Package 'LBL'

July 18, 2019

Type Package

Title Logistic Bayesian Lasso for identifying genetic association between common diseases and rare haplotypes

Version 0.9.0

Description LBL uses Bayesian Lasso to detect rare haplotypes that are associated with common diseases. The current implementation considers dichotomous traits. A future release will include quantitative and survival traits. LBL is capable of handling different study designs: this version of the software is capable of handling independent cases and controls, case-parent trios and a mixture of both (provided that the family data is independent of the case-control data). LBL uses algorithms from pre.hapassoc function from hapassoc package to obtain all possible haplotype pairs compatible with an indivdiual's set of genotypes.

Author Swati Biswas, Meng Wang, Xiaofei Zhou, Han Zhang, Shuang Xia, Yuan Zhang, and Shili Lin

Maintainer Shili Lin <shili@stat.osu.edu>

Depends R (>= 3.0.0)

License GPL-3

Encoding UTF-8

LazyData true

Imports stats, utils

Suggests knitr, rmarkdown

VignetteBuilder knitr, rmarkdown

RoxygenNote 6.1.1

RemoteType github

RemoteHost api.github.com

RemoteRepo LBL

RemoteUsername mxw010

RemoteRef master

RemoteSha 50f17ff5e717052b52c2b3ae108b87223dffa647

GithubRepo LBL

GithubUsername mxw010

GithubRef master

GithubSHA1 50f17ff5e717052b52c2b3ae108b87223dffa647

NeedsCompilation yes

R topics documented:

`LBL-package` *LBL: Logistic Bayesian Lasso for Detecting Rare (or Common) Haplotype Association*

Description

LBL uses the Bayesian LASSO framework to detect association between a phenotype and haplotypes given the (unphased) genotypes of individuals.

Details

LBL uses Bayesian Lasso to detect rare haplotypes that are associated with common diseases. The current implementation considers dichotomous traits. A future release will include quantitative and survival traits. LBL is capable of handling different study designs: this version of the software is capable of handling independent cases and controls, case-parent trios and a mixture of both (provided that the family data is independent of the case-control data).

A function from the hapassoc package is first used to acquire all compatible haplotypes. The posterior samples are then obtained via Markov Chain Monte Carlo (MCMC) algorithm and inference on the parameters of interest can be carried out (Bayes Factor, Credible Interval, etc.) based on these posterior samples.

Functions

`LBL`: MCMC algorithm to obtain posterior samples for independent case-control data. `famLBL`: MCMC algorithm to obtain posterior samples for case-parent trio data. `cLBL`: MCMC algorithm to obtain posterior samples for combined data.

`LBL_summary` provides model summary (in the form of list) based on posterior samples. `print_LBL_summary` prints model summary in a user-friendly format from the list result of `LBL_summary`.

Author(s)

Swati Biswas, Meng Wang, Xiaofei Zhou, Han Zhang, Shuang Xia, Yuan Zhang, and Shili Lin <shili@stat.osu.edu>

References

Biswas S. and Lin S. (2012). Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics*, 68(2): 587-97.

Wang, M. and Lin, S. (2014). FamLBL: detecting rare haplotype disease association based on common SNPs using case-parent triads. *Bioinformatics*, 30(18), 2611-2618.

Zhou, X., Wang, M., and Lin, S. (2019). cLBL: Combined logistic Bayesian LASSO for detecting rare associated haplotypes using independent case, control and family trio data. *Manuscript*.

`cac` *An example file consisting of independent cases/controls*

Description

A dataset containing phenotypes and genotypes of 5 SNPs for 500 independent cases and controls, and is presented in a pedigree format for easier integration with family data.

Usage

```
data(cac)
```

Format

A data frame with 500 rows (250 cases and 250 controls) and 16 (= 6 fixed columns + 2 * # of SNPs) columns:

column 1 family ID, integers

column 2 individual ID, integers

column 3 father's ID, 0 for unknown father.

column 4 mother's ID, 0 for unknown mother.

column 5 individual's sex: 1 = male and 2 = female.

column 6 affection status: 0 = unknown, 1 = unaffected, and 2 = affected.

column 7 - 16 marker genotypes. Each SNP is represented by two columns: one for each allele.

`cLBL` *Bayesian Lasso for Detecting Rare (or Common) Haplotype Association in Population and Family Based Studies*

Description

`cLBL` is an MCMC algorithm that generates posterior samples for a dataset containing both case-control and family trio designs. This function takes standard pedigree format as input. The input does not allow missing observations and subjects with missing data are removed. The function returns an object containing posterior samples after the burn-in period.

