Package 'ZIBBSeqDiscovery'

March 23, 2018

Type Package

Title Zero-Inflated Beta-Binomial Modeling of Microbiome Count Data

Version 1.0

Date 2018-03-22

Author Tao Hu, Yihui Zhou

Maintainer Yihui Zhou <yihui_zhou@ncsu.edu>

Description Microbiome count data (Operational Taxonomic Unit, OTUs) is usually overdispersed and has excessive zero counts. The 'ZIBBSeqDiscovery' assumes a zero-inflated beta-binomial model for the distribution of the count data, and employes link functions to adjust interested covariates. To fit the model, two approaches are proposed (i) a free approach which treats the overdispersion parameters for OTUs as independent, and (ii) a constrained approach which proposes a mean-overdispersion relationship to the count data. This package can be used to test the association between the composition of the microbiome counts and the interested covariates.

License GPL-2

Imports mcc

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2018-03-23 13:48:19 UTC

R topics documented:

`ZIBBSeqDiscovery-package`
Zero-Inflated Beta-Binomial Modeling of Microbiome Count Data

---

Description

Microbiome count data (Operational Taxonomic Unit, OTUs) is usually overdispersed and has excessive zero counts. The 'ZIBBSeqDiscovery' assumes a zero-inflated beta-binomial model for the distribution of the count data, and employes link functions to adjust interested covariates. To fit the model, two approaches are proposed (i) a free approach which treats the overdispersion parameters for OTUs as independent, and (ii) a constrained approach which proposes a mean-overdispersion relationship to the count data. This package can be used to test the association between the composition of the microbiome counts and the interested covariates.

Details

Index of help topics:

```
ZIBBSeqDiscovery-package
                        Zero-Inflated Beta-Binomial Modeling of
                        Microbiome Count Data
constrained.estimate    Estimate parameters with constrained approach
constrained.loglikelihood
                        Define the objective function in optimization
                        procedure for estimating parameters with
                        constrained approach
data.Y                  Data set for ZIBBSeqDiscovery
fitZIBB                 The main function to fit ZIBB model
free.estimate           Estimate parameters with free approach
free.loglikelihood      Define the objective function in optimization
                        procedure for estimating parameters with free
                        approach
kostic.x                Data set for ZIBBSeqDiscovery
kostic.y                Data set for ZIBBSeqDiscovery
mcc.adj                 Using MCC method to replace NAs in the p values
regression              Simple linear regression
```

Author(s)

Tao Hu, Yihui Zhou

Maintainer: Yihui Zhou <yihui_zhou@ncsu.edu>

`constrained.estimate` *Estimate parameters with constrained approach*

Description

Estimate unknown parameters with constrained approach.

Usage

```
constrained.estimate(m, p, q, n, betastart, bvarstart, psi.start,
                     eta.start, gamma.start, Y, X, Y.c, ziMatrix,
                     gn = 3)
```

Arguments

`m`	Number of OTUs.
`p`	Number of covariates for count model (e.g., beta-binomial).
`q`	Number of covariates for zero model.
`n`	Number of samples.
`betastart`	Matrix of estimated betas, which are the effects/coefficients for the count model, with dimension p by m. It is used as initial values for the optimization procedure to estimate betas.
`bvarstart`	Matrix of variance of estimated betas with dimension p by m.
`psi.start`	Estimated vector of logit of overdispersion parameters with length m. And psi.start will be used as initial values for the optimization procedure to estimate psi.
`eta.start`	Matrix of estimated etas, which are the effects/coefficients for the zero model, with dimension q by m. It is used as initial values for the optimization procedure to estimate etas.
`gamma.start`	Estimation vector of the coefficients in the polynomial mean-overdispersion relationship in constrained approach.
`Y`	Count matrix with dimension n by m.
`X`	The design matrix (n by p, p is the number of covariates) for the count model (e.g., beta-binomial), and intercept is included.
`Y.c`	Vector of library size with length n.
`ziMatrix`	The design matrix (n by q) for the zero model, and intercept is included.
`gn`	We use a polynomial with degree of freedom gn to fit the mean-overdispersion relationship.

