Package 'vbdm'

February 20, 2015

Type Package

Title Variational Bayes Discrete Mixture Model

Version 0.0.4

Date 2014-02-01

Author Benjamin Logsdon

Maintainer Benjamin Logsdon <blogsdon@uw.edu>

Depends R (>= 3.0.0)

Description Efficient algorithm for solving discrete mixture regression model for rare variant association analysis. Uses variational Bayes algorithm to efficiently search over model space. Outputs an approximate likelihood ratio test as well as variant level posterior probabilities of association.

License GPL-2

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-02-01 11:18:59

R topics documented:

`burdenPlot` *plotting function for rare variant vbdm test*

Description

This function produces a plot depicting the phenotype distribution, the distribution of carriers of the rare alleles, and also can include the posterior probaiblity of association as estimated by the vbdm algorithm.

Usage

```
burdenPlot(y, G, annotation = rep('missense',ncol(G)), title='',
          order='mean', legend='keep', type='lines',
          post=NULL, name.snp=NULL)
```

Arguments

`y`	A vector of continuous phenotypes.
`G`	A matrix of genotypes that are encoded as 0, 1, or 2.
`annotation`	A vector of SNP annotations for the columns of G. The default is `'missense'`. Possible annotations are `'missense'`, `'nonsense'`,`'splice'`, and `'synonymous'`. Default is for all variants to be missense.
`title`	An optional title for the plot.
`order`	How the variants should be ordered in the bottom quadrant of the plot. Possible options include `order='mean'` which orders based on the mean phenotypic value of carriers of the rare variants, `order='MAF'` which orders the variants based on minor allele frequency, `order='MAF.mean'` which orders variants first by MAF, then by mean phenotypic value, `order='anno'` which orders by annotation first, then by mean phenotypic value, and `order=''` which removes any reordering.
`legend`	If `legend='keep'` then a legend is depicted with variant annotations.
`type`	If `type='lines'` then the range of phenotype values for carriers are shown with a horizontal lines, with phenotypic values at the vertical tick marks. If `type='points'` then the phenotypic values are shown with points.
`post`	Optional vector of posterior probabilities from `vbdm` result.
`name.snp`	Optional vector of snp names.

Author(s)

Paul L. Auer (paul.wl.auer@gmail.com), Benjamin A. Logsdon (blogsdon@uw.edu)

References

Logsdon, B.A., et al. (2014) *A Variational Bayes Discrete Mixture Test for Rare Variant Association.*, *Genetic Epidemiology, Vol. 38(1), 21-30 2014*

See Also

vbdm, vbdmR

Examples

```
#generate some test data
library(vbdm)
set.seed(1)
n <- 1000
m <- 30
G <- matrix(rbinom(n*m,2,.01),n,m);
beta1 <- rbinom(m,1,.2)
y <- G%*%beta1+rnorm(n,0,2)
res <- vbdm(y=y,G=G,scaling=FALSE);
bp<-burdenPlot(y=y,G=G,post=res$pvec,name.snp=1:30);
```

---

vbdm *fit a discrete mixture model*

Description

Fits a discrete mixture model for rare variant association analysis. Uses an approximate variational Bayes coordinate ascent algorithm for a computationally efficient solution.

Usage

```
vbdm(y, G, X=NULL, thres=0.05, genotypes=TRUE,
     include.mean=TRUE, minor.allele=TRUE, impute="MEAN",
     eps=1e-4, scaling=TRUE, nperm=0, maxit=1000, hyper=c(2,2))
```

Arguments

y	A vector of continuous phenotypes.
G	A matrix of genotypes or variables of interest. Rows are treated as individuals and columns are treated as genotypes.
X	An optional matrix of covariates.
thres	If the matrix is of genotypes, then this specifies a minor allele frequency threshold. Variants with a MAF greater than this threshold are excluded from the analysis.
genotypes	This specifies whether or not to treat G as a matrix of genotypes. If it is treated as genotypes then it will be filtered based on thres, and there are more options for missing data imputations. The default genotype encoding is additive (e.g. genotypes are encoded as 0,1,2). Also if G is a genotype matrix vbdm will flip

	the encoding such that the homozygous major allele genotype is encoded as 0, the heterozygote as 1, and the homozygous minor allele genotype as 2 unless `minor.allele=FALSE`
`include.mean`	This specifies whether to add an interecept term to the model. If no covariates are provided it is automatically added, but if there are covariates provided it can be optional.
`minor.allele`	When `minor.allele=TRUE` and `genotypes=TRUE` the genotypes are flipped so that the major allele genotype is encoded as 0.
`impute`	If there is missing data in `G` this specifies the method with which to impute the missing data. There are two options `impute="MEAN"` which sets any missing genotype to the expected dosage given the MAF, or `impute="MAJOR"` which sets any missing genoypte to the homozygous genotype of the major allele. If the matrix is not treated as a genotype matrix (e.g. `genotype=FALSE`), then only `impute="MEAN"` will work. Also, missing data is not allowed in the covariates X.
`eps`	The tolerance for convergence of the coordinate ascent algorithm based on the change in the lower bound of the log marginal likeilhood.
`scaling`	Whether or not to scale the genotypes to have mean 0 and variance 1.
`nperm`	Optional parameter defining the number of null permutations of the vbdm likelihood ratio test. This can be used to generate an exact p-value
`maxit`	The maximum number of iterations allowed for the vbdm algorithm.
`hyper`	The hyperparameters for the prior defined over the mixing probability parameter. The first hyperparameter is the alpha parameter, and the second is the beta parameter.

