Package ‘ldlasso’

February 20, 2015

Type     Package
Title    LD LASSO Regression for SNP Association Study
Version  3.2
Date     Jan. 2, 2013
Author   Samuel G. Younkin
Maintainer Samuel G. Younkin <samuel.younkin@gmail.com>
Description ldlasso requires data be of class gwaa.data from the the
package GenABEL
License  GPL (>= 2)
Depends  R (>= 2.6.0), GenABEL, quadprog
LazyLoad yes
Repository CRAN
Date/Publication 2013-01-03 08:23:20
NeedsCompilation no

R topics documented:

ldlasso-package .............................................. 2
block.bounds ............................................... 3
block.cood .................................................. 4
block.map.matrix ........................................... 4
block.obj ................................................... 5
chi2 ......................................................... 6
get.cp ....................................................... 6
get.s1 ....................................................... 7
heatmap ..................................................... 9
ldlasso.obj ............................................... 10
ld_lasso .................................................. 10
ld_lasso_method .......................................... 12
p0tos1 .................................................... 14
plot_beta .................................................. 15
plot_ldlasso ............................................... 16
**Description**

Performs a modification of the fused LASSO for SNP subset selection in genetic association studies. The LD LASSO requires densely spaced SNP data and haplotype block structure.

**Details**

The ldlasso package is a collection of functions designed to apply the LD LASSO methodology as described in Younkin and Rao 2010. Functions useful for displaying the results are also included.

```r
Package: ldlasso
Version: 2.0
Date: February 1, 2011
```

**Note**

See the vignette 'ldlasso_vignette.pdf' in ldlasso package directory doc/.

For a package useful for creating haplotype block boundaries see MATILDE by Pattaro et al., available at http://astor.som.jhmi.edu/~gp/software/matilde/

The following function may be used to create block boundaries with MATILDE.

```r
find.bounds <- function( block.obj, prob.threshold = 0.95, B = 1e4, frac = 2 ) {
  Xa <- as.double.snp.data(block.obj@gtdata)
  X1 <- ifelse( Xa == 0, "A", "B" )
  X2 <- ifelse( Xa == 2, "B", "A" )
  X <- makeGenotypes(matrix(paste(X1,X2,sep="/"), byrow = F, ncol = dim(Xa)[2]))
  LD.obj <- LD(X)
  null.X <- null.density(X)
  nXr <- data.frame(null.X$x, null.X$yR)
  names(nXr) <- c("x", "y")
  matilde.obj <- matilde(LD.obj$r^2, nXr, B=B)
  n <- length(matilde.obj$L)
  b.vec <- apply(matilde.obj$b[(n/frac):n], 2, mean)
  block.cood <- as.numeric((b.vec >= prob.threshold))
  return( list( matilde.obj = matilde.obj, prob.threshold = prob.threshold, block.cood = block.cood ) )
}
```
block.bounds

**Author(s)**

Samuel G. Younkin  
<samuel.younkin@gmail.com>

**References**


**See Also**

ld_lasso_method, ld_lasso, plot_ldlasso, plot_beta, heatmap

**Examples**

```r
# Load example data
data(ldlasso_example)
ls()

plot_ldlasso(ldlasso.obj)
heatmap(ldlasso.obj)
```

---

**block.bounds**  
*Creates a vector of haplotype block boundary basepair positions.*

**Description**

This function takes a map vector and a block coordinate vector and creates a vector of block boundary positions for use when plotting block boundaries, as in plot_ldlasso.

**Usage**

```r
block.bounds(map, block.cood)
```

**Arguments**

- `map`  
a vector of SNP map positions in basepairs (or kilo-basepairs)

- `block.cood`  
A vector of length p+1, where p is the number of SNPs. block.cood is an indicator vector that indicates block boundaries at all p+1 SNP bounded intervals. Use find.bounds to create this vector.

