Package 'CPBayes'

January 23, 2017

Title Bayesian Meta Analysis for Studying Cross-Phenotype Genetic Associations

Version 0.1.0

Date 2017-01-22

Author Arunabha Majumdar <statgen.arunabha@gmail.com> [aut, cre], Tanushree Haldar <tanushree.haldar@gmail.com> [aut], John Witte [ctb]

Maintainer Arunabha Majumdar <statgen.arunabha@gmail.com>

Description A Bayesian meta-analysis method for studying cross-phenotype genetic associations. It uses summary-level data across multiple phenotypes to simultaneously measure the evidence of aggregate-level pleiotropic association and estimate an optimal subset of traits associated with the risk locus. CPBayes is based on a spike and slab prior and is implemented by Markov chain Monte Carlo technique Gibbs sampling.

Depends R (>= 3.2.0)

License GPL-3

LazyData TRUE

URL https://github.com/ArunabhaCodes/CPBayes

BugReports https://github.com/ArunabhaCodes/CPBayes/issues

RoxygenNote 5.0.1

Suggests testthat, knitr, rmarkdown

VignetteBuilder knitr

Imports MASS, stats

NeedsCompilation no

Repository CRAN

Date/Publication 2017-01-23 13:49:46

R topics documented:

`CPBayes` *CPBayes: An R-package implemeting a Bayesian meta analysis method for studying cross-phenotype genetic associations.*

Description

Simultaneous analysis of genetic associations with multiple phenotypes may reveal shared genetic susceptibility across traits (pleiotropy). CPBayes is a Bayesian meta analysis approach for studying cross-phenotype genetic associations. It uses summary-level data across multiple phenotypes to simultaneously measure the evidence of aggregate-level pleiotropic association and estimate an optimal subset of traits associated with the risk locus. CPBayes is based on a spike and slab prior and is implemented by Markov chain Monte Carlo (MCMC) technique Gibbs sampling.

Details

The package consists of four main functions: `cpbayes_uncor`, `cpbayes_cor`, `post_summaries`, `estimate_corln`.

Functions

`cpbayes_uncor` This function implements CPBayes for uncorrelated summary statistics. The summary statistics across traits/studies are uncorrelated when the studies have no overlapping subject.

`cpbayes_cor` This function implements CPBayes for correlated summary statistics. The summary statistics across traits/studies are correlated when the studies have overlapping subjects or the phenotypes were measured in a cohort study.

`post_summaries` This function summarizes the MCMC data produced by the main two functions `cpbayes_uncor` or `cpbayes_cor` listed above. It computes additional summaries to provide a better insight into a pleiotropic signal. It works in the same way for both `cpbayes_uncor` and `cpbayes_cor`.

`estimate_corln` This function computes an approximate correlation matrix of the beta-hat vector for multiple overlapping case-control studies or a cohort study using the sample-overlap matrices. The estimated correlation matrix can be passed as an argument into `cpbayes_cor`.

References

Arunabha Majumdar, Tanushree Haldar, Sourabh Bhattacharya, John Witte. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations (submitted). Available at: http://biorxiv.org/content/early/2017/01/18/101543.

`cpbayes_cor` *Run correlated version of CPBayes.*

Description

Run correlated version of CPBayes when the main genetic effect (beta/log(odds ratio)) estimates across studies/traits are correlated.

Usage

```
cpbayes_cor(BetaHat, SE, Corln, Phenotypes, Variant, UpdateDE = TRUE,
  MCMCiter = 20000, Burnin = 10000)
```

Arguments

`BetaHat`	A numeric vector of dimension K where K is the number of phenotypes. It contains the beta-hat values across studies/traits. No default is specified.
`SE`	A numeric vector with the same dimension as BetaHat providing the standard errors corresponding to BetaHat. Every element of SE must be positive. No default is specified.
`Corln`	A numeric square matrix of order K by K providing the correlation matrix of BetaHat. The number of rows of Corln must be the same as the dimension of the BetaHat vector. No default is specified.
`Phenotypes`	A character vector of the same dimension as BetaHat providing the name of the phenotypes. Default is specified as trait1, trait2, . . . , traitK. Note that BetaHat, SE, Corln, and Phenotypes must be in the same order.
`Variant`	A character vector of length 1 providing the name of the genetic variant. Default is 'Variant'.
`UpdateDE`	A logical vector of length 1. Default is TRUE. If FALSE, the variance of the slab distribution is considered to be fixed at 1. If TRUE, it is updated at each MCMC iteration in a range (0.8-1.2).
`MCMCiter`	A positive integer greater than or equal to 10,000. Default is 20,000.
`Burnin`	A positive integer greater than or equal to 5,000. Default is 10,000. Note that the MCMC sample size (MCMCiter - Burnin) must be at least 5,000.

