Package 'BayRepulsive'

October 11, 2018

Type Package

Title BayRepulsive: A Bayesian Repulsive Deconvolution Model for Inferring Tumor Heterogeneity

Version 0.1.0

Date 2018-10-07

Depends mvtnorm, alabama, psych, optimx, Rdpack

Author Yuliang Li and Yanxun Xu

Maintainer Yuliang Li <yli193@jhu.edu>

Description A Bayesian matrix factorization model making use of the determinantal point process as a prior on latent factors to induce repulsiveness among them.

License Johns Hopkins University

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

RdMacros Rdpack

R topics documented:

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

Description

This function gives deconvolution results of the observed matrix, along with the square sum of the residuals.

Usage

```
BayRepulsive_known(DATA, K, Nobs, Nfeature,
                   Niter = 100, epsilon = 0.0001, tau = 100,
                   a.theta = 0.01, b.theta = 0.01, seed = 1 )
```

Arguments

DATA	The observed data matrix. Each row represents a feature (gene); each column represents a sample.
K	The number of subclones.
Nobs	The number of samples, i.e., the number of columns of the DATA.
Nfeature	The number of features, i.e., the number of rows of the DATA.
Niter	The number of maximum iterations.
epsilon	Tolerance for convergence. We determine whether to break based on the estimated weight matrix. We decide to break if the distance induced by L2 norm between two successive estimated weight matrices is less than epsilon.
tau	The hyperparameter for DPP. A large number is preferred. See more in **Details**.
a.theta	The hyperparameter for DPP. See more in **Details**.
b.theta	The hyperparameter for DPP. See more in **Details**.
seed	The random seed.

Details

Given an observed matrix, whose columns are mixed samples of known number of subclones, the function returns the deconvolution results.

The deconvolution model is based on the assumption that

$$Y = ZW + E.$$

Here Y is the observed matrix DATA; Z is the subclone-specific expression matrix; W is the weight matrix; E is the matrix whose entries are independent white noises, with unknown variance σ^2. We assume each column of W, W_j has a prior $W_j \sim \mathrm{Dir}(\alpha)$, where α is a vector with elements 1. We also assume an improper uniform prior for σ^2: $\sigma^2 \sim \mathrm{Uniform}(0, 10^6)$. We use a fixed-size determinant point process (Kulesza and Taskar 2011) as a prior for the subclone-specific expression matrix Z. Suppose there are K subclones and let Z_k be the expression profile of subclone k. Mean zero multivariate normal density functions are commonly used as quality functions in DPP. Since the subclone-specific expression matrix is nonnegative, we cosider a transformation, $\tilde{Z}_k = Z_k - \bar{Y}$, where the vector $\bar{Y}$ is the mean of average expression level in the DATA of each gene. The prior of $(\tilde{Z}_1, \ldots, \tilde{Z}_K)$ is proportional to the determinant of a $K \times K$ matrix L, defined by $L_{ij} = q(\tilde{Z}_i)\phi(\tilde{Z}_i, \tilde{Z}_j)q(\tilde{Z}_j)$, where $\phi(\tilde{Z}_i, \tilde{Z}_j) = \exp\{-\frac{\|\tilde{Z}_i - \tilde{Z}_j\|^2}{\theta^2}\}$ and $q(\tilde{Z}_j)$ is the density function of a multivariate normal distribution with mean being the zero vector and variance being $\tau^2 I$. Here τ is the parameter tau in the function, and θ is an unknown parameter with a prior $\theta \sim \mathrm{Gamma}(a_\theta, b_\theta)$. Here a_θ and b_θ are the parameters a.theta and b.theta in the function.

Value

A list of following components:

Z	The estimated subclone-specific expression matrix.
W	The estimated weight matrix.
C	Square sum of the residuals used as a measure of performance.

Source

BayRepulsive: A Bayesian Repulsive Deconvolution Model for Inferring Tumor Heterogeneity

References

Kulesza A and Taskar B (2011). "k-DPPs: Fixed-size determinantal point processes." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1193–1200.