Usage

```
cLBL(data.fam, data.cac, input.freq, baseline = "missing", a = 15,
  b = 15, start.beta = 0.01, lambda = 1, D = 0, seed = NULL,
  burn.in = 10000, num.it = 40000, summary = TRUE, e = 0.1,
  ci.level = 0.95)
```

Arguments

`data.fam`	The family portion of data. This dataset should consist of "3n" rows and 6+2*p columns, where n is the number of case-parent trios, and p is the number of SNPs. The data should be in standard pedigree format, with the first 6 columns representing the family ID, individual ID, father ID, mother ID, sex, and affection status. The other 2*p columns are genotype data in allelic format, with each allele of a SNP taking up one column. An example can be found in this package under the name `"fam"`. For more information about the format, type `"?fam"` into R, or see "Linkage Format" section of https://www.broadinstitute.org/haploview/input-file-formats.
`data.cac`	The case-control portion of data. This dataset should consist of "n" rows and 6+2*p columns, where n is the number of individuals of the independent cases and controls, and p is the number of SNPs. The data should be in standard pedigree format, with the first 6 columns representing the family ID, individual ID, father ID, mother ID, sex, and affection status. The other 2*p columns are genotype data in allelic format, with each allele of a SNP taking up one column. An example can be found in this package under the name `"cac"`. For more information about the format, type `"?cac"` into R, or see "Linkage Format" section of https://www.broadinstitute.org/haploview/input-file-formats.
`input.freq`	Optional. Specify frequency distribution of haplotypes. If not provided, the algorithm will use the estimated frequencies.
`baseline`	Haplotype to be used for baseline coding; default is the most frequent haplotype according to the initial haplotype frequency estimates. This argument should be a character, starting with an h and followed by the baseline haplotype.
`a`	First hyperparameter of the prior for regression coefficients, β. The prior variance of β is $2/\lambda^2$ and λ has Gamma(a,b) prior. The Gamma prior parameters a and b are formulated such that the mean and variance of the Gamma distribution are a/b and a/b^2. The default value of a is 15.
`b`	Second hyperparameter of the Gamma(a,b) distribution described above; default is 15.
`start.beta`	Starting value of all regression coefficients, β; default is 0.01.
`lambda`	Starting value of the λ parameter described above; default is 1.
`D`	Starting value of the D parameter, which is the within-population inbreeding coefficient; default is 0.
`seed`	Seed to be used for the MCMC in Bayesian Lasso; default is a random seed. If exact same results need to be reproduced, seed should be fixed to the same number.
`burn.in`	Burn-in period of the MCMC sampling scheme; default is 10000.
`num.it`	Total number of MCMC iterations including burn-in; default is 40000.
`summary`	Logical. If TRUE, cLBL will return a summary of the analysis. If FALSE, cLBL will return the posterior samples of MCMC. Default is set to be TRUE.

e — A (small) number ϵ in the null hypothesis of no association, $H_0 : |\beta| \leq \epsilon$. The default is 0.1. Changing e from the default of 0.1 may necessitate choosing a different threshold for Bayes Factor (one of the outputs) to infer association. Only used if `summary = TRUE`.

ci.level — Credible probability. The probability that the true value of β will be within the credible interval. Default is 0.95, which corresponds to a 95% posterior credible interval. Only used if `summary = TRUE`.

Value

If `summary = FALSE`, return a list with the following components:

haplotypes The list of haplotypes used in the analysis. The last column is the reference haplotype.

beta Posterior samples of betas stored in a matrix.

lambda A vector of (num.it-burn.in) posterior samples of lambda.

freq Posterior samples of the frequencies of haplotypes stored in a matrix format, in the same order as haplotypes.

init.freq The haplotype distribution used to initiate the MCMC.

If `summary = TRUE`, return the result of LBL_summary. For details, see the description of the `LBL_summary` function.

See Also

LBL, famLBL, LBL_summary, print_LBL_summary, LBL-package.

Examples

```
data(fam)
data(cac)
combined.obj<-cLBL(data.fam=fam,data.cac=cac)
combined.obj
print_LBL_summary(combined.obj)
```

---

fam — *An example file of case-parent trios*

Description

A dataset containing the pedigree information, phenotypes and genotypes of 250 case-parent trios in pedigree format.

Usage

```
data(fam)
```

Format

A data frame with 750 rows (250 trios) and 16 (= 6 fixed columns + 2 * # of SNPs) columns:

column 1 family ID, integers

column 2 individual ID, integers

column 3 father's ID, 0 for unknown father.

column 4 mother's ID, 0 for unknown mother.

column 5 individual's sex. 1 = male and 2 = female.

column 6 affection status. 0 = unknown, 1 = unaffected and 2 = affected.

column 7 - 16 marker genotypes. Each marker is represented by two columns: one for each allele.

