfeature`, `sigma0`, `mu0`, `K`, `Theta`, `W.star`, `Z.star`, `W_temp`, `sigma.square`, `data.now`, `Z.now`, `i`. Thus, users should avoid these variable names when using BayRepulsive_unknown, if they don't want the variables to be overwritten. Especially `i`, which is commonly used in loops.

Value

`W` the estiamted signature matrix.

`Z` and the estiamted number of subclones.

`K` the estimated number of subclones.

Source

BayRepulsive: A Bayesian Repulsive Deconvolution Model for Inferring Tumor Heterogeneity

Examples

```
rm(list=ls())
library(BayRepulsive)
data(CCLE)
set.seed(1)
Nobs     <- dim(CCLE$DATA)[1]
Nfeature <- dim(CCLE$DATA)[2]
error    <- matrix(rnorm(Nobs * Nfeature, mean = 0, sd = 0.1), nrow = Nobs)
DATA     <- CCLE$DATA + error
DATA     <- pmax(DATA,0)
result1  <- BayRepulsive_unknown(Datause = DATA, K_min = 2, K_max = 6, Nobs = Nobs,
                                 Nfeature = Nfeature)
cor(as.vector(result1$W), as.vector(CCLE$W))
#----------------
rm(list=ls())
library(BayRepulsive)
data(Inhouse)
Nobs      <- dim(Inhouse$DATA)[1]
Nfeature  <- dim(Inhouse$DATA)[2]
result1  <- BayRepulsive_unknown(Datause = Inhouse$DATA, K_min = 2, K_max = 6, Nobs = Nobs,
                                 Nfeature = Nfeature, seed = 12)
# handle the label swithing issue
W_est     <- result1$W
W_est[,1] <- result1$W[,2]
W_est[,2] <- result1$W[,1]
cor(as.vector(W_est), as.vector(Inhouse$W))
```

CCLE *CCLE*

Description

This dataset was generated from the pure cell line expression data from CCLE .

Usage

```
data(CCLE)
```

Format

This is a data frame with three components: (a) the pure cell line expression, (b) the sample proportion, (c) the mixed data. Z: the pure cell line expression of three cancer cell lines – NCIH524_LUNG, NCIH209_LUNG, SBC5_LUNG. Each line is for one cancer cell line. We selected top 100 differentially expressed gene. W: the sample proportion. Each row is the proportion of one sample. We used this sample porportion to mix 24 mixed samples. DATA: the mixed data. Each line is the gene expression for one sample. DATA = WZ.

Examples

```
# import the data
data(CCLE)
# get the mixed data
CCLE$DATA
# get the gene expression level of pure cell line
CCLE$Z
# get the proportion
CCLE$W
```

Inhouse *Inhouse*

Description

This was generated the pure cell line expression data from one prostate cancer patient at Johns Hopkins hospital.

Usage

```
data(Inhouse)
```

Format

This is a data frame with three components: (a) the pure cell line expression, (b) the sample proportion, (c) the mixed data. Z: the pure cell line expression of three cancer cell lines – naive CD4+ T cell, naive CD8+ T cell, and activated CD4+ T cell in tumor sample. Each line is for one cancer cell line. We selected top 100 differentially expressed gene. W: the sample proportion. Each row is the proportion of one sample. We used this sample porportion to mix 10 mixed samples. DATA: the mixed data with noise. Each line is the gene expression for one sample.

Examples

```
# import the data
data(Inhouse)
# get the mixed data
Inhouse$DATA
# get the gene expression level of pure cell line
Inhouse$Z
# get the proportion
Inhouse$W
# get the added noise
Inhouse$DATA - Inhouse$W%*%Inhouse$Z
```

Index