Value

`betahat`	Estimation matrix of beta (p by m).
`bvar`	Estimation matrix of the variance of estimated betahat (p by m).
`psi`	Estimation vector of the logit of the overdispersion parameters (with length m).
`eta`	Estimation matrix of eta (q by m).

Author(s)

Tao Hu, Yihui Zhou

`constrained.loglikelihood` *Define the objective function in optimization procedure for estimating parameters with constrained approach*

Description

The objective function is the negative of log likelihood function.

Usage

```
constrained.loglikelihood(para, X, Y.col, coeff, Y.c, ziMatrix)
```

Arguments

`para`	Vector of optimized parameters with length p+q, where p is the number of covariates for count model (e.g., beta-binomial), q is the number of covariates for zero model. The first p elements are betas which are the effects/coefficients for the count model. The last q elements are etas which are the effects/coefficients for the zero model.
`X`	The design matrix (n by p, p is the number of covariates) for the count model (e.g., beta-binomial), and intercept is included.
`Y.col`	Vector of counts corresponding to an OTU, with length n.
`coeff`	Vector of coefficients in the polynomial mean-overdispersion relationship in constrained approach.
`Y.c`	Vector of library size with length n.
`ziMatrix`	The design matrix (n by q) for the zero model, and intercept is included.

Author(s)

Tao Hu, Yihui Zhou

data.Y *Data set for ZIBBSeqDiscovery*

Description

Data set for ZIBBSeqDiscovery. The data set data.Y is a count matrix with 100 OTUs and 20 samples randomly selected from a real microbiome data set.

Usage

```
data(data.Y)
```

Examples

```
## Load the data
data(data.Y)
dim(data.Y)
```

fitZIBB *The main function to fit ZIBB model*

Description

We use zero-inflated beta-binomial (ZIBB) to account for overdispersion and excessive zeros in the microbiome count data. The parameter estimation method is maximum likelihood. Two approaches are proposed to estimate the overdispersion parameters: free approach and constrained approach. For free approach, user does not need to provide initial values for unknown parameters because our program will come up with naive initial values automatically. For constrained approach, user should provide the estimations from free approach as the initial values. This function gives the estimations of the parameters, as well as the corresponding p values, which can be used to test the association between the composition of microbiome counts data and the interested covariates.

Usage

```
fitZIBB(dataMatrix, X, ziMatrix, mode = "free", gn = 3,
        betastart = matrix(NA, 0, 0),
        psi.start = vector(mode = "numeric", length = 0),
        eta.start = matrix(NA, 0, 0))
```

Arguments

dataMatrix	The count matrix (m by n, m is the number of OTUs and n is the number of samples).
X	The design matrix (n by p, p is the number of covariates) for the count model (e.g., beta-binomial), and intercept is included. The second column is assumed to be the covariate of interest.

ziMatrix	The design matrix (n by q, q is the number of covariates) for the zero model, and intercept is included.
mode	Indicates which approach is used to estimate overdispersion parameters. mode can be set as either "free" or "constrained".
gn	In constrained approach, we use a polynomial with degree of freedom gn to fit the mean-overdispersion relationship. The default value for gn is 3. Note that gn is only valid when mode = "constrained".
betastart	Initial values for beta estimation matrix (p by m), where beta are the effects (or coefficients) for the count model. betastart is required only in constrained approach, and it should be assigned as the beta estimation matrix from free approach.
psi.start	Initial values for the logit of overdispersion parameters vector (with length m). psi.start is required only in constrained approach, and it should be assigned as the psi estimation vector from free approach.
eta.start	Initial values for eta estimation matrix (q by m), where eta are the effects (or coefficients) for the zero model. eta.start is required only in constrained approach, and it should be assigned as the eta estimation matrix from free approach.

Details

In this package, we refer the covariate of interest as phenotype (only one phenotype is assumed currently, thouugh we can extend it to include multiple phenotypes), and the phenotype is assumed to corresonding to the second column of the degisn matrix X (note that the first column corresonds to the intercept). Assuming the parameters corresonding to the phenotype are $\{\beta_{1i}\}_{i=1,\ldots,m}$, this function tests the null hypothesis $H_0 : \beta_{1i} = 0$ for each OTU i.