`famLBL` *Bayesian Lasso for Detecting Rare (or Common) Haplotype Association in Case-Parent Triad Designs*

Description

`famLBL` is an MCMC algorithm that generates posterior samples for family trio data. This function takes standard pedigree format as input. The input does not allow missing observations and subjects with missing data are removed. The function returns an object containing posterior samples after the burn-in period.

Usage

```
famLBL(data.fam, baseline = "missing", start.beta = 0.01, lambda = 1,
  D = 0, seed = NULL, a = 15, b = 15, burn.in = 10000,
  num.it = 40000, summary = TRUE, e = 0.1, ci.level = 0.95)
```

Arguments

`data.fam`	The input data. It should consist of "3n" rows and 6+2*p columns, where n is the number of of case-parent trios, and p is the number of SNPs. The data should be in standard pedigree format, with the first 6 columns representing the family ID, individual ID, father ID, mother ID, sex, and affection status. The other 2*p columns are genotype data in allelic format, with each allele of a SNP taking up one column. An example can be found in this package under the name `"fam"`. For more information about the format, type `"?fam"` into R, or see "Linkage Format" section of https://www.broadinstitute.org/haploview/input-file-formats.
`baseline`	Haplotype to be used for baseline coding; default is the most frequent haplotype according to the initial haplotype frequency estimates. This argument should be a character, starting with an h and followed by the baseline haplotype.
`start.beta`	Starting value of all regression coefficients, β; default is 0.01.
`lambda`	Starting value of the λ parameter described above; default is 1.
`D`	Starting value of the D parameter, which is the within-population inbreeding coefficient; default is 0.

seed	Seed to be used for the MCMC in Bayesian Lasso; default is a random seed. If exact same results need to be reproduced, seed should be fixed to the same number.
a	First hyperparameter of the prior for regression coefficients, β. The prior variance of β is $2/\lambda^2$ and λ has Gamma(a,b) prior. The Gamma prior parameters a and b are formulated such that the mean and variance of the Gamma distribution are a/b and a/b^2. The default value of a is 15.
b	Second hyperparameter of the Gamma(a,b) distribution described above; default is 15.
burn.in	Burn-in period of the MCMC sampling scheme; default is 10000.
num.it	Total number of MCMC iterations including burn-in; default is 40000.
summary	Logical. If TRUE, famLBL will return a summary of the analysis. If FALSE, famLBL will return the posterior samples of MCMC. Default is set to be TRUE.
e	A (small) number ϵ in the null hypothesis of no association, $H_0 : \|\beta\| \leq \epsilon$. The default is 0.1. Changing e from the default of 0.1 may necessitate choosing a different threshold for Bayes Factor (one of the outputs) to infer association. Only used if `summary = TRUE`.
ci.level	Credible probability. The probability that the true value of $beta$ will be within the credible interval. Default is 0.95, which corresponds to a 95% posterior credible interval. Only used if `summary = TRUE`.

Value

If `summary = FALSE`, return a list with the following components:

haplotypes The list of haplotypes used in the analysis. The last column is the reference haplotype.

beta Posterior samples of betas stored in a matrix.

lambda A vector of (num.it-burn.in) posterior samples of lambda.

freq Posterior samples of the frequencies of haplotypes stored in a matrix format, in the same order as haplotypes.

init.freq The haplotype distribution used to initiate the MCMC.

If `summary = TRUE`, return the result of LBL_summary. For details, see the description of the `LBL_summary` function.

See Also

`LBL`, `cLBL`, `LBL_summary`, `print_LBL_summary`, `LBL-package`.

Examples

```
data(fam)
fam.obj<-famLBL(fam)
fam.obj
print_LBL_summary(fam.obj)
```

LBL *Logistic Bayesian Lasso for Detecting Rare (and Common) Haplotyptic Association in Population Based Designs*

Description

LBL is a Bayesian LASSO method developed to detect association between common/rare haplotypes and dichotomous disease phenotype, based on MCMC algorithm. This function will handle independent case/control study design. For other types of study designs, see famLBL and cLBL. This function takes standard pedigree format as input with an individual's genotypes, phenotype and familiar relationships. The input does not allow missing observations, and therefore subjects with missing data are removed. This function returns an object containing posterior samples after the burn-in period.