Value

The output produced by cpbayes_cor is a list which consists of various components.

`variantName` It is the name of the genetic variant provided by the user. If not specified by the user, default name is 'variant'.

`log10_BF` It provides the log10(Bayes factor) produced by CPBayes that measures the evidence of the overall pleiotropic association.

`PPNA` It provides the posterior probability of null association produced by CPBayes (a Bayesian analog of the p-value) which is another measure of the evidence of aggregate-level pleiotropic association.

`subset` It provides the optimal subset of associated/non-null traits estimated by CPBayes. It is NULL if no phenotype is selected by CPBayes.

`important_traits` It provides the traits which yield a trait-specific posterior probability of association (PPAj) > 25%. Even if a phenotype is not selected in the optimal subset of non-null traits, it can produce a non-negligible value of PPAj. We note that 'important_traits' is expected to include the traits already contained in 'subset'. It provides both the name of the important traits and their corresponding value of PPAj. Always check out 'important_traits' even if 'subset' contains no trait or a single trait. It helps to better explain an observed pleiotropic signal.

`auxi_data` It contains supplementary data generated by the MCMC which is used later by the post_summaries function to provide additional insights into a pleiotropic signal. The supplementary data contained in auxi_data are as follows.

1. traitNames: Name of all the phenotypes.
2. K: Total number of phenotypes.
3. mcmc.samplesize: MCMC sample size.
4. asso.pr: Trait-specific posterior probability for all the traits.
5. Z.data: MCMC data on the latent association status (Z).
6. sim.beta: MCMC data on the unknown true genetic effect (beta) on each trait.

`uncor_use` 'Yes' or 'No'. Whether the combined strategy of CPBayes (implemented for correlated summary statistics) used the uncorrelated version or not.

`runtime` It provides the runtime (in seconds) taken by CPBayes. It will help the user to plan the whole analysis.

References

Arunabha Majumdar, Tanushree Haldar, Sourabh Bhattacharya, John Witte. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations (submitted). Available at: http://biorxiv.org/content/early/2017/01/18/101543.

See Also

estimate_corln, post_summaries, cpbayes_uncor

Examples

```
data(ExampleDataCor)
BetaHat <- ExampleDataCor$BetaHat
BetaHat
SE <- ExampleDataCor$SE
SE
cor <- ExampleDataCor$cor
cor
traitNames <- paste("Disease", 1:10, sep = "")
SNP1 <- "rs1234"
result <- cpbayes_cor(BetaHat, SE, cor, Phenotypes = traitNames, Variant = SNP1)
str(result)
```

---

cpbayes_uncor *Run uncorrelated version of CPBayes.*

Description

Run uncorrelated version of CPBayes when the main genetic effect (beta/log(odds ratio)) estimates across studies/traits are uncorrelated.

Usage

```
cpbayes_uncor(BetaHat, SE, Phenotypes, Variant, UpdateDE = TRUE,
  MCMCiter = 20000, Burnin = 10000)
```

Arguments

BetaHat	A numeric vector of dimension K where K is the number of phenotypes. It contains the beta hat values across studies/traits. No default is specified.
SE	A numeric vector with the same dimension as BetaHat providing the standard errors corresponding to BetaHat. Every element of SE must be positive. No default is specified.
Phenotypes	A character vector of the same dimension as BetaHat providing the name of the phenotypes. Default is specified as trait1, trait2, . . . , traitK. Note that BetaHat, SE, and Phenotypes must be in the same order.
Variant	A character vector of length 1 specifying the name of the genetic variant. Default is 'Variant'.
UpdateDE	A logical vector of length 1. Default is TRUE. If FALSE, the variance of the slab distribution is considered to be fixed at 1. If TRUE, it is updated at each MCMC iteration in a range (0.8-1.2).
MCMCiter	A positive integer greater than or equal to 10,000. Default is 20,000.
Burnin	A positive integer greater than or equal to 5,000. Default is 10,000. Note that the MCMC sample size (MCMCiter - Burnin) must be at least 5,000.

Value

The output produced by the function is a list which consists of various components.

`variantName` It is the name of the genetic variant provided by the user. If not specified by the user, default name is 'variant'.