**Value**

A vector of block boundary positions in basepairs
Author(s)

Samuel G. Younkin

Examples

data("ldlasso_example")
bmap <- (block.obj@gtdata@map-block.obj@gtdata@map[1])/1e3
block.bounds.vec <- block.bounds( map = bmap, block.cood = block.cood )
block.bounds.vec

Description

This is an indicator vector for haplotype block boundaries. If there are p SNPs in the data then this vector will be of length p+1. Each element in the vector indicates whether the inter-SNP interval is a block boundary. The first and last element must always be one.

Usage

block.cood

Format

A vector containing 86 entries

References


Description

Simple function that maps the block boundary vector to an indicator matrix for use in the definition of constraint matrix. This matrix ensures that only within block SNP pairs are considered.

Usage

block.map.matrix(block.cood)
Arguments

block.cood  A vector of length p+1, where p is the number of SNPs. block.cood is an indicator vector that indicates block boundaries at all p+1 SNP bounded intervals. Use find.bounds to create this vector.

Value

A matrix of logical variables. If the (i,j) entry is TRUE than SNP i and SNP j are in the same haplotype block.

Author(s)

Samuel G. Younkin

See Also

ld_lasso

---

block.obj  GenABEL Object with Genotype and Phenotype Data

Description

This object is of class `gwaa.data` from the package GenABEL. It contains both genotype and phenotype data in slots 'gtdata' and 'phdata' respectively. It is a small portion of a case-control genome-wide association study of Late-Onset Alzheimer's Disease performed at Mayo Clinic Florida.

Usage

block.obj

Format

An object of class `gwaa.data` from the package GenABEL

References


chi2

Create a vector of chi-squared values from the allelic test of association

Description

One degree of freedom Chi-squared test of allelic counts

Usage

chi2(Xa, Y)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xa</td>
<td>A matrix of genotype values coded as 0, 1 or 2 for homozygous major, heterozygous, or homozygous minor, respectively. Rows are subjects and columns are SNPs. May be created with method as.double.snp.data from the GenABEL package.</td>
</tr>
<tr>
<td>Y</td>
<td>A vector of disease phenotypes, diseased = 1, non-diseased = 0</td>
</tr>
</tbody>
</table>

Value

This function returns a vector of chi-squared, one degree of freedom, statistics for the allelic test of association. The length of this vector is the number of SNPs.

Author(s)

Samuel G. Younkin

---

get.cp

Uses bootstrap sampling over a vector of LD LASSO constraint parameters, s2, to compute a vector of cp estimates.

Description

The vector of cp estimates is used to identify the cp-optimal solution.

Usage

get.cp(s2low, s2high, s2.vec.length, block.obj, Xa = NA, Y = NA, s1, r2.cut, block.cood, B = 20)
get.s1

Arguments

s2low  The lower limit for the s2 vector.
s2high  The upper limit for the s2 vector.
s2.vec.length  The number of exponentially spaced values in the s2 vector.
block.obj  An object of class gwaa.data from GenABEL.
xa  If block.obj is NA then a genotype matrix must be provided. Xa is a matrix of genotype values codes as 0, 1 or 2 for homozygous major, heterozygous, or homozygous minor, respectively.
y  If block.obj is NA then a phenotype vector Y must be provided. Y is a vector of diagnoses, where 0 is non-diseased and 1 is diseased.
s1  The LASSO parameter
r2.cut  Only SNP pairs with correlation greater than r2.cut are bounded by the LD LASSO constraint.
block.cood  A vector of length p+1, where p is the number of SNPs. block.cood is an indicator vector that indicates block boundaries at all p+1 SNP bounded intervals. Use find.bounds to create this vector.
b  Number of bootstrap samples

Value

s2.vec  A vector of s2 values
cp.vec  A vector of cp estimates
beta0.mat  A matrix of LD LASSO estimates
s1  The LASSO parameter

Author(s)

Samuel G. Younkin

See Also

ld_lasso_method

Description

get.s1  Finds an estimate for s1 based on an estimate of the expected false positive rate. This function is a wrapper for p0tos1.

