Package 'genMOSS'

February 19, 2015

Title Functions for the Bayesian Analysis of GWAS Data

Version 1.2

Date 2014-12-01

Author Matthew Friedlander, Adrian Dobra, Helene Massam, and Laurent Briollais

Maintainer Matthew Friedlander <friedlander.matthew@gmail.com>

Depends R (>= 2.10), ROCR

Description

Implements the Mode Oriented Stochastic Search (MOSS) algorithm as well as a simple moving window approach to look for combinations of SNPs that are associated with a response.

License GPL (>= 2)

LazyLoad yes

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-12-03 07:54:14

R topics documented:

genMOSS *Functions for the Bayesian analysis of GWAS data*

Description

The genMOSS package implements the Mode Oriented Stochastic Search (MOSS) of Dobra and Massam (2010) as well as a simple moving window approach (see Sun et. al (2006) and Wu et. al (2010)) to identify combinations of SNPs that are associated with a response.

Details

Package:	genMOSS
Type:	Package
Version:	1.2
Date:	2014-12-01
License:	GPL-2
LazyLoad:	yes

The function MOSS_GWAS implements the MOSS algorithm while the mWindow implements the moving window approach.

Author(s)

Authors: Matthew Friedlander, Adrian Dobra, Helene Massam, and Laurent Briollais, Maintainer: Matthew Friedlander <friedlander.matthew@gmail.com>

References

[1] Massam, H., Liu, J. and Dobra, A. (2009). A conjugate prior for discrete hierarchical log-linear models. Annals of Statistics, 37, 3431-3467.

[2] Dobra, A., Briollais, L., Jarjanazi, H., Ozcelik, H. and Massam, H. (2010). Applications of the mode oriented stochastic search (MOSS) algorithm for discrete multi-way data to genomewide studies. Bayesian Modeling in Bioinformatics, Taylor & Francis (Dey, D., Ghosh, S., and Mallick, B., eds.), 63-93.

[3] Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. Statistical Methodology, 7, 240-253.

[4] Sun, J., Levin, A., Boerwinkle, E., Robertson, H., and Kardia, S. (2006). A scan statistic for identifying chromosomal patterns of SNP association. Genetic Epidemiology, 30, 627-635.

[5] Wu, M., Kraft, P., Epstein, M., Taylor, D., Chanock, S., Hunter, D., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. The American Journal of Human Genetics, 86, 929-942.

Examples

```
data(simuCC)
data <- simuCC[,c(1002,2971,rep(5978:6001))]
# The SNPs in columns 1002 and 2971 of simuCC called rs4491689 and rs6869003 cause the disease.
set.seed(7)
MOSS_GWAS (alpha = 1, c = 0.1, cPrime = 0.0001, q = 0.1, replicates = 1,
           maxVars = 3, data, dimens = c(rep(3,25),2), confVars = NULL, k = NULL)
s <- mWindow (data, dimens = c(rep(3,25),2), alpha = 1, windowSize = 2)
head (s, n = 5)
```

MOSS_GWAS *Analyzing GWAS data using MOSS*

Description

The Mode Oriented Stochastic Search (MOSS) of Dobra at. al (2010) is a two stage Bayesian variable selection procedure for analyzing GWAS data. If we let Y be a response, X be a small group of SNPs, and r = Y|X be called the regression of Y on X, the first stage of the procedure is to find regressions such that the marginal likelihood, P(r) = P(Y|X), is high. In the second stage, a hierarchical log-linear model search is performed to identify the most relevant interactions among the variables in each of the top regressions. Using the top log-linear models, model averaging is then used to construct a classifier for predicting the response and its capability is assessed via k-fold cross validation. The prior used in Bayesian computations is the generalized hyper Dirichlet of Massam et. al (2009).

Usage

```
MOSS_GWAS(alpha = 1, c = 0.1, cPrime = 0.0001, q = 0.1,
          replicates = 5, maxVars = 3, data, dimens, confVars = NULL, k = NULL)
```

Arguments

alpha	A hyperparameter of the prior representing the total of a fictive contingency table with counts equal to alpha divided by the number of cells. Alpha must be a positive real number.
c, cPrime, q	Tuning parameters for MOSS. All 3 must be real numbers between 0 and 1 and cPrime must be smaller than c.
replicates	The number of instances the first stage of the MOSS procedure will be run. The top regressions are culled from the results of all the replicates.
maxVars	The maximum number of variables allowed in a regression (including the response). Must be an integer from 3 to 6.
data	A data frame containing the genotype information for a given set of SNPs. The data frame should be organized such that each row refers to a subject and each column to a SNP. The last column must be a binary response for each subject. The data frame must contain at least 8 columns. Rows containing any missing values (i.e. NAs) are omitted from the analysis.

dimens	The number of possible values for each column of data. Each possible value does not need to occur in data. All entries of dimens must be greater than or equal to 2.
confVars	The parameter confVars (for confounding variables) is a character vector specifying the names of SNPs which, other than the response, will be forced to be in every regression. If no confounding variables are desired, confVars can be set to NULL. A maximum of (maxVars - 2) confounding variables may be specified.
k	The fold of the cross validation. If k is NULL then only the first stage of MOSS is carried out.