Value

betahat	Estimation matrix of beta (p by m) for count model.
bvar	Estimation matrix of the variance of estimated betahat (p by m).
p	The vector (with length of m) of p values corresponding to the phenotype (aka, covariate of interest). In this package, we assume it corresponds to the second column of design matrix X (because intercept is included), though you can always include multiple phenotypes or other covariates. The i'th p value corresponds to the hypothesis test of $H_0 : \beta_{1i} = 0$ for OTU i
psi	Estimation vector of the logit of the overdispersion parameters (with length m).
zeroCoef	Estimation matrix of eta (q by m) for zero model.
gamma	Estimation vector of the coefficients in the mean-overdispersion relationship in constrained approach (with length gn+1). So gamma is only available when mode="constrained".

Author(s)

Tao Hu, Yihui Zhou

Examples

```
## Load the data
## data.Y is a count matrix with 100 OTUs and 20 samples randomly selected
##   from kostic data
data(data.Y)

## set random seed
set.seed(1)

## construct design matrix for count model
## data.X is a 20-by-2 matrix, phenotype is group, and the first 10 samples
##   come from group 1 and the rest samples come from group 2
data.X <- matrix(c(rep(1, 20), rep(0,10), rep(1, 10)), 20, 2)

## construct design matrix for zero model
## data.ziMatrix is a 20-by-2 matrix, the covariate is log of library size
data.ziMatrix <- matrix(1, 20, 2)
data.ziMatrix[, 2] <- log(colSums(data.Y))

## fit ZIBB with free approach
out.free <- fitZIBB(data.Y, data.X, data.ziMatrix, mode = "free")

## fit ZIBB with constrained approach
out.constrained <- fitZIBB(data.Y, data.X, data.ziMatrix,
                           mode = "constrained", gn = 3,
                           betastart = out.free$betahat,
                           psi.start = out.free$psi,
                           eta.start = out.free$zeroCoef)

## print OTUs which has p values smaller than 0.05
out.constrained$p[which(out.constrained$p < 0.05)]
```

`free.estimate` *Estimate parameters with free approach*

Description

Estimate unknown parameters with free approach.

Usage

```
free.estimate(m, p, q, n, betastart, bvarstart, psi.start, eta.start,
              Y, X, Y.c, ziMatrix)
```

Arguments

`m`	Number of OTUs.
`p`	Number of covariates for count model (e.g., beta-binomial).
`q`	Number of covariates for zero model.

n	Number of samples.
betastart	Matrix of estimated betas, which are the effects/coefficients for the count model, with dimension p by m. It is used as initial values for the optimization procedure to estimate betas.
bvarstart	Matrix of variance of estimated betas with dimension p by m.
psi.start	Estimated vector of logit of overdispersion parameters with length m. And psi.start will be used as initial values for the optimization procedure to estimate psi.
eta.start	Matrix of estimated etas, which are the effects/coefficients for the zero model, with dimension q by m. It is used as initial values for the optimization procedure to estimate etas.
Y	Count matrix with dimension n by m.
X	The design matrix (n by p, p is the number of covariates) for the count model (e.g., beta-binomial), and intercept is included.
Y.c	Vector of library size with length n.
ziMatrix	The design matrix (n by q) for the zero model, and intercept is included.

Value

betahat	Estimation matrix of beta (p by m).
bvar	Estimation matrix of the variance of estimated betahat (p by m).
psi	Estimation vector of the logit of the overdispersion parameters (with length m).
eta	Estimation matrix of eta (q by m).

Author(s)

Tao Hu, Yihui Zhou

---

free.loglikelihood *Define the objective function in optimization procedure for estimating parameters with free approach*

Description

The objective function is the negative of log likelihood function.

Usage

```
free.loglikelihood(para, X, Y.col, Y.c, ziMatrix)
```

Arguments

para	Vector of optimized parameters with length p+q+1, where p is the number of covariates for count model (e.g., beta-binomial), q is the number of covariates for zero model. The first p elements are betas which are the effects/coefficients for the count model. The (p+1)'th element is the logit of the overdispersion parameter. The last q elements are etas which are the effects/coefficients for the zero model.
X	The design matrix (n by p, p is the number of covariates) for the count model (e.g., beta-binomial), and intercept is included.
Y.col	Vector of counts corresponding to an OTU, with length n.
Y.c	Vector of library size with length n.
ziMatrix	The design matrix (n by q) for the zero model, and intercept is included.