This function calls the function p0tos1, which finds an estimate for s1, given p0, where p0 is the number of nonzero estimates found with a permuted phenotype vector.
Usage

get.s1(i, block.obj, Xa = NA, Y = NA, r2.cut, s1high, s1low, max.iter, tol)

Arguments

i  Number of nonzero estimates in permuted model. If the expected false positive rate is desired to be 0.10, then set $i = 0.10 \times p$, where $p$ is the number of SNPs.

block.obj  An object of class gwaa.data from GenABEL.

Xa  If block.obj is NA then a genotype matrix must be provided. Xa is a matrix of genotype values codes as 0, 1 or 2 for homozygous major, heterozygous, or homozygous minor, respectively.

Y  If block.obj is NA then a phenotype vector Y must be provided. Y is a vector of diagnoses, where 0 is non-diseased and 1 is diseased.

r2.cut  Value for the cutoff value of r-squared. Can be found with r2.cut.fn.

s1high  The initial upper limit in the bisection algorithm

s1low  The lower upper limit in the bisection algorithm

max.iter  The maximum number of iterations allowed in the bisection algorithm before NA is returned

tol  If $p$ never equals $i$ the bisection algorithm stops when $|s1.old - s1.new| < tol$. Otherwise algorithm stops when $p = i$.

Details

This function increases the value for $p_0$ if NA is returned from p0tos1. The bisection algorithm is contained in the function p0tos1

Value

Given the number of nonzero SNPs allowed with permuted phenotype vector, $p_0$, this function returns an estimate for $s_1$.

Author(s)

Samuel G. Younkin

See Also

p0tos1
heatmap

Plots the heat map for the LD LASSO solution space

Description

This function uses the matrix beta0.mat, computed by the function ld_lasso_method, to plot a heat map of LD LASSO solutions. Increasing darkness corresponds to increasing values for the SNP LD LASSO estimate.

Usage

heatmap( ldlasso.obj, s2.indx )

Arguments

ldlasso.obj An ld lasso object
s2.indx s2

Details

See vignette.

Value

A heat map is a visual representation of the matrix of LD LASSO estimates. Rows correspond to s2 values and columns correspond to SNP indices. The magnitude of the SNP estimate is represented by the degree to which the element is shaded. Increasing darkness corresponds to increasing magnitude.

Author(s)

Samuel G. Younkin

See Also

plot_ldlasso, plot_beta, image
**ldlasso.obj**

*LD LASSO Object created by the function ld_lasso_method*

**Description**

This is a data frame created by the function ld_lasso_method. It contains information about the parameters used to generate the three LD LASSO solutions, as well as the solutions themselves. It also contains the results from the grid search performed to estimate the cp-optimal solution. (see cp.obj) Some basic map information is also stored here for use when plotting.

**Usage**

ldlasso.obj

**Format**

data.frame

- **beta1** numeric cp-optimal solution
- **beta2** numeric fused solution
- **beta3** numeric un-fused solution
- **s2star** numeric s2 parameter used for cp-optimal solution
- **cp.obj** data.frame s2 and cp vectors along with solution matrix used to find s2star
- **log10p** numeric log base 10 of p-values from allelic test of association for each SNP
- **bpmap** numeric vector of physical SNP positions in kb from left boundary
- **block.bounds.vec** numeric position of block boundaries in kb from left boundary (does not include endpoints)
- **s1** numeric LASSO parameter
- **B** numeric number of bootstrap samples used by ld_lasso_method to find s2star
- **s2.vec.length** numeric length of vector of s2 values used in ld_lasso_method

**References**


**ld_lasso**

*The main function of the LD LASSO method*

**Description**

The LD LASSO uses the correlation of SNP genotypes in a penalized least squares regression framework. The estimator is the solution to a convex optimization problem, and here we use the solution from the package quadprog.
Usage

\texttt{ld\_lasso(block.obj, block.cood = NA, Xa = NA, Y = NA, s1, s2, r2.cut = 0.5, delta = 1e-10, form = 3, ytype = 1, solve = TRUE )}