Value

A list with 4 data frame elements:

topRegressions	The top regressions identified together with their log marginal likelihood
postIncProbs	The posterior inclusion probabilities of each SNP that appears in one of the top regressions. This is obtained by adding the marginal likelihoods of the regressions in which each SNP appears and then normalizing over all the regressions.
interactionModels	The best (in terms of marginal likelihood) hierarchical log-linear model containing the variables in each of the top regressions.
fits	The fitted interaction models (using the glm function).
cvMatrix	A matrix with the overall results of the k-fold cross validation. This table is typically called a confusion matrix.
cvDiag	Some diagnostic information based on the cross validation: 'acc' is the accuracy, 'tpr' is the true positive rate, 'fpr' is the false postive rate, and 'auc' is the area under the ROC curve.

Note

The function recode_data can decide whether diallelic SNPs (i.e. SNPs with three categories) should be recoded as binary and in which way. If desired this function should be used prior to MOSS_GWAS.

Author(s)

Matthew Friedlander, Adrian Dobra, Helene Massam, and Laurent Briollais

References

[1] Massam, H., Liu, J. and Dobra, A. (2009). A conjugate prior for discrete hierarchical log-linear models. Annals of Statistics, 37, 3431-3467.

[2] Dobra, A., Briollais, L., Jarjanazi, H., Ozcelik, H. and Massam, H. (2010). Applications of the mode oriented stochastic search (MOSS) algorithm for discrete multi-way data to genomewide studies. Bayesian Modeling in Bioinformatics, Taylor & Francis (Dey, D., Ghosh, S., and Mallick, B., eds.), 63-93.

[3] Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. Statistical Methodology, 7, 240-253.

See Also

recode_data

Examples

```
data(simuCC)
data <- simuCC[,c(1002,2971,rep(5978:6001))]
# The SNPs in columns 1002 and 2971 of simuCC called rs4491689 and rs6869003 cause the disease.
set.seed(7)
MOSS_GWAS (alpha = 1, c = 0.1, cPrime = 0.0001, q = 0.1, replicates = 1,
           maxVars = 3, data, dimens = c(rep(3,25),2), confVars = NULL, k = NULL)
```

---

mWindow *Analyzing GWAS data in sequence over a moving window*

Description

A simple alternative to stochastically searching through all combinations of SNPs (as MOSS_GWAS does) is to group SNPs together according to the sequence that they appear in a genetic region (if that information is available). Definining a window size w we first group SNPs 1 to w together and then SNPs 2 to w + 1 together and so on. If we let Y be a response, X be a group of SNPs, and r = Y|X be called the regression of Y on X, the mWindow function computes the marginal likelihood, P(r) = P(Y|X), of the regression corresponding to each group (or window). The aim is to identify those regressions such that the marginal likelihood is the highest. The groups of SNPs (or genetic regions) contained in these regressions are most associated with the response. The prior used in Bayesian computations is the generalized hyper Dirichlet of Massam et. al (2009).

Usage

```
mWindow (data, dimens, alpha = 1, windowSize = 2)
```

Arguments

data	A data frame containing the genotype information for a given set of SNPs. The data frame should be organized such that each row refers to a subject and each column to a SNP. The last column must be a binary response for each subject. The data frame must contain at least 8 columns. Rows containing any missing values (i.e. NAs) are omitted from the analysis.
dimens	The number of possible values for each column of data. Each possible value does not need to occur in data. All entries of dimens must be greater than or equal to 2.
alpha	A hyperparameter of the prior representing the total of a fictive contingency table with counts equal to alpha divided by the number of cells. Alpha must be a positive real number.
windowSize	The size of the moving window. Must be an integer from 1 to 5.

Value

A data frame listing the regression formed in each window and its corresponding log marginal likelihood. The data frame is sorted in descending order by log marginal likelihood.

Note

The function recode_data can decide whether diallelic SNPs (i.e. SNPs with three categories) should be recoded as binary and in which way. If desired this function should be used prior to mWindow.

Author(s)

Matthew Friedlander, Adrian Dobra, Helene Massam, and Laurent Briollais

References

[1] Massam, H., Liu, J. and Dobra, A. (2009). A conjugate prior for discrete hierarchical log-linear models. Annals of Statistics, 37, 3431-3467.

