Package 'sp.gwas'

November 6, 2019

Type Package

Title Selection probabilities using glmnet for GWAS

Version 0.2.0

Author Kipoong Kim

Maintainer Kipoong Kim <kkp7700@gmail.com>

Description Selection probabilities using generalized linear model with regularization for a SNP data in the hapmap format.

License GPL-2 | GPL-3

Imports stats, ggplot2, compiler, glmnet, dplyr, CMplot, readxl, data.table, writexl, bestNormalize, gridExtra, tidyr, gtools

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

R topics documented:

`sp.gwas-package` *Selection probabilities using generalized linear model with regularization for a SNP data in the hapmap format.*

Description

sp.gwas

Details

The penalty function of `elastic-net` is defined as

$$\alpha||\beta||_1 + (1-\alpha)||\beta||_2/2,$$

where α is a mixing proportion of ridge and the lasso, and β is regression coefficients. This penalty is equivalent to the Lasso penalty if `alpha=1`.

Value

A list of data files(genotype, phenotype, etc.), results for selection probabilities, and manhattan plot for multiple traits.

References

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320.

---

sp.gwas *Selection probabilities using generalized linear model with regularization for a SNP data in the hapmap format.*

Description

For analysis of high-dimensional genomic data, penalized regression can be a solution to accomodate correlations between predictors. Moreover, selection probabilities do not depend on tuning parameter selection so that it produces a stablity selection. Thresholds are also generated to control the false positives(errors). Thresholds vary with the expected number of false positives to be controlled by the user. Input data is the hapmap formatted SNP data and phenotype data corresponding to SNP data. Output includes files of three types: (1) Matched data files (genotype, numerical, snp info, and phenotype), (2) Results file (selection probabilities and thresholds), (3) Circular manhattan plot with (blue dotted) significant line corresponding to the largest value among user-defined false discoveries.

Usage

```
sp.gwas(genotype.path = NULL, phenotype.path = NULL,
  save.path = "./sp.folder", y.col = 4, y.id.col = 2,
  method = "lasso", family = "gaussian", Falsediscovery = c(1, 5,
  10), plot.ylim = NULL, lambda.min.quantile = 0.5, n.lambda = 10,
  K = 100, psub = 0.5, manhattan.type = "c", plot.name = "",
  plot.type = "jpg", plot.dpi = 300)
```

Arguments

`genotype.path` A path of a snp data which is a `p` by `(n+11)` matrix of genotypes. It is formatted by hapmap which has (rs, allele, chr, pos) in the first four(1-4) columns, (strand, assembly, center, protLSID, assayLSID, panel, Qcode) in the following seven(5-11) columns. If NULL, user can choose a path in interactive use.

`phenotype.path` A path of a phenotype data which is an `n` by `p` matrix of phenotypes. Since the first some columns can display attributes of the phenotypes, you should enter the arguments, y.col and y.id.col, which represent the columns of phenotypes to be analyzed and the column of sample ID. If NULL, user can choose a path in interactive use.

`save.path` A save.path which has all output files. If there exists save.path, sp.gwas will check if there is an output file. If then, sp.gwas is not going to generate the results but just loading the output files(.RData).

`y.col` The columns of phenotypes. At most 4 phenotypes can be considered, because the plot of them will be fine. Default is 4.

`y.id.col` The column of sample ID in the phenotype data file. Default is 2.

`method` A method of penalized regression. It includes "lasso" for the lasso and "enet" for the elastic-net.

`family` A family of response variable(phenotype). It is "gaussian" for continuous response variable, "binomial" for binary, "poisson" for count, etc. Now you can use only the same family for the multi phenotypes. For more details, see the function(`stats::glm`). Default is "gaussian".

`Falsediscovery` The expected number of false discovery to be controlled. The larger it is, the higher threshold becomes. Default is c(1, 5, 10).

`plot.ylim` A range of the y-axis. If NULL, automatic range in the y-axis will be provided. For plot.ylim=c(0,1), the y-axis has a range of 0 and 1.

`lambda.min.quantile`
A range of lambda sequence. Default is 0.5 (median). If the range is so small that it can have many tied selection probabilities which is 1.

`n.lambda` The length of lambda sequence. The larger n.lambda, the more detailed lambda sequence will be.

`K` The number of iterations in resampling when calculating the selection probabilities.

`psub` The subsampling proportion. For efficiency, default is 0.5.

`manhattan.type` A type of manhattan plot to be drawn includes circular('c') and rectangular('m').

`plot.name` A name of plot file.

`plot.type` A type of plot file which includes "jpg", "pdf", "tiff", etc.

`plot.dpi` A resolution of plot. If you want to get a high-resolution image, plot.dpi should be large.

Details

The penalty function of `elastic-net` is defined as

$$\alpha||\beta||_1 + (1-\alpha)||\beta||_2/2,$$

where α is a mixing proportion of ridge and the lasso, and β is regression coefficients. This penalty is equivalent to the Lasso penalty if `alpha=1`.

`An algorithm of selection probabilities with elastic-net.`

```
Let us assume that a genomic data has n samples and p variables.
For all $\Lambda = (\alpha, \lambda)$, where $\alpha in [0,1]$, $\lambda > 0$.
for k=1 to K do.
    ——- Subsample $I_k$ with size $[n/2]$.
    ——- Compute $\hat{\beta}_j^{\Lambda}(I_k)$ with regularization model.
end for.
$SP_j^{\Lambda} = \frac{1}{K}\#\{k \le K : \hat{\beta}_j^{\Lambda}(I_k) \neq 0\}$.
$SP_j = \max_{\Lambda} SP_j^{\Lambda},\ \ j = 1, \cdots, p$.
return SP=$(SP_1, \cdots, SP_p)$.
```

Value

`Histogram of original and transformed phenotypes`
Histogram of phenotypes with p-value by Shapiro-test on the top right corner.

`myDATA` A list of myX, myGD, myGM, myGT, myY, and myY.original(for "gaussian").

`sp.res` A list of sp.df and threshold.

`Circular Manhattan plot`
Manhattan plot for the first phenotype is the innermost circle. Colors for chromosome is fixed, so that if you want to change colors, you would edit the R code of sp.manhattan function.

The dataframe with new mean and sum columns

Author(s)

Kipoong Kim <kkp7700@gmail.com>

Examples

```
# Not run
# sp.gwas(genotype.path = "./input.snp.csv",
#         phenotype.path = "./input.phenotype.csv",
#         save.path = "./Test",
#         y.col=5:8,
#         y.id.col=2,
#         method="lasso",
#         family="gaussian",
#         Falsediscovery = c(1,5,10),
#         plot.ylim = NULL,
#         lambda.min.quantile = 0.5,
#         n.lambda = 10,
#         K = 100,
#         psub = 0.5,
#         manhattan.type = c("c", "r")[2],
#         plot.name = "Test",
#         plot.type = "jpg",
#         plot.dpi = 300)
#' A function
```

Index