Arguments

- \texttt{block.obj} An object of class \texttt{gwaa.data} from \texttt{GenABEL}.
- \texttt{block.cood} A vector of length \(p+1\), where \(p\) is the number of SNPs. \texttt{block.cood} is an indicator vector that indicates block boundaries at all \(p+1\) SNP bounded intervals. Use \texttt{find.bounds} to create this vector.
- \texttt{Xa} If \texttt{block.obj} is \texttt{NA} then a genotype matrix must be provided. \texttt{Xa} is a matrix of genotype values codes as 0, 1 or 2 for homozygous major, heterozygous, or homozygous minor, respectively.
- \texttt{Y} If \texttt{block.obj} is \texttt{NA} then a phenotype vector \texttt{Y} must be provided. \texttt{Y} is a vector of diagnoses, where 0 is non-diseased and 1 is diseased.
- \texttt{s1} The LASSO constraint parameter – the sum of the magnitude of the estimates is bounded by \(s1\).
- \texttt{s2} The LD LASSO constraint parameter – the absolute difference of SNP pair estimates is bounded by \(s2 \log(r^2)\) times the log of \(r^2\)
- \texttt{r2.cut} Only SNP pairs with correlation greater than \(r^2.cut\) are bounded by the LD LASSO constraint.
- \texttt{delta} Included so that optimization is numerically feasible in cases when \(r^2 = 1\)
- \texttt{form} Form of constraint matrix. \texttt{form} is either 1, 2 or 3: 1 for \texttt{cpcc.vec <- 1e6*rep(1,length(r2))} – LASSO solution 2 for \texttt{cpcc.vec <- -s2*log(r2) + delta, s1 <- 1e6} – LD fused solution 3 for \texttt{cpcc.vec <- -s2*log(r2) + delta} – LD LASSO
- \texttt{ytype} If \texttt{ytype} is 1 then \texttt{Y} is a vector of binary disease phenotypes, 0 for non-disease, 1 for diseased. If \texttt{ytype} is 2 then \texttt{Y} is the normalized log OR.
- \texttt{solve} logical variable indicating whether or not to solve regression problem. Useful when \texttt{ld\_lasso} is used to construct constraint matrix, and the solution is not necessary, as in the selection of the \(r^2\) cutoff.

Details

This function performs the ld lasso regression with parameters \(s1, s2\) and \(r^2.cut\) on \texttt{block.obj} with haplotype block boundaries defined by \texttt{block.cood}.

Value

- \texttt{qp} List from the function \texttt{solve.QP} in the package \texttt{quadprog}. This object contains the solutions for \(c(\beta, \beta^+, \beta^-)\) and so the LD LASSO estimates are the first \(p\) elements, or \texttt{qp$solution[1:p]}.
- \texttt{y} A vector of normalized log OR
- \texttt{A} The constraint matrix
The vector of r-squared values used to define constraint matrix. Elements in this vector are the correlation estimates for all inter-block SNP pairs.

The vector of ld lasso constraints with length 3p

A vector of odds ratios

Author(s)

Samuel G. Younkin

References


See Also

ld_lasso_method, quadprog, solve.QP

Examples

data("ldlasso_example")
ldlasso.test <- ld_lasso( block.obj, block.cood, s1 = 1, s2 = 0.5 )

ld_lasso_method

The LD LASSO function

Description

This function implements a method for the automatic selection of parameters for the LD LASSO. It returns three solutions, the fused, cp-optimal, and unfused solutions. It also creates a matrix of solutions needed for creating the trace plot.

Usage

ld_lasso_method(block.obj, block.cood = NA, Xa = NA, Y = NA, bmap = NA, maxcol = 5e3, p.frac = 0.10, B = 5, s2low = 5e-3, s2high = 5e1, s2.vec.length = 4, null = FALSE)

Arguments

block.obj An object of class gwaa.data from GenABEL.
block.cood A vector of length p+1, where p is the number of SNPs. block.cood is an indicator vector that indicates block boundaries at all p+1 SNP bounded intervals. Use find.bounds to create this vector.
Xa If block.obj is NA then a genotype matrix must be provided. Xa is a matrix of genotype values codes as 0, 1 or 2 for homozygous major, heterozygous, or homozygous minor, respectively.
Y  If block.obj is NA then a phenotype vector Y must be provided. Y is a vector of diagnoses, where 0 is non-diseased and 1 is diseased.

bpmap  A vector of map positions in terms of kB from the left boundary

maxcol  The upper limit on the number of columns in the constraint matrix. This is to prevent computational overload. Increasing maxcol may increase computation time and memory needed.

p.frac  The fraction of SNPs allowed in LASSO model under null hypothesis. This parameter is used in the function get.s1.

