

# Package ‘SummaryLasso’

November 15, 2019

**Type** Package

**Title** Building Polygenic Risk Score Using GWAS Summary Statistics

**Version** 1.2.1

**Author** Ting-Huei Chen

**Maintainer** Ting-Huei Chen <tingstat22@gmail.com>

**Depends** R(>= 3.5.0), gtools

## Description

Shrinkage estimator for polygenic risk prediction models based on summary statistics of genome-wide association studies.

**License** GPL-3

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2019-11-15 12:40:02 UTC

## R topics documented:

funcIndex . . . . .	2
gsfPEN . . . . .	2
gsPEN . . . . .	5
Nvec . . . . .	7
plinkLD . . . . .	8
summaryZ . . . . .	8
<b>Index</b>	<b>10</b>

---

funcIndex	<i>Inputs for the functional annotations of SNPs.</i>
-----------	---

---

### Description

A 3614 x 3 matrix with (0,1) entry with 3614 SNPs and 3 functional annotations. For the element at i-th row, j-th column, the entry 0 means SNP i without j-th functional annotation; entry 1 means otherwise. follows:

- f1: The binary index for functional annotation 1.
- f2: The binary index for functional annotation 2.
- f3: The binary index for functional annotation 3.

### Usage

```
data(summaryZ)
```

### Format

A matrix with 3614 rows for the 3614 SNPs and 3 columns for functional annotations.

---

gsfPEN	<i>SummaryLasso incorporating multiple traits and functional annotations of SNPs.</i>
--------	---

---

### Description

SummaryLasso to model pleiotropy by introducing a group-Lasso type penalty, which is sensitive to select SNPs modestly associated with multiple traits and to incorporate functional annotations of SNPs simultaneously.

### Usage

```
gsfPEN(summaryZ, Nvec, plinkLD, NumIter = 1000, RupperVal = NULL,
breaking = 1, numChrs = 22, ChrIndexBeta = 0, Init_summaryBetas = 0,
Zscale = 1, tuningMatrix = NULL, penalty = "mixLOG", funcIndex,
numfunc, p.Threshold = NULL, p.Thresholdpara = c(0.5, 10^-4, 4),
taufactor = c(1/25, 1, 3), llim_length = 4, subtuning = 4,
Lambda_limit = c(0.5, 0.9), Lenlam = 4, lambdavec_func = NULL,
lambdavec_func_limit_len = c(1.5, 3), dfMax = NULL, outputAll = 0,
warmStart = 0, customed = 0, AllTuningMatrix = NULL, SDvec = NULL,
IniBeta = 0)
```

**Arguments**

summaryZ	The Z statistics of p SNPs from q GWA studies. A matrix with dimension p x q for p SNPs and q traits. The first column corresponds to the primary trait and the rest columns correspond to the secondary traits.
Nvec	A vector of length q for the sample sizes of q GWA studies.
plinkLD	.ld file obtained from the LD calculation from plink.
NumIter	The number of maximum iterations for the estimation procedure.
RupperVal	The maximum tolerable magnitude of the estimates of coefficients during the iterations. This is to avoid certain estimates of coefficients to diverge during the iterations. This may happen when the signs of the correlation coefficients were estimated incorrectly. The default value is 50 times the maximum of coefficients from the input in absolute values.
breaking	A binary (0,1) variable to check if there are some certain estimates of coefficients to diverge during the iterations. This may happen when the signs of the correlation coefficients were estimated incorrectly. The default value is 1.
numChrs	The number of chromosomes used in the analysis. Current version of package does not use this argument.
ChrIndexBeta	The chromosome index for each SNP. Current version of package does not use this argument.
Init_summaryBetas	Can be used to set the initial values of the coefficients for the iterative estimation.
Zscale	A binary (0,1) variable to make the coefficients from different GWA studies with unequal sample sizes comparable. The default value is 1.
tuningMatrix	Inputs for the tuning values of the tuning parameters. Default is null and it will be generated automatically.
penalty	Current version of package does not use this argument.
funcIndex	Inputs for the functional annotations of SNPs. A p x k matrix with (0,1) entry; p is the number of SNPs and k is the number of functional annotations. For the element at i-th row, j-th column, the entry 0 means SNP i without j-th functional annotation; entry 1 means otherwise.
numfunc	The number of functional annotations.
p.Threshold	The p-values threshold to set up the tuning values of the baseline tuning parameter.
p.Thresholdpara	When p.Threshold is null, p.Threshold will be generated automatically based on the values of p.Thresholdpara. The default values are c(0.5, 10 <sup>4</sup> , 5), where the first element is the maximum of the p-value threshold, the second element is the minimum, and the third element is total number of p-value thresholds to be generated from the minimum to the maximum.
taufactor	The weights to generate the tuning values for the tuning parameter "tau" and the default is c(1/25, 1, 10) times the median of the p summation of the coefficients for each SNP across q traits.
llim_length	The argument to set up the number of tuning values for lambdas between the lower and upper bound. The default value is 10.

