# Package 'bastah'

June 2, 2016

**Type** Package

**Title** Big Data Statistical Analysis for High-Dimensional Models

**Version** 1.0.7

**Date** 2016-05-22

**Author** Ehsan Ullah [aut, cre], Reda Rawi [aut], Lukas Meier [ctb], Ruben Dezeure [ctb], Nicolai Meinshausen [ctb], Martin Maechler [ctb], Peter Buehlmann [ctb], Tingni Sun [ctb]

**Maintainer** Ehsan Ullah <eullah@qf.org.qa>

**Repository** CRAN

**Description** Big data statistical analysis for high-dimensional models is made possible by modifying lasso.proj() in 'hdi' package by replacing its nodewise-regression with sparse precision matrix computation using 'BigQUIC'.

**License** GPL (== 2)

**LazyData** TRUE

**Enhances** doMC, rPython

**Imports** BigQuic, foreach, glmnet, lars, MASS, stats, Matrix, scalreg

**Depends** R (>= 3.2.0)

**NeedsCompilation** no

**Date/Publication** 2016-06-02 14:44:01

## R topics documented:

bastah *Big Data Statistical Analysis for High-Dimensional Models*

## Description

Big data statistical analysis for high-dimensional models is made possible by modifying lasso.proj() in 'hdi' package by replacing its nodewise-regression with sparse precision matrix computation using 'BigQUIC'.

## Usage

```
bastah (X, y, categorical = FALSE, family = "gaussian", mcorr = "holm",
N = 10000, ncores = 4, verbose = FALSE)
```

## Arguments

| | |
|---|---|
| X | An n by p numeric design matrix with p columns for p predictor variables and n rows corresponding to n observations. |
| y | A numeric response variable of length n. |
| categorical | Type of data in the design matrix. (default = FALSE) |
| family | Family of the response variable. It should be either "gaussian" or "binomial". (default = "gaussian") |
| mcorr | Multiple correction method. It can be either "WY" or any of p.adjust.methods. (default = "holm") |
| N | It is the number of samples to take for the empirical distribution which is used to correct the pvalues if multiple correction method is "WY" (Westfall-Young). (default = 10000) |
| ncores | Maximum number of cores to be used for parallel execution. (default = 4) |
| verbose | Prints more information if this is set to TRUE. (default = FALSE) |

## Details

In this package lasso.proj function of hdi package is updated for application on big data. The original lasso.proj is updated by replacing node-wise regression with scaled lasso. BigQUIC is used for sparse precision matrix calculation. Data is always normalized before processing. Normalization technique used by Vlaming and Groenen (2014) is used. The method has been successfully used on large SNP (Single Nucleotide Polymorphism) datasets for GWAS (Genomewide Association Study).

The package can use scikit-learn (http://scikit-learn.org) for a better performance. It is advised to install doMC, rPython, python, numpy and scikit-learn. The package uses scikit-learn at runtime, therefore, python, numpy and scikit-learn are not required for package installation and can be installed after installation of the package.

NOTE: We have noticed that lars package in R crashes, so it is recommended to use scikit-learn.

NOTE: In preprocessing step, variables having a constant value are not considered. The list of varaibles used is returned in selection varaible of the result.

## Value

An object with Class "bastah"

| | |
|---|---|
| pval | Calculated p-values |
| pval.corr | Corrected p-values |
| sigmahat | Estimated standard deviation |
| bhat | Estimated coefficients |
| selection | Indicies of variables selected for analysis |

## Author(s)

Ehsan Ullah [aut, cre], Reda Rawi [aut], Lukas Meier [ctb], Ruben Dezeure [ctb], Nicolai Meinshausen [ctb], Martin Maechler [ctb], Peter Buehlmann [ctb], Tingni Sun [ctb]

Maintainer: Ehsan Ullah <eullah@qf.org.qa>

## References

C. Hsieh, M. Sustik, I. Dhillon, P. Ravikumar, R. Poldrack. In Neural Information Processing Systems (NIPS), December 2013. (Oral)

S. van de Geer, P. Buhlmann, Y. Ritov and R. Dezeure (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics 42, 1166-1202.

C. Zhang, S. Zhang(2014) Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B76, 217-242.

P. Buhlmann and S. van de Geer(2015) High-dimensional inference in misspecified linear models. Electronic Journal of Statistics 9, 1449-1473.

R. de Vlaming and P. J. F. Groenen (2014) The current and future use of ridge regression for prediction in quantitative genetics. BioMed Research International, 2015, 143712.

## Examples

```
# The package is accompanied with a simulated genome-wide association
# study dataset "snps" containing n=100 observations of p=500 predictors
   data(snps)
# The association of SNPs to the phenotype can be identified using bastah
# NOTE: We have noticed that lars package in R crashes,
# so it is recommended to use scikit-learn (see package details).
## Not run:
      result = bastah(X = snps$X, y = snps$y, family = "binomial", verbose = TRUE)

## End(Not run)
```

snps *SNP data set*

## Description

Simulated Single Polymorphism Nucleotide (SNP) dataset containing $n = 100$ observations of $p = 500$ predictors (SNPs, 1=Homozygote1, 0=Heterozygote, -1=Homozygote2) and a one-dimensional response (1=case, 0=control). The dataset is generated using GWAsimulator.

## Usage

```
data(snps)
```

## Format

**y** Phenotype (1=case, 0=control) of 100 individuals.

**x** SNP genotype data (SNPs, 1=Homozygote1, 0=Heterozygote, -1=Homozygote2) of 500 simulated SNPs.

## References

C. Li and M. Li (2008) GWAsimulator: a rapid whole-genome simulation program. Bioinformatics 24 (1): 140-142.

## Examples

```
data(snps)
```

# Index