B  The number of bootstrap iterations for cp estimate.

s2low  The lower limit of the s2 vector

s2high  The upper limit of the s2 vector

s2: vec.length  The number of exponentially spaced values of s2.

null  A logical variable that indicates if analysis should be performed on permuted phenotype vector.

Details

Use function find.bounds to create block.cood vector with MATILDE MCMC methods. See ldlasso help page for details on the package MATILDE.

Value

beta1  The ld lasso solution with s2 that minimizes cp (cp-optimal solution)

beta2  The ld lasso solution for lower limit of s2 interval (fused solution)

beta3  The ld lasso solution for upper limit of s2 interval (unfused solution)

s2star  The s2 value that minimizes cp

cp.obj  A list that contains information used for cp estimation

log10p  A vector of log10 p values for test of allelic association

bpmap  A vector of map positions in base pairs

block.bounds.vec  A vector of block boundaries in kB from left boundary

s1  the LASSO constraint

b  Number of bootstrap samples

s2: vec.length  The length of the s2 vector

Author(s)

Samuel G. Younkin

References

See Also

ld_lasso

Examples

# Load example data
# data(ldlasso_example)

# Run the method with low B and s2.vec.length first to test.
# ldlasso.obj <- ld_lasso_method( block.obj, block.cood, B = 3, s2.vec.length = 2 )

p0tos1

Finds the LASSO parameter s1 that corresponds to desired false positive rate

Description

Estimates the LASSO parameter that corresponds to desired false positive rate. More specifically, it is a bisection algorithm designed to find an s1 that corresponds to p0 nonzero estimates from the ld lasso with a permuted phenotype vector.

Usage

p0tos1(p0, block.obj, Xa = NA, Y = NA, r2.cut = 0.01, s1high, s1low, max.iter = 100, tol = 0.1)

Arguments

p0 The number of nonzero estimates in permuted model. If the expected false positive rate is set at 0.10, then i = 0.10*p, where p is the number of SNPs.
block.obj An object of class gwaa.data from GenABEL.
Xa If block.obj is NA then a genotype matrix must be provided. Xa is a matrix of genotype values codes as 0, 1 or 2 for homozygous major, heterozygous, or homozygous minor, respectively.
Y If block.obj is NA then a phenotype vector Y must be provided. Y is a vector of diagnoses, where 0 is non-diseased and 1 is diseased.
r2.cut The value for the cutoff value of r-squared. Can be found with r2.cut.fn.
s1high The initial upper limit in the bisection algorithm
s1low The initial lower upper limit in the bisection algorithm
max.iter The maximum number of iterations allowed in the bisection algorithm before NA is returned
tol If p never equals i the bisection algorithm stops when |s1.old - s1.new| < tol. Otherwise algorithm stops when p =i.
plot_beta

Value
An estimate for s1

Author(s)
Samuel G. Younkin

See Also
get.s1

plot_beta : Plot the LD LASSO solution by physical position.

Description
Creates a plot of LD LASSO estimates using the physical map, bpmap

Usage
plot_beta( beta, bpmap=NA, block.bounds.vec=NA )

Arguments
beta A vector of ldlasso estimates.
bpmap A vector of map positions in terms of kb from the left boundary
block.bounds.vec A vector of block boundaries positions in kilobases from the left boundary. Boundaries are plotted at the midpoint of SNP intervals.

Details
Includes haplotpye block boundaries as vertical broken lines.

Value
A plot object.

