Package ‘aml’

February 19, 2015

Version 0.1-1
Date 2013-11-26
Title Adaptive Mixed LASSO
Author Dong Wang
Maintainer Dong Wang <dwangstat@gmail.com>
Depends R (>= 2.10), lars
Suggests MASS
Description This package implements the adaptive mixed lasso (AML) method proposed by Wang et al.(2011). AML applies adaptive lasso penalty to a large number of predictors, thus producing a sparse model, while accounting for the population structure in the linear mixed model framework. The package here is primarily designed for application to genome wide association studies or genomic prediction in plant breeding populations, though it could be applied to other settings of linear mixed models.
License GPL (>= 2)
URL http://www.r-project.org
NeedsCompilation no
Repository CRAN
Date/Publication 2013-11-28 08:14:36

R topics documented:

aml.estimate .......................................................... 2
aml.prd.outside ....................................................... 3
amltest ................................................................. 5
cleanclust ............................................................. 7
epigen ................................................................. 9
wheat ................................................................. 11
Index 12
aml.estimate

Estimate Genetic Values Using AML Fit

Description

This function calculate the genetic values using the result from amltest.

Usage

aml.estimate(fit, marker, response, kin)

Arguments

fit An object generated by amltest to be used to calculate genetic values.
marker A matrix or data frame for the markers (or genetic effects). It should be the same one used by amltest to generate fit.
response A numerical vector of trait (phenotype) values. It should be the same one used by amltest to generate fit.
kin The kinship matrix representing relationships between lines. It should be the same one used by amltest to generate fit.

Details

This function is used to genetic values, i.e., fitted phenotypic values using genetic marker information. It requires an adaptive mixed LASSO model has been fitted to the lines using amltest and the result is given in fit. Thus this function will only calculate genetic values for lines with observed phenotypic values. To make prediction for lines with only genetic information but no observed phenotypic values, use the function aml.pred.outside.

Value

A numeric vector containing estimated genetic values for lines analyzed by amltest to generate fit.

References


See Also

amltest, aml.pred.outside.
Examples

```r
## estimate genetic values for lines in the wheat data set
data("wheat")
clmarker <- cleanclust(wheat$marker, nafrac=0.2, mafb=0.1, corbnd=0.5, method="complete")
intermat <- epigen(wheat$y, clmarker$newmarker, wheat$a, numkeep=100, selectvar=30, corbnd=0.5, mafb=0.04)
resep <- amltest(wheat$y, intermat$effects, wheat$a, numkeep=80, selectvar=40)
predall <- aml.estimate(resep, intermat$effects, wheat$y, wheat$a)
```

aml.pred.outside  Prediction With Adaptive Mixed LASSO

Description

This function is used to predict the genetic values for lines with marker information after fitting adaptive mixed LASSO on a training set using amltest.

Usage

```r
aml.pred.outside(marker, response, kin, which.pred, numkeep, selectvar)
```

Arguments

- **marker**: A matrix or data frame for the markers (or genetic effects). It should include both lines with observed phenotypic information (those in the training set) and those lines for which the genetic values will be predicted.
- **response**: A numerical vector of trait (phenotype) values, corresponding to the lines in marker. For lines for which prediction will be made, the trait values can simply be set to NA. But for lines to be used in the training set, the trait value cannot be missing.
- **kin**: The kinship matrix representing relationships between lines. It should correspond to the rows in marker and represent the relationships between lines in the training set as well as those to be predicted.
- **which.pred**: A vector of integers specifying for which lines (which rows in marker) the prediction should be made. Lines not in which.pred will be used as the training set.
- **numkeep**: This parameter is passed to amltest. It should be less than the number of lines in the training set.
- **selectvar**: This parameter is passed to amltest. It should be less than numkeep.
Details

This function uses both marker effects and kinship to predict genetic values. Thus the kinship matrix should include both lines in the training set and the lines on which predictions are to be made. An adaptive mixed LASSO model is fitted for the training set including lines not in which.pred. The regression coefficients provided by amltest are then used for prediction. Besides performing prediction for lines with genetic marker genotypes but no phenotypic values, this function is especially convenient for performing cross-validation.

Value

A list of the following:

- predict.vl: The vector of predicted genetic values for lines specified in which.pred.
- response.vl: The vector of observed phenotypic values for lines specified in which.pred. This is useful for cross-validation when comparing predicted and observed values. Otherwise it might be a vector of NAs.

References


See Also

amltest, aml.estimate.

Examples

```r
## Predict the phenotype values of ten lines using the rest of the population in the wheat data
data("wheat")
clmarker<- cleanclust(wheat$marker, nafrac=0.2, mafb=0.1, corbnd=0.5, method="complete")
intermat <- epigen(wheat$y, clmarker$newmarker, wheat$a, numkeep=100, selectvar=30, corbnd=0.5, maf=-0.05)
which10<- sample(1:282, 10)
pred10<- aml.pred.outside(intermat$effects, wheat$y, wheat$a, which10, 80, 40)
```
Adaptive Mixed LASSO Analysis

Description

Perform adaptive mixed LASSO analysis. The function is designed for association mapping or genomic prediction in structured populations, though other applications are possible.