subtuning	The argument to set up the number of tuning values for lambdas between the lower and upper bound. The default value is 50.
Lambda_limit	The quantiles to set up the tuning values of lambda. The default value is c(0.5, 0.9).
Lenlam	The number of tuning values for lambda parameter without using the Log penalty. In other words, the initial Lenlam rows of the tuningMatrix are for summayLasso single trait analysis.
lambdavec_func	The tuning values for the tuning parameters associated with the functional annotations.
lambdavec_func_limit_len	When lambdavec_func is null, lambdavec_func will be generated automatically based on the arguments of lambdavec_func_limit_len. The default values are c(1.5, 4). The first element is the maximum of the tuning value and the second element is the total number of the tuning values to be generated from 0 to the maximum.
dfMax	The upper bound of the number of non-zero estimates of coefficients for the primary trait.
outputAll	For internal checking usage. The default value is 0.
warmStart	For internal checking usage. The default value is 0.
customed	For internal checking usage. The default value is 0.
AllTuningMatrix	For internal checking usage. The default value is NULL.
SDvec	The matrix of the standard error for regression coefficients. When the input of SummaryZ is at Z scale, let SDvec = NULL and it will be computed internally.
IniBeta	A binary (0,1) variable to indicate if the regression coefficients need to be initialized or not. 1 is for yes.

## Details

Note that the tuning values for the tuning parameters may need to be modified manually when the selected optimal tuning parameters are at the boundary of the inputs.

## Value

BetaMatrix	The output of the coefficients matrix with dimensions (total number of combinations of the tuning values times (pq)). Each column represents the vectorization of the p x q coefficients matrix given a particular combination of the tuning values (stacking its columns into a column vector).
Numitervec	This vector shows the number of iterations to converge for each combination of the tuning values.
AllTuningMatrix	This matrix shows all combination of tuning values used in the estimation process. Its dimension is that total number of combinations of the tuning values times total number of tuning parameters.

**Author(s)**

Ting-Huei Chen

**References**

This R packages is based on the method introduced in the manuscript "A comprehensive statistical framework for building polygenic risk prediction models based on summary statistics of genome-wide association studies."

**Examples**

```
data("summaryZ")
data("Nvec")
data("plinkLD")
data("funcIndex")
output = gsFPEN(summaryZ=summaryZ, Nvec=Nvec, plinkLD=plinkLD, funcIndex=funcIndex,
numfunc=ncol(funcIndex))
```

gsPEN

*SummaryLasso incorporating multiple traits***Description**

SummaryLasso to model pleiotropy by introducing a group-Lasso type penalty, which is sensitive to select SNPs modestly associated with multiple traits.

**Usage**

```
gsPEN(summaryZ, Nvec, plinkLD, NumIter = 100, breaking = 1, numChrs = 22,
ChrIndexBeta = 0, Init_summaryBetas = 0, Zscale = 1, RupperVal = NULL,
tuningMatrix = NULL, penalty = c("mixLOG"), taufactor = c(1/25, 1, 10),
llim_length = 10, subtuning = 50, Lambda_limit = c(0.5, 0.9),
Lenlam_singleTrait = 200, dfMax = NULL, IniBeta = 0, inverseTuning = 0,
outputAll = 0, warmStart = 1)
```

**Arguments**

summaryZ	The Z statistics of p SNPs from q GWA studies. A matrix with dimension p x q for p SNPs and q traits. The first column corresponds to the primary trait and the rest columns correspond to the secondary traits.
Nvec	A vector of length q for the sample sizes of q GWA studies.
plinkLD	.ld file of the LD calculation from plink.
NumIter	The number of maximum iterations for the estimation procedure.
breaking	A binary (0,1) variable to check if there are some certain estimates of coefficients to diverge during the iterations. This may happen when the signs of the correlation coefficients were estimated incorrectly. The default value is 1.