[2] Dobra, A., Briollais, L., Jarjanazi, H., Ozcelik, H. and Massam, H. (2010). Applications of the mode oriented stochastic search (MOSS) algorithm for discrete multi-way data to genomewide studies. Bayesian Modeling in Bioinformatics, Taylor & Francis (Dey, D., Ghosh, S., and Mallick, B., eds.), 63-93.

[3] Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. Statistical Methodology, 7, 240-253.

[4] Sun, J., Levin, A., Boerwinkle, E., Robertson, H., and Kardia, S. (2006). A scan statistic for identifying chromosomal patterns of SNP association. Genetic Epidemiology, 30, 627-635.

[5] Wu, M., Kraft, P., Epstein, M., Taylor, D., Chanock, S., Hunter, D., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. The American Journal of Human Genetics, 86, 929-942.

See Also

recode_data

Examples

```
data(simuCC)
data <- simuCC[,c(1002,2971,rep(5978:6001))]
# The SNPs in columns 1002 and 2971 of simuCC called rs4491689 and rs6869003 cause the disease.
set.seed(7)
s <- mWindow (data, dimens = c(rep(3,25),2), alpha = 1, windowSize = 2)
head (s, n = 5)
```

recode_data *Finding the optimal coding of diallelic SNPs*

Description

Let Y be a response, X be a SNP, and r = Y|X be called the regression of Y on X. For a diallelic SNP (i.e. a SNP with 3 categories), it may be that the marginal likelihood of the regression, P(r) = P(Y|X), is higher when the SNP is recoded as binary. Using the coding that maximizes this marginal likelihood may increase the power. Trinary variables can be recoded as binary in three different ways (or can be left as is). The function recode_data finds the optimal coding for each diallelic SNP in a given data frame and returns a revised data frame in the same order as the original. SNPs that are not diallelic are inserted into the new data frame unchanged. A vector containing the dimension of each SNP in the revised data frame is also returned. The prior used in Bayesian computations is the generalized hyper Dirichelt of Massam et. al (2009).

Usage

```
recode_data (data, dimens, alpha = 1)
```

Arguments

data	A data frame containing the genotype information for a given set of SNPs. The data frame should be organized such that each row refers to a subject and each column to a SNP. The last column must be a binary response for each subject. The data frame must contain at least 8 columns. Rows containing any missing values (i.e. NAs) are omitted.
dimens	The number of possible values for each column of data. Each possible value does not need to occur in data. All entries of dimens must be greater than or equal to 2.
alpha	A hyperparameter of the prior representing the total of a fictive contingency table with counts equal to alpha divided by the number of cells. Alpha must be a positive real number.

Value

A list with a data frame and a vector:

recoded_data	The recoded dataset.
recoded_dimens	The revised dimension vector.

Author(s)

Matthew Friedlander, Adrian Dobra, Helene Massam, and Laurent Briollais

References

[1] Massam, H., Liu, J. and Dobra, A. (2009). A conjugate prior for discrete hierarchical log-linear models. Annals of Statistics, 37, 3431-3467.

[2] Dobra, A., Briollais, L., Jarjanazi, H., Ozcelik, H. and Massam, H. (2010). Applications of the mode oriented stochastic search (MOSS) algorithm for discrete multi-way data to genomewide studies. Bayesian Modeling in Bioinformatics, Taylor & Francis (Dey, D., Ghosh, S., and Mallick, B., eds.), 63-93.

[3] Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. Statistical Methodology, 7, 240-253.

Examples

```
data(simuCC)
data <- simuCC[,c(1002,2971,rep(5978:6001))]
# The SNPs in columns 1002 and 2971 of simuCC called rs4491689 and rs6869003 cause the disease.
set.seed(7)
r <- recode_data (data, dimens = c(rep(3,25),2), alpha = 1)
s <- mWindow (data = r$recoded_data, dimens = r$recoded_dimens, alpha = 1, windowSize = 2)
head (s, n = 5)
```

---

simuCC *A simulated case-control sample.*

Description

The data was simulated using code from the simuPOP cookbook on simupop.sourceforge.net. It is a sample of 1000 cases and 1000 controls from a simulated but realistic population. It contains the genotype information at 6000 diallelic SNPs (i.e. SNPs with three categories) and the disease status for each individual (which is the last column). The markers 'rs4491689' and 'rs6869003' are associated with the disease.

Usage

```
data(simuCC)
```

Source

The python code for simulating the data is example2.py on http://simupop.sourceforge.net/cookbook/pmwiki.php/Cookbook/
See the references for more information on how the data were generated.

References

[1] Peng, B. and Amos, C. (2010). Forward-time simulation of realistic samples for genome-wide association studies. BMC Bioinformatics, 11, 442-453.

[2] Peng, B. and Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. Bioinformatics, 21, 3686-3687.

Index