Author(s)

Tao Hu, Yihui Zhou

---

kostic.x *Data set for ZIBBSeqDiscovery*

Description

Data set for ZIBBSeqDiscovery. The data set kostic.x is a count matrix with 2505 OTUs and 185 samples randomly selected from a real microbiome data set kostic.

Usage

```
data("kostic.x")
```

References

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., et al. (2012). Genomic analysis identifies association of fusobacterium with colorectal carcinoma. Genome research 22, 292-298.

Examples

```
## load the data
data(kostic.x)
dim(kostic.x)
```

kostic.y *Data set for ZIBBSeqDiscovery*

Description

Data set for ZIBBSeqDiscovery. The data set kostic.y is the phenotype of the 185 samples from a real microbiome data set kostic.

Usage

```
data("kostic.y")
```

References

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., et al. (2012). Genomic analysis identifies association of fusobacterium with colorectal carcinoma. Genome research 22, 292-298.

Examples

```
## load the data
data(kostic.y)
length(kostic.y)
```

mcc.adj *Using MCC method to replace NAs in the p values*

Description

When fitting the ZIBB model, some parameter estimations may fail due to numerical issues. In that case, a NA will be given as the corresponding p value. Here, a Moment Corrected Correlation (MCC) approach is employed to replace the NAs in the p values.

Usage

```
mcc.adj(out.fitZIBB, dataMatrix, X, ziMatrix, K = 4)
```

Arguments

out.fitZIBB	The output from function fitZIBB.
dataMatrix	The count matrix (m by n, m is the number of OTUs and n is the number of samples).
X	The design matrix (n by p, p is the number of covariates) for the count model (e.g., beta-binomial), and intercept is included. The second column is assumed to be the covariate of interest.

ziMatrix	The design matrix (n by q, q is the number of covariates) for the zero model, and intercept is included.
K	Divide covariate in ziMatrix (second colunm in default) into K stratum, under the requirement of MCC approach. The default value of K is 4.

Value

The output has the exact same format as function fitZIBB, with corrected p values.

Author(s)

Tao Hu, Yihui Zhou

References

Zhou, Y. H., & Wright, F. A. (2015). Hypothesis testing at the extremes: fast and robust association for high-throughput data. Biostatistics, 16(3), 611-625.

See Also

fitZIBB

Examples

```
## Load the data
## data.Y is a count matrix with 100 OTUs and 20 samples randomly selected
##   from kostic data
data(data.Y)

## set random seed
set.seed(1)

## construct design matrix for count model
## data.X is a 20-by-2 matrix, phenotype is group, and the first 10 samples
##   come from group 1 and the rest samples come from group 2
data.X <- matrix(c(rep(1, 20), rep(0,10), rep(1, 10)), 20, 2)

## construct design matrix for zero model
## data.ziMatrix is a 20-by-2 matrix, the covariate is log of library size
data.ziMatrix <- matrix(1, 20, 2)
data.ziMatrix[, 2] <- log(colSums(data.Y))

## fit ZIBB with free approach
out.free <- fitZIBB(data.Y, data.X, data.ziMatrix, mode = "free")

## count how many NAs in the p values
sum(is.na(out.free$p))

## MCC adjustment
out.free.mcc <- mcc.adj(out.free, data.Y, data.X, data.ziMatrix, K=4)

## count how many NAs in the p values after MCC adjustment
```

```
sum(is.na(out.free.mcc$p))
```

`regression` *Simple linear regression*

Description

Use simple linear regression to find the initial value of betas in free approach, where beta are the effects/coefficients for the count model.

Usage

```
regression(y, x)
```

Arguments

`y`	The response vector.
`x`	The design matrix (n by p, p is the number of covariates) for the count model (e.g., beta-binomial), and intercept is included.

Value

`betahat`	Initial values for beta estimation matrix (p by m, p is the number of involved covariates and m is the number of OTUs).
`se`	Standard errors for the estimated betas.

Author(s)

Tao Hu, Yihui Zhou

Index