Value

`y`	The phenotype vector passed to vbdm.
`G`	The genotype matrix passed to vbdm. Note that any variables that were dropped will be dropped from this matrix.
`X`	The covariate matrix passed to vbdm. Will include intercept term if it was added earlier.
`keep`	A vector of indices of the kept variables in G (if any were excluded based on `thres`)
`pvec`	The vector of estimated posterior probabilities for each variable in G.
`gamma`	A vector of additive covariate effect estimates.
`theta`	The estimated effect of the variables in G.
`sigma`	The estimated error variance.
`prob`	The estimated mixing parameter.
`lb`	The lower bound of the marginal log likelihood.
`lbnull`	The lower bound of the marginal log likelihood under the null model.
`lrt`	The approximate likelihood ratio test based on the lower bounds.
`p.value`	A p-value computed based on `lrt` with the assumption that `lrt~chi^2_1`

`lbperm`	If `nperm>0`, the lower bound of the fitted null permutations.
`lrtperm`	If `nperm>0`, the likelihood ratio test of the fitted null permutations.
`p.value.perm`	If `nperm>0`, the empirical p-value based on the fitted null permutations.

Author(s)

Benjamin A. Logsdon (blogsdon@uw.edu)

References

Logsdon, B.A., et al. (2014) *A Variational Bayes Discrete Mixture Test for Rare Variant Association.*, *Genetic Epidemiology, Vol. 38(1), 21-30 2014*

See Also

`vbdmR`,`burdenPlot`

Examples

```
#generate some test data
library(vbdm)
set.seed(3)
n <- 1000
m <- 20
G <- matrix(rbinom(n*m,2,.01),n,m);
beta1 <- rbinom(m,1,.2)
y <- G%*%beta1+rnorm(n,0,1.3)

#with scaling:
res <- vbdm(y=y,G=G);
T5 <- summary(lm(y~rowSums(scale(G))))$coef[2,4];
cat('vbdm p-value:',res$p.value,'\nT5 p-value:',T5,'\n')
#vbdm p-value: 0.001345869
#T5 p-value: 0.9481797

#without scaling:
res <- vbdm(y=y,G=G,scaling=FALSE)
T5 <- summary(lm(y~rowSums(G)))$coef[2,4];
cat('vbdm p-value:',res$p.value,'\nT5 p-value:',T5,'\n')
#vbdm p-value: 0.0005315836
#T5 p-value: 0.904476

#run 100 permutations
set.seed(2)
res <- vbdm(y=y,G=G,scaling=FALSE,nperm=1e2);
cat('vbdm approximate p-value:',res$p.value,'\nvbdm permutation p-value <',res$p.value.perm,'\n');
#vbdm approximate p-value: 0.0005315836
#vbdm permutation p-value: 0
```

vbdmR *fit a discrete mixture model (R implementation)*

Description

Fits a discrete mixture model for rare variant association analysis. Uses an approximate variational Bayes coordinate ascent algorithm for a computationally efficient solution. This is the slow but well documented R implementation.

Usage

```
vbdmR(y, G, X=NULL, tol=1e-4, thres=0.05, scaling=TRUE,
      hyper=c(2,2))
```

Arguments

y	A vector of continuous phenotypes.
G	A matrix of genotypes or variables of interest.
X	An optional matrix of covariates.
tol	The tolerance for convergence based on the change in the lower bound on the marginal log likelihood in the vbdm algorithm.
thres	If the matrix is of genotypes, then this specifies a minor allele frequency threshold. Variants with a MAF greater than this threshold are excluded from the analysis.
scaling	Whether or not to scale the genotypes to have mean 0 and variance 1.
hyper	The hyperparameters for the prior defined over the mixing probability parameter. The first hyperparameter is the alpha parameter, and the second is the beta parameter.

Details

This function contains the much slower, but well documented R implementation of the vbdm algorithm. This function does not have all of the sanity checks that vbdm has, and should therefore only be used for diagnostic purposes.

Value

y	The phenotype vector passed to vbdmR.
G	The genotype matrix passed to vbdmR. Note that any variables that were dropped will be dropped from this matrix.
X	The covariate matrix passed to vbdmR. Will include intercept term if it was added earlier.
keep	A vector of indices of the kept variables in G (if any were excluded based on thres)

pvec	The vector of estimated posterior probabilities for each variable in G.
gamma	A vector of additive covariate effect estimates.
theta	The estimated effect of the variables in G.
sigma	The estimated error variance.
prob	The estimated mixing parameter.
lb	The lower bound of the marginal log likelihood.
lbnull	The lower bound of the marginal log likelihood under the null model.
lrt	The approximate likelihood ratio test based on the lower bounds.
p.value	A p-value computed based on lrt with the assumption that lrt~chi^2_1

Author(s)

Benjamin A. Logsdon (blogsdon@uw.edu)

References

Logsdon, B.A., et al. (2014) *A Variational Bayes Discrete Mixture Test for Rare Variant Association.*, *Genetic Epidemiology, Vol. 38(1), 21-30 2014*

See Also

vbdm, burdenPlot

Examples

```
#generate some test data
library(vbdm)
set.seed(3)
n <- 1000
m <- 20
G <- matrix(rbinom(n*m,2,.01),n,m);
beta1 <- rbinom(m,1,.2)
y <- G%*%beta1+rnorm(n,0,1.3)

#compare implementations
res1 <- vbdm(y=y,G=G);
res2 <- vbdmR(y=y,G=G);
T5 <- summary(lm(y~rowSums(scale(G))))$coef[2,4];
cat('vbdm p-value:',res1$p.value,
  '\nvbdmR p-value:',res2$p.value,
  '\nT5 p-value:',T5,'\n')
#vbdm p-value: 0.001345869
#vbdmR p-value: 0.001345869
#T5 p-value: 0.9481797
```

Index