Package 'MixRF'

January 20, 2025

Title A Random-Forest-Based Approach for Imputing Clustered Incomplete Data

Version 1.0

Date 2016-04-05

Author Jiebiao Wang and Lin S. Chen

Maintainer Jiebiao Wang <randel.wang@gmail.com>

Description It offers random-forest-based functions to impute clustered incomplete data. The package is tailored for but not limited to imputing multitissue expression data, in which a gene's expression is measured on the collected tissues of an individual but missing on the uncollected tissues.

License GPL

Depends doParallel, randomForest, lme4, foreach

URL https://github.com/randel/MixRF

BugReports https://github.com/randel/MixRF/issues

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2016-04-06 09:43:04

Contents

`MixRF-package` *A random-forest-based algorithm for imputing clustered incomplete data*

Description

This package offers random-forest-based functions to impute clustered incomplete data. The package is tailored for but not limited to imputing multitissue expression data, in which a gene's expression is measured on the collected tissues of an individual but missing on the uncollected tissues.

Details

Package:	MixRF
Type:	Package
Version:	1.0
Date:	2016-04-05
License:	GPL
LazyLoad:	yes

Author(s)

Jiebiao Wang and Lin S. Chen

Maintainer: Jiebiao Wang <randel.wang@gmail.com>

References

Wang, J., Gamazon, E.R., Pierce, B.L., Stranger, B.E., Im, H.K., Gibbons, R.D., Cox, N.J., Nicolae, D.L. and Chen, L.S. (2016) Imputing gene expression in uncollected tissues within and beyond GTEx. http://dx.doi.org/10.1016/j.ajhg.2016.02.020

See Also

`MixRF.impute`

`MixRF` *Mixed Random Forest*

Description

The function to fit a random forest with random effects.

Usage

```
MixRF(Y, X, random, data, initialRandomEffects = 0, ErrorTolerance = 0.001,
    MaxIterations = 1000)
```

Arguments

`Y`	The outcome variable.
`X`	A data frame or matrix contains the predictors.
`random`	A string in lme4 format indicates the random effect model.
`data`	The data set as a data frame.
`initialRandomEffects`	The initial values for random effects.
`ErrorTolerance`	The tolerance for log-likelihood.
`MaxIterations`	The maximum iteration times.

Value

A list contains the random forest ($forest), mixed model ($MixedModel), and random effects ($RandomEffects). See the example below for the usage.

Examples

```
data(sleepstudy)

tmp = MixRF(Y = sleepstudy$Reaction, X = as.data.frame(sleepstudy$Days),
    random = "(Days|Subject)", data = sleepstudy, initialRandomEffects = 0,
    ErrorTolerance = 0.01, MaxIterations = 100)

# tmp$forest

# tmp$MixedModel

# tmp$RandomEffects
```

`MixRF.impute` *Impute a large number of genes using the MixRF algorithm with parallel computing*

Description

This function impute the expression of a large number of genes using the MixRF algorithm with parallel computing.

Usage

```
MixRF.impute(Ydat, eqtl.lis, snp.dat, cov = NULL, iPC = TRUE,
    idx.selected.gene.iPC = NULL, parallel.size = 1, correlation = FALSE,
    nCV = 3)
```

Arguments

`Ydat`	An array of expression data of dimension sample-by-gene-by-tissue, nxpxT, where n is sample size. p is the number of genes, and T is the number of tissues. Ydat[,1,] is a matrix of the first gene expression in T tissues for n individuals, nxT. Ydat[,,1] is a nxp matrix of the expression data of p genes in the first tissue.
`eqtl.lis`	A list of eQTL names of length p. Each element in the list contains the name of the eQTLs for the corresponding gene. The order of the list should correspond to the order of genes in Ydat. The code and example to calculate eQTLs can be found at https://github.com/randel/MixRF/blob/master/R/eqtl.r.
`snp.dat`	A matrix of genotype. Each row is a sample and each column corresponds to one SNP. The column names should match eqtl.lis.
`cov`	A matrix of covariates. Each row is a sample and each column corresponds to one covariate. For example, age, gender.
`iPC`	An option. When it is TRUE, the imputed PCs (iPCs) for each tissue type will be constructed based on the combined observed and imputed data on the selected genes. The iPCs will be adjusted as covariates in the imputation.
`idx.selected.gene.iPC`	The option is used only when iPC=TRUE. When it is, one may select a subset of genes and impute those first to construct iPCs.
`parallel.size`	A numerical value specifying the number of CPUs/cores/processors available for parallel computing.
`correlation`	The option to calculate the imputation correlation using cross-validation or not. The default is FALSE.
`nCV`	The option is used only when correlation=TRUE. The number of folds for cross-validation. The default is 3 folds.

Value

An nxpxT array of imputed and observed expression data. The observed values in Ydat are still kept and the missing values in Ydat are imputed. When the user chooses to calculate the imputation correlation using cross-validation (correlation=TRUE), the estimated imputation correlation (cor) will also be returned in a list together with the imputed data (Yimp).

Examples

```
## Not run:
data(sim)

idx.selected.gene.iPC = which(sapply(sim$eqtl.lis, length) >= 1)

Yimp = MixRF.impute(sim$Ydat, sim$eqtl.lis, sim$snp.dat, sim$cov, iPC = TRUE,
    idx.selected.gene.iPC, parallel.size = 4)

## End(Not run)
```

---

sim *Simulated data list*

Description

This simulated data list is for demonstration.

Value

`Ydat`	An array of expression data of dimension sample-by-gene-by-tissue, nxpxT, where n is sample size. p is the number of genes, and T is the number of tissues. Ydat[,1,] is a matrix of the first gene expression in T tissues for n individuals, nxT. Ydat[,,1] is a nxp matrix of the expression data of p genes in the first tissue.
`eqtl.lis`	A list of eQTL names of length p. Each of the element in the list contains the name of the eQTLs for the corresponding gene. The order of the list should correspond to the order of genes in Ydat.
`snp.dat`	A matrix of genotype. Each row is a sample and each column corresponds to one SNP. The column names should match eqtl.lis.
`cov`	A matrix of covariates. Each row is a sample and each column corresponds to one covariate. For example, age, gender.

See Also

`MixRF.impute`

Index