Usage

```
LBL(data.cac, baseline = "missing", a = 15, b = 15,
  start.beta = 0.01, lambda = 1, D = 0, seed = NULL,
  burn.in = 10000, num.it = 40000, summary = T, e = 0.1,
  ci.level = 0.95)
```

Arguments

data.cac	Input data. data.cac should be either a data frame or a matrix, consisting of "n" rows and 6+2*p columns, where n is the number of cases and controls, and p is the number of SNPs. The data should be in standard pedigree format, with the first 6 columns representing the family ID, individual ID, father ID, mother ID, sex, and affection status. The other 2*p columns are genotype data in allelic format, with each allele of a SNP taking up one column. An example can be found in this package under the name "cac". For more information about the format, type "?cac" into R, or see "Linkage Format" section of https://www.broadinstitute.org/haploview/input-file-formats. Note that since these are independent case-control data, the father ID and mother ID are missing (coded as 0) and each individual has an unique family ID.
baseline	Haplotype to be used for baseline coding; default is the most frequent haplotype according to the initial haplotype frequency estimates. This argument should be a character, starting with an h and followed by the SNPs at each marker locus, for example, if the desired baseline haplotype is 0 1 1 0 0, then baseline should be coded as "h01100".
a	First hyperparameter of the prior for regression coefficients, β. The prior variance of β is $2/\lambda^2$ and λ has Gamma(a,b) prior. The Gamma prior parameters a and b are formulated such that the mean and variance of the Gamma distribution are a/b and a/b^2. The default value of a is 15.
b	Second hyperparameter of the Gamma(a,b) distribution described above; default is 15.
start.beta	Starting value of all regression coefficients, β; default is 0.01.
lambda	Starting value of the λ parameter described above; default is 1.
D	Starting value of the D parameter, which is the within-population inbreeding coefficient; default is 0.

seed	Seed to be used for the MCMC in Bayesian Lasso; default is a random seed. If exact same results need to be reproduced, seed should be fixed to the same number.
burn.in	Burn-in period of the MCMC sampling scheme; default is 10000.
num.it	Total number of MCMC iterations including burn-in; default is 40000.
summary	Logical. If TRUE, LBL will return a summary of the analysis in the form of a list. If FALSE, LBL will return the posterior samples for all parameters. Default is set to be TRUE.
e	A (small) number ϵ in the null hypothesis of no association, $H_0 : \|\beta\| \leq \epsilon$. The default is 0.1. Changing e from the default of 0.1 may necessitate choosing a different threshold for Bayes Factor (one of the outputs) to infer association. Only used if summary = TRUE.
ci.level	Credible probability. The probability that the true value of $beta$ will be within the credible interval. Default is 0.95, which corresponds to a 95% posterior credible interval. Only used if summary = TRUE.

Value

If summary = FALSE, return a list with the following components:

haplotypes The list of haplotypes used in the analysis. The last column is the reference haplotype.

beta Posterior samples of betas stored in a matrix.

lambda A vector of (num.it-burn.in) posterior samples of lambda.

freq Posterior samples of the frequencies of haplotypes stored in a matrix format, in the same order as haplotypes.

init.freq The haplotype distribution used to initiate the MCMC.

If summary = TRUE, return the result of LBL_summary. For details, see the description of the LBL_summary function.

See Also

famLBL, cLBL, LBL_summary, print_LBL_summary, LBL-package.

Examples

```
data(cac)
cac.obj<-LBL(cac)
cac.obj
print_LBL_summary(cac.obj)
```

LBL_summary *Posterior Inference for LBL*

Description

LBL_summary provides inferences based on the posterior samples. Specifically, this function will return posterior means, credible intervals, and Bayes Factors (BF) estimates for the haplotypic effect coefficients.

Usage

```
LBL_summary(output, a, b, e = 0.1, ci.level = 0.95)
```

Arguments

output	An object returned by LBL functions (LBL, famLBL, and cLBL) with 'summary=FALSE'.
a	First hyperparameter of the prior for regression coefficients, β. The prior variance of β is $2/\lambda^2$ and λ has Gamma(a,b) prior. The Gamma prior parameters a and b are such that the mean and variance of the Gamma distribution are a/b and a/b^2. The value will be transferred from LBL functions.
b	Second hyperparameter of the Gamma(a,b) distribution described above. The value will be transferred from LBL functions.
e	A (small) number ϵ in the null hypothesis of no association, $H_0 : \|\beta\| \leq \epsilon$. The default is 0.1. Changing e from the default of 0.1 may necessitate choosing a different threshold for Bayes Factor (one of the outputs) to infer association.
ci.level	Credible probability.. The probability that the true value of $beta$ will be within the credible interval. Default is 0.95 which corresponds to a 95% posterior credible interval. Only used if 'summary = TRUE'.

Value

A list with the following components:

haplotypes The list of haplotypes used in the analysis.

OR Posterior mean of odds ratio.

OR.CI 95% posterior credible sets for the ORs.

BF Bayes Factor estimates based on posterior samples. if the posterior samples are all greater than e, then BF is set to be 999.

print_LBL_summary *Posterior Inference for LBL Functions*

Description

print_LBL_summary prints list result from LBL_summary function in a more legible style.

Usage

```
print_LBL_summary(LBL_summary_object)
```

Index