`log10_BF` It provides the log10(Bayes factor) produced by CPBayes that measures the evidence of the overall pleiotropic association.

`PPNA` It provides the posterior probability of null association produced by CPBayes (a Bayesian analog of the p-value) which is another measure of the evidence of the aggregate-level pleiotropic association.

`subset` It provides the optimal subset of associated/non-null traits estimated by CPBayes. It is NULL if no phenotype is selected by CPBayes.

`important_traits` It provides the traits which yield a trait-specific posterior probability of association (PPAj) > 25%. Even if a phenotype is not selected in the optimal subset of non-null traits, it can produce a non-negligible value of trait-specific posterior probability of association (PPAj). We note that 'important_traits' is expected to include the traits already contained in 'subset'. It provides both the name of the important traits and their corresponding values of PPAj. Always check out 'important_traits' even if 'subset' contains no trait or a single trait. It helps to better explain an observed pleiotropic signal.

`auxi_data` It contains supplementary data generated by the MCMC which is used later by the `post_summaries` function to provide additional insights into a pleiotropic signal. The supplementary data contained in auxi_data are as follows.

1. traitNames: Name of all the phenotypes.
2. K: Total number of phenotypes.
3. mcmc.samplesize: MCMC sample size.
4. PPAj: Trait-specific posterior probability of association for all the traits.
5. Z.data: MCMC data on the latent association status of all the traits (Z).
6. sim.beta: MCMC data on the unknown true genetic effect (beta) on all the traits.

`runtime` It provides the runtime (in seconds) taken by CPBayes. It will help the user to plan the whole analysis.

References

Arunabha Majumdar, Tanushree Haldar, Sourabh Bhattacharya, John Witte. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations (submitted). Available at: http://biorxiv.org/content/early/2017/01/18/101543.

See Also

`post_summaries`, `cpbayes_cor`, `estimate_corln`

Examples

```
data(ExampleDataUncor)
BetaHat <- ExampleDataUncor$BetaHat
BetaHat
SE <- ExampleDataUncor$SE
SE
traitNames <- paste("Disease", 1:10, sep = "")
SNP1 <- "rs1234"
result <- cpbayes_uncor(BetaHat, SE, Phenotypes = traitNames, Variant = SNP1)
str(result)
```

`estimate_corln` *Estimate correlation structure of beta-hat vector for multiple overlapping case-control studies or a cohort study using sample-overlap matrix.*

Description

Compute an approximate correlation matrix of the beta-hat vector for multiple overlapping case-control studies or a cohort study using the sample-overlap matrices which describe the number of cases or controls shared between studies/traits, and the number of subjects who are case for one study/trait but control for another study/trait. This approximation is more accurate when none of the diseases/traits is associated with the environmental covariates present in the study.

Important note on the estimation of correlation structure of correlated beta-hat vector: In general, environmental covariates are expected to be present in a study and associated with the phenotypes of interest. Hence the above approximation of the correlation matrix may not be completely accurate. So, in presence of environmental covariates, we recommend an alternative strategy to estimate the correlation matrix using the genome-wide summary statistics data across traits as follows. First, extract all the SNPs for each of which the trait-specific univariate association p-value across all the traits are > 0.1. The trait-specific univariate association p-values can be obtained based on the beta-hat and standard error for each trait. Each of the SNPs selected in this way is either weakly or not associated with any of the phenotypes (null SNP). Next, select a set of independent null SNPs from the initial set of null SNPs by using a threshold of r^2 < 0.01 (r: the correlation between the genotypes at a pair of SNPs). Finally, compute the correlation matrix of the effect estimates (beta-hat vector) as the sample correlation matrix of the beta-hat vector across all the selected independent null SNPs. This strategy is more general and applicable to a cohort study or multiple overlapping studies for binary or quantitative traits with arbitrary distributions. Misspecification of the correlation structure can affect the results produced by CPBayes to some extent. Hence, if genome-wide summary statistics data across traits is available, we recommend to use this alternative strategy to estimate the correlation matrix of the beta-hat vector. See our paper for more details at: http://biorxiv.org/content/early/2017/01/18/101543.