Author(s)
Samuel G. Younkin

See Also
plot_ldlasso
plot_ldlasso  

*Diagnostic plot for an ldlasso object*

**Description**

Creates four plots; Top-left: Allelic test of association; Top-right: cp-optimal, fused, and unfused solutions (black, red, green); Bottom-left: Trace plot with fixed s1; Bottom-right: cp estimates vs. s2

**Usage**

```r
plot_ldlasso( ldlasso.obj )
```

**Arguments**

- `ldlasso.obj`  
  An ld lasso object

**Details**

Note that often two solutions overlap, and one is therefore concealed. The vertical line in the cp plot is drawn at the value of s2 used in the cp-optimal solution.

**Value**

A plot object

**Author(s)**

Samuel G. Younkin

**See Also**

ld_lasso_method, ld_lasso

---

r2.cut.fn  

*Find the r-squared cutoff*

**Description**

Returns the minimum value of r2 in r2.vec that keeps the number of constraint matrix columns less than maxcol.

**Usage**

```r
r2.cut.fn(block.obj, block.cood, Xa, Y, maxcol, r2.cut.min, r2.cut.max, r2.vec.length)
```
Arguments

block.obj  An object of class gwaa.data from GenABEL.

block.cood  A vector of length p+1, where p is the number of SNPs. block.cood is an indicator vector that indicates block boundaries at all p+1 SNP bounded intervals. Use find.bounds to create this vector.

Xa  If block.obj is NA then a genotype matrix must be provided. Xa is a matrix of genotype values codes as 0, 1 or 2 for homozygous major, heterozygous, or homozygous minor, respectively.

Y  If block.obj is NA then a phenotype vector Y must be provided. Y is a vector of diagnoses, where 0 is non-diseased and 1 is diseased.

maxcol  The maximum number of columns allowed in the constraint matrix. Default is 5,000.

r2.cut.min  The lower bound of the r2.cut interval

r2.cut.max  The upper bound of the r2.cut interval

r2.vec.length  The number of r2.cut values uniformly spaced in r2.cut interval

Value

The function r2.cut.fn returns the minimum value of r2.cut that satisfies, ncol(A) < maxcol, where A is the constraint matrix

Author(s)

Samuel G. Younkin

See Also

ld_lasso

---

trace.plot  Creates a trace plot for the LD LASSO

Description

This function creates a trace plot as a function of s2, the LD LASSO constraint

Usage

trace.plot(beta0.mat, s2.vec, type = "l", indx = NA, s2star = NA, abs = TRUE)
Arguments

- **beta0.mat**: A matrix where rows are ld lasso estimates. This matrix has number of rows equal to length(s2.vec) and number of columns equal to the number of SNPs.
- **s2.vec**: The vector of s2 values used to generate the rows of beta0.mat.
- **type**: Type of plot, "l" for lines, "p" for points.
- **indx**: If index is NA then indx, is a vector of SNP indices to be plotted in red.
- **s2star**: Value of s2star from cp object. If s2star is not NA, a vertical line is plotted at this value.
- **abs**: A logical variable indicating whether to plot absolute value of beta instead of beta.

Details

In this trace plot each line represents a SNP estimate and we see the behavior of the solution as the LD LASSO constraint is varied.

Value

Creates a trace plot.

Author(s)

Samuel G. Younkin
Index

*Topic **datasets**
  - block.cood, 4
  - block.obj, 5
  - ldlasso.obj, 10

*Topic **package**
  - ldlasso-package, 2

block.bounds, 3
block.cood, 4
block.map.matrix, 4
block.obj, 5
blockcood(block.cood), 4
blockobj(block.obj), 5

chi2, 6
get.cp, 6
get.s1, 7

heatmap, 9

ld_lasso, 10
ld_lasso_method, 12
ldlasso(ldlasso-package), 2
ldlasso-package, 2
ldlasso.obj, 10
ldlassoobj(ldlasso.obj), 10

p0tos1, 14
plot_beta, 15
plot_ld_lasso(plot_ldlasso), 16
plot_ld_lasso_method(plot_ldlasso), 16
plot_ldlasso, 16

r2.cut.fn, 16

trace.plot, 17