numChrs	The number of chromosomes used in the analysis. Current version of package does not use this argument.
ChrIndexBeta	The chromosome index for each SNP. Current version of package does not use this argument.
Init_summaryBetas	Can be used to set the initial values of the coefficients for the iterative estimation.
Zscale	A binary (0,1) variable to make the coefficients from different GWA studies with unequal sample sizes comparable. The default value is 1.
RupperVal	The maximum tolerable magnitude of the estimates of coefficients during the iterations. This is to avoid a certain estimates of coefficients to diverge during the iterations. This may happen when the signs of the correlation coefficients were estimated incorrectly. The default value is 50 times the maximum of coefficients from the input in absolute values.
tuningMatrix	Inputs for the tuning values of the tuning parameters. Default is null and it will be generated automatically.
penalty	Current version of package does not use this argument.
taufactor	The weights to generate the tuning values for the tuning parameter "tau" and the default is c(1/25, 1, 10) times the median of the p summation of the coefficients for each SNP across q traits.
llim_length	The argument to set up the number of tuning values for lambdas between the lower and upper bound. The default value is 10.
subtuning	The argument to set up the number of tuning values for lambdas between the lower and upper bound. The default value is 50.
Lambda_limit	The quantiles to set up the tuning values of lambda. The default value is c(0.5, 0.9).
Lenlam_singleTrait	The quantiles to set up the tuning values of lambda for single trait analysis.
dfMax	The upper bound of the number of non-zero estimates of coefficients for the primary trait.
IniBeta	A binary (0,1) variable to indicate if the regression coefficients need to be initialized or not. 1 is for yes.
inverseTuning	For internal checking usage. The default value is 0.
outputAll	For internal checking usage. The default value is 0.
warmStart	For analysis with single trait or multiple traits without functional annotations, it is recommended to use warmStart = 1 to enhance computations.

## Details

Note that the tuning values for the tuning parameters may need to be modified manually when the selected optimal tuning parameters are at the boundary of the inputs.

**Value**

BetaMatrix	The output of the coefficients matrix with dimensions (total number of combinations of the tuning values times (pq)). Each column represents the vectorization of the $p \times q$ coefficients matrix given a particular combination of the tuning values (stacking its columns into a column vector).
Numitervec	This vector shows the number of iterations to converge for each combination of the tuning values.
AllTuningMatrix	This matrix shows all combination of tuning values used in the estimation process. Its dimension is that total number of combinations of the tuning values times total number of tuning parameters.

**Author(s)**

Ting-Huei Chen

**References**

This R packages is based on the method introduced in the manuscript "A comprehensive statistical framework for building polygenic risk prediction models based on summary statistics of genome-wide association studies."

**Examples**

```
data("summaryZ")
data("Nvec")
data("plinkLD")
output = gsPEN(summaryZ=summaryZ, Nvec=Nvec, plinkLD=plinkLD)
```

---

Nvec

*A vector of sample sizes for the q traits of the summaryZ.*

---

**Description**

A vector of q sample sizes for the q set of Z statistics corresponding to the q columns of summaryZ.

**Usage**

```
data(Nvec)
```

**Format**

A vector with q elements, where q is the number of columns of summaryZ.

---

 plinkLD

*The LD info from output of the software (plink)*


---

### Description

The LD information is crucial for the analysis by SummaryLasso. The reference alleles used to obtain for the Z statistics or the regression coefficients have to be the same as those used for the LD calculation. This file can be obtained directly from the output of the LD calculation by the software (plink); for example the output can be like plink.ld. On the other hand, the user can calculate the LD based on their preferred tools. The variables are as follows:

- CHR\_A: The chromosome of SNP\_A
- BP\_A: The positions of SNP\_A
- SNP\_A: The names of SNP\_A
- CHR\_B: The chromosome of SNP\_B
- BP\_B: The positions of SNP\_B
- SNP\_B: The names of SNP\_B
- R: The correlation between SNP\_A and SNP\_B

### Usage

```
data(plinkLD)
```

### Format

A data frame with 205959 rows and 7 columns

### References

- Purcell S, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81**.

---

 summaryZ

*The Z statistics from the univariate analysis of the association between 3614 SNPs and three traits respectively.*


---

### Description

These Z statistics are obtained from simulated datasets. The variables are as follows:

- Z1: The Z statistics from trait 1; the primary trait.
- Z2: The Z statistics from trait 2; the secondary trait.
- Z2: The Z statistics from trait 3; the secondary trait.



*summaryZ*

9

**Usage**

```
data(summaryZ)
```

**Format**

A matrix with 3614 rows for the 3614 SNPs and 3 columns for 3 traits.

# Index

## \*Topic **datasets**

funcIndex, [2](#)

Nvec, [7](#)

plinkLD, [8](#)

summaryZ, [8](#)

funcIndex, [2](#)

gsfPEN, [2](#)

gsPEN, [5](#)

Nvec, [7](#)

plinkLD, [8](#)

summaryZ, [8](#)