Usage

```
estimate_corln(n11, n00, n10)
```

Arguments

n11	An integer square matrix (number of rows must be the same as the number of studies/traits) providing the number of cases shared between all possible pairs of studies/traits. So (k,l)-th element of n11 is the number of subjects who are case for both k-th and l-th study/trait. Note that the diagonal elements of n11 are the number of cases across studies/traits. In case, no case is shared between studies/traits, the off-diagonal elements of n11 will be zero. No default is specified.
n00	An integer square matrix (number of rows must be the same as the number of studies/traits) providing the number of controls shared between all possible pairs of studies/traits. So (k,l)-th element of n00 is the number subjects who are control for both k-th and l-th study/trait. Note that the diagonal elements of n00 are the number of controls across studies/traits. In case, no control is shared between studies/traits, the off-diagonal elements will be zero. No default is specified.
n10	An integer square matrix (number of rows must be the same as the number of studies/traits) providing the number of subjects who are case for one study/trait and control for another study/trait. Clearly, the diagonal elements will be zero. An off diagonal element, e.g., (k,l)-th element of n10 is the number of subjects who are case for k-th study/trait and control for l-th study/trait. If there is no such overlap, all the elements of n10 will be zero. No default is specified.

Value

This function returns an approximate correlation matrix of the beta-hat vector for multiple overlapping case-control studies or a cohort study. See the example below.

References

Arunabha Majumdar, Tanushree Haldar, Sourabh Bhattacharya, John Witte. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations (submitted). Available at: http://biorxiv.org/content/early/2017/01/18/101543.

See Also

cpbayes_cor

Examples

```
data(SampleOverlapMatrix)
n11 <- SampleOverlapMatrix$n11
n11
n00 <- SampleOverlapMatrix$n00
n00
n10 <- SampleOverlapMatrix$n10
n10
cor <- estimate_corln(n11, n00, n10)
cor
```

ExampleDataCor *An example data for correlated summary statistics.*

Description

ExampleDataCor is a list consisting of three components: BetaHat, SE, cor. ExampleDataCor$BetaHat is a numeric vector that contains the main genetic effect (beta/log(odds ratio)) estimates for a SNP across 10 overlapping case-control studies for 10 different diseases. Each of the 10 studies has a distinct set of 7000 cases and a common set of 10000 controls shared across all the studies. In each case-control study, we fit a logistic regression of the case-control status on the genotype coded as the minor allele count for all the individuals in the sample. One can also include various covariates, such as, age, gender, principal components (PCs) of ancestries in the logistic regression. From each logistic regression for a disease, we obtain the estimate of the main genetic association parameter (beta/log(odds ratio)) along with the corresponding standard error. Since the studies have overlapping subjects, the beta-hat across traits are correlated. ExampleDataCor$SE contains the standard error vector corresponding to the correlated beta-hat vector. ExampleDataCor$cor is a numeric square matrix providing the correlation matrix of the correlated beta-hat vector.

Usage

```
data(ExampleDataCor)
```

Format

A list consisting of two numeric vectors (each of length 10) and a numeric square matrix of dimension 10 by 10:

BetaHat beta hat vector of length 10.

SE standard error vector corresponding to the beta-hat vector.

cor correlation matrix of the beta-hat vector.

Examples

```
data(ExampleDataCor)
BetaHat <- ExampleDataCor$BetaHat
BetaHat
SE <- ExampleDataCor$SE
SE
cor <- ExampleDataCor$cor
cor
cpbayes_cor(BetaHat, SE, cor)
```

ExampleDataUncor *An example data for uncorrelated summary statistics.*

Description

ExampleDataUncor is a list which has two components: BetaHat, SE. The numeric vector ExampleDataUncor$BetaHat contains the main genetic effect (beta/log(odds ratio)) estimates for a single nucleotide polymorphism (SNP) obtained from 10 separate case-control studies for 10 different diseases. In each case-control study comprising a distinct set of 7000 cases and 10000 controls, we fit a logistic regression of the case-control status on the genotype coded as the minor allele count for all the individuals in the sample. One can also include various covariates, such as, age, gender, principal components (PCs) of ancestries in the logistic regression. From each logistic regression for a disease, we obtain the estimate of the main genetic association parameter (beta/log(odds ratio)) along with the corresponding standard error. Since the studies do not have any overlapping subject, the beta-hat across the traits are uncorrelated. ExampleDataUncor$SE is the second numeric vector that contains the standard errors corresponding to the uncorrelated beta-hat vector.

Usage

```
data(ExampleDataUncor)
```

Format

A list of two numeric vectors each of length 10 (for 10 studies):

BetaHat beta hat vector of length 10.

SE standard error vector corresponding to beta-hat vector.