Examples

```
rm(list=ls())
library(BayRepulsive)
data(CCLE)
set.seed(1)
Nobs     <- 24
Nfeature <- 100
K0       <- 3
### randomly generate weight matrix W for 24 mixing samples
W        <- matrix(0,nrow = K0, ncol = Nobs)
for(i in 1:Nobs){
  Theta <- rgamma(K0,1/K0,1)
  W[,i] <- Theta/sum(Theta)
}
### add some noise
error    <- t(matrix(rnorm(Nfeature * Nobs, mean = 0, sd = 0.5), nrow = Nobs))
DATA     <- CCLE$Z%*%W + error
### Note: please make sure that there are no negative values after adding the noise
result1  <- BayRepulsive_known(DATA = DATA, K = K0, Nobs = Nobs,
                               Nfeature = Nfeature)
cor(as.vector(result1$W), as.vector(W))
```

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

Description

This function gives the estimated number of subclones, along with deconvolution results of the observed matrix.

Usage

```
BayRepulsive_unknown(DATA, K_min, K_max, Nobs, Nfeature,
                     Niter = 100, epsilon = 0.0001, tau = 100,
                     a.theta = 0.01, b.theta = 0.01, seed = 1 )
```

Arguments

`DATA`	The observed data matrix. Each row represents a feature (gene); each column represents 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 columns of the DATA.
Nfeature	The number of features, i.e., the number of rows of the DATA.
Niter	The number of maximum iterations.
epsilon	Tolerance for convergence. We determine whether to break based on the estimated weight matrix. We decide to break if the distance induced by L2 norm between two successive estimated weight matrices is less than epsilon.
tau	The hyperparameter for DPP. A large number is preferred. See BayRepulsive_known for more detials.
a.theta	The hyperparameter for DPP. See BayRepulsive_known for more detials.
b.theta	The hyperparameter for DPP. See BayRepulsive_known for more detials.
seed	The random seed.

Details

Given an observed matrix, whose columns are mixed samples of unknown number of subclones, this function gives an estimation of number of subclones along with deconvolution results.

We first use the algorithm in BayRepulsive_known to fit the data for every possible number of subclones. Let $S(k)$ denote the square sum of the residuals when the number of subclones is fixed at k. We define the second-order finite difference $\Delta^2 S(k)$ of the residual by $\Delta^2 S(k) = [S(k+1) - S(k)] - [S(k) - S(k-1)]$, for $k = K_{min} + 1, \ldots, K_{max} - 1$. Then the optimal $\hat{K}$ estimated by BayRepulsive is

$$\hat{K} = \text{argmax}_k \Delta^2 S(k).$$

And the deconvolution results are the corresponding results when the number of subclones is fixed at $\hat{K}$.

Value

A list of following components:

Z	The estimated subclone-specific expression matrix.
W	The estimated weight matrix.
K_hat	The number of estimated subclones.

Source

BayRepulsive: A Bayesian Repulsive Deconvolution Model for Inferring Tumor Heterogeneity

See Also

BayRepulsive_known

Examples

```
rm(list=ls())
library(BayRepulsive)
data(CCLE)
set.seed(1)
Nobs     <- 24
Nfeature <- 100
K0       <- 3
```

```
### randomly generate weight matrix W for 24 mixing samples
W        <- matrix(0,nrow = K0, ncol = Nobs)
for(i in 1:Nobs){
  Theta <- rgamma(K0,1/K0,1)
  W[,i] <- Theta/sum(Theta)
}
### add some noise
error    <- t(matrix(rnorm(Nfeature * Nobs, mean = 0, sd = 0.5), nrow = Nobs))
DATA     <- CCLE$Z%*%W + error
### Note: please make sure that there are no negative values after adding the noise
result1  <- BayRepulsive_unknown(DATA = DATA, K_min = 2, K_max = 6, Nobs = Nobs,
                                 Nfeature = Nfeature)
cor(as.vector(result1$W), as.vector(W))
```

CCLE *The CCLE Dataset*

Description

This dataset is the pure cell line expression from *Cancer Cell Line Encyclopedia* (CCLE) (Barretina et al. 2012).

Usage

```
data(CCLE)
```

Format

From the CCLE dataset, we randomly chose three lung-related cell lines, NCIH524_LUNG, NCIH209_LUNG and SBC5_LUNG. We then selected the top 100 differentially expressed genes by ranking the standard deviations of genes across pure samples. The expression levels of these 100 genes in the selected three cell lines compose our simluated Z matrix. This is a data frame with one components: the pure cell line expression, Z.

References

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D and others (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." *Nature*, **483**(7391), pp. 603–607.

Examples

```
# import the data
data(CCLE)
# get the gene expression level of pure cell line
CCLE$Z
# get the name of cell lines included in the dataset
colnames(CCLE$Z)
```

Index