Examples

```
data(ExampleDataUncor)
BetaHat <- ExampleDataUncor$BetaHat
BetaHat
SE <- ExampleDataUncor$SE
SE
cpbayes_uncor(BetaHat, SE)
```

post_summaries *Post summary of the MCMC data generated by the uncorrelated or correlated version of CPBayes.*

Description

Run the post_summaries function to summarize the MCMC data produced by cpbayes_uncor or cpbayes_cor and obtain meaningful insights into an observed pleiotropic signal.

Usage

```
post_summaries(mcmc_output, level = 0.05)
```

Arguments

mcmc_output	A list returned by either of the two main CPBayes functions cpbayes_uncor and cpbayes_cor. This list contains all the primary results and MCMC data produced by cpbayes_uncor or cpbayes_cor. No default is specified. See the example below.
level	A numeric value. (1-level)% credible interval (Bayesian analog of the confidence interval) is computed for the true unknown genetic effect (beta/odds ratio) on the traits. Default choice is 0.05.

Value

The output produced by this function is a list that consists of various components.

variantName	It is the name of the genetic variant provided by the user. If not specified by the user, default name is 'variant'.
log10_BF	It provides the log10(Bayes factor) produced by CPBayes that measures the evidence of the overall pleiotropic association.
PPNA	It provides the posterior probability of null association produced by CPBayes (a Bayesian analog of the p-value) which is another measure of the evidence of aggregate-level pleiotropic association.
subset	A data frame providing the optimal subset of associated/non-null traits along with their trait-specific posterior probability of association (PPAj) and direction of associations. It is NULL if no phenotype is selected by CPBayes.
important_traits	It provides the traits which yield a trait-specific posterior probability of association (PPAj) > 25%. Even if a phenotype is not selected in the optimal subset of non-null traits, it can produce a non-negligible value of trait-specific posterior probability of association. We note that 'important_traits' is expected to include the traits already contained in 'subset'. It provides the name of the important traits and their trait-specific posterior probability of association (PPAj) and the direction of associations. Always check out 'important_traits' even if 'subset' contains no trait or a single trait. It helps to better explain an observed pleiotropic signal.
traitNames	It returns the name of all the phenotypes specified by the user. Default is trait1, trait2, ... , traitK.
PPAj	Data frame provaiding the trait-specific posterior probability of association for all the phenotypes.
poste_summary_beta	Data frame providing the posterior summary of the unknown true genetic effect (beta) on each trait. It gives posterior mean, median, standard error, credible interval (lower and upper limits) of the unknown true beta corresponding to each trait.

poste_summary_OR
: Data frame providing the posterior summary of the unknown true genetic effect (odds ratio) on each trait. It gives posterior mean, median, standard error, credible interval (lower and upper limits) of the unknown true odds ratio corresponding to each trait.

References

Arunabha Majumdar, Tanushree Haldar, Sourabh Bhattacharya, John Witte. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations (submitted). Available at: http://biorxiv.org/content/early/2017/01/18/101543.

See Also

cpbayes_uncor, cpbayes_cor, estimate_corln

Examples

```
data(ExampleDataUncor)
BetaHat <- ExampleDataUncor$BetaHat
BetaHat
SE <- ExampleDataUncor$SE
SE
traitNames <- paste("Disease", 1:10, sep = "")
SNP1 <- "rs1234"
result <- cpbayes_uncor(BetaHat, SE, Phenotypes = traitNames, Variant = SNP1)
PleioSumm <- post_summaries(result, level = 0.05)
str(PleioSumm)
```

SampleOverlapMatrix *An example data of sample-overlap matrices.*

Description

An example data of sample-overlap matrices for five different diseases in the Kaiser GERA cohort (a real data). SampleOverlapMatrix is a list that contains an example of the sample overlap matrices for five different diseases in the Kaiser GERA cohort. SampleOverlapMatrix$n11 provides the number of cases shared between all possible pairs of diseases. SampleOverlapMatrix$n00 provides the number of controls shared between all possible pairs of diseases. SampleOverlapMatrix$n10 provides the number of subjects who are case for one disease and control for another disease.

Usage

```
data(SampleOverlapMatrix)
```

Format

A list consisting of three integer square matrices (each of dimension 5 by 5):

n11 number of cases shared between all possible pairs of diseases.

n00 number of controls shared between all possible pairs of diseases.

n10 number of subjects who are case for one disease and control for another disease.

Examples

```
data(SampleOverlapMatrix)
n11 <- SampleOverlapMatrix$n11
n11
n00 <- SampleOverlapMatrix$n00
n00
n10 <- SampleOverlapMatrix$n10
n10
estimate_corln(n11,n00,n10)
```

Index