Usage

```r
amltest(response, marker, kin, numkeep=floor(length(response)*.5), selectvar)
```

Arguments

- **response**
  A numerical vector of the trait (phenotype) to be analyzed.
- **marker**
  A matrix or data frame for the marker (or more generally, genetic effect) information. The number of rows should equal the number of lines and the number of columns should equal the number of markers. The values of each element should be between 0 and 1 with minor allele encoded as 1 and majority allele as 0. If minor allele is encoded as 1 instead for some markers, `cleanclust` can be used to re-encode it. The function `cleanclust` should also be used to preprocess the marker data to remove marker with a high proportion of missing values or very low minor allele frequency as well as impute missing values with the sample mean. It is also recommend that `cleanclust` be used to filter the markers so that no markers are highly correlated.
- **kin**
  The kinship matrix representing relationships between lines. It should be symmetric and positive definite, and have the number of rows and columns equal to the number of rows of `marker`.
- **numkeep**
  The number of markers that should be retained after the preliminary screening. It should be less than the number of lines. The default value is a half of the number of lines. See Details.
- **selectvar**
  The number of markers to be included in the model. Strictly speaking, it is the number of iterations for the fitting procedure. The number of markers in the output could be slightly less than `selectvar`. See Details.

Details

In adaptive mixed LASSO fitting, `amltest` first performs a preliminary screening to retain a set of markers (predictors) numbering at most `numkeep`, which should be less than the number of lines. This step relies on LASSO fitting using `lars`. The quantity `numkeep` is the maximum steps of iterations in LASSO fit. Due to the nature of the `lars` algorithm, the number of markers retained after the screening might be slightly less than `numkeep`. Then `amltest` will perform adaptive mixed LASSO fit by iteratively estimating the fixed effects and random effects up to the number of iterations defined by `selectvar`. Again, the number of markers in the output might be slightly less than `selectvar` as determined by the behavior of the `lars` algorithm. So if an exact number of markers are required in the model, some trial and error might be needed.
Value

A list containing the following:

- **estimate**: A matrix of two columns. The first column indicates which column in marker is included in the model fit and the second column is the effect for each marker in the model.

- **AIC**: A vector of AIC values for models using different number of markers. The first entry is for model with zero markers (only random line effects) and the last entry corresponding to the model with markers specified in `estimate`.

- **BIC**: A vector of BIC values for models using different number of markers. The first entry is for model with zero markers (only random line effects) and the last entry corresponding to the model with markers specified in `estimate`.

- **EBIC**: A vector of EBIC values for models using different number of markers. The first entry is for model with zero markers (only random line effects) and the last entry corresponding to the model with markers specified in `estimate`.

- **vars**: The vector for variance components of random effects. The first entry is the genetic variance $\sigma_g^2$ and the second entry is the ratio of the error variance over the genetic variance. Thus the product of these two entries gives the error variance $\sigma_e^2$.

- **mcount**: The vector of the number of markers in each step. This is mainly used in conjunction with AIC, BIC, or EBIC.

References


See Also

- `cleanclust`.

Examples

```r
## analyze the wheat data with main marker effects.
data("wheat")
clmarker<- cleanclust(wheat$marker, nafrac=P.2, mafb=P.1, corbnd=P.5, method="complete")
resmain <- amltest(wheat$y, clmarker$newmarker, wheat$a, numkeep=80, selectvar=40)
```
**cleanclust**  
*Clean, Impute, and Filter Markers*

**Description**
Prepare marker data for use for amltest. This function can be used to remove markers with a high proportion of missing values, impute missing values with sample average, remove markers with very little variation, and if necessary, re-encode the minor allele as 1 and the majority allele as 0.

**Usage**
```r
cleanclust(marker, nafrac=0.2, mafb=0.1, corbnd=0.5, method="complete")
```

**Arguments**
- `marker`: A matrix or data frame for the marker information. The number of rows should equal the number of lines and the number of columns should equal the number of markers. The values of each element should be between 0 and 1 preferably with minor allele encoded as 1 and majority allele as 0. If minor allele is encoded as 1 instead for a marker, `cleanclust` change its value to 1 minus the original column. Each column has to have a unique name to identify the marker.
- `nafrac`: The maximum proportion of missing values for a marker. Markers with higher proportion of missing values will be removed. The default is 0.2.
- `mafb`: The minimum minor allele frequency, markers with lower minor allele frequency will be removed. The default is 0.1.
- `corbnd`: The bound used for cutting the dendrogram after the hierarchical clustering, the default is 0.5. See Details.
- `method`: The method of clustering passed to `hclust`. The values could be one of "complete", "average" or "single". The default is "complete".

**Details**
This is a simplified version of the Hclust method described in the paper *Characterization of Multilocus Linkage Disequilibrium* by Rinald, et al. (2005), tailored for use with amltest and other functions in this package. The R code for the original Hclust package can be find at [http://www.epic.Pitt.ed/Accompaniment/hclust/hclust.ht](http://www.epic.Pitt.ed/Accompaniment/hclust/hclust.ht), which provides more functionality.

The function `cleanclust` provides two main utilities. The first is to clean and impute the marker data, including removing markers with a high proportion of missing values or very low minor allele frequency as well as impute the remaining missing values by the sample mean regarding each marker. The second is to remove some markers when necessary so that no markers will be highly correlated. Like other LASSO type method, the performance of adaptive mixed LASSO can be improved when predictors are not highly correlated. This process follows that of Rinald et al. (2005). The correlation between each pair of markers are calculated and \( r = 1 - \text{cor}^2 \) is used as the distance between markers to perform hierarchical clustering with `hclust`. The resulted dendrogram is cut to form clusters according to the bound on \( \text{cor}^2 \), `corbnd`. Specifically, higher `corbnd` values
will result in less clusters being formed and less markers in the output. One marker is retained for each cluster in \texttt{newmarker}.

**Value**

A list containing the following:

- \texttt{newmarker} The new marker matrix after removing markers with a high proportion of missing values or low minor allele frequency, with missing values replaced with sample means, and possibly removing some markers to avoid multiple highly correlated markers.

- \texttt{flip} A vector of marker names for which the minor allele and major allele has been flipped. Other functions in this package require the minor allele to be encoded as 1 and major allele as 0. If the opposite is the case for a marker, the value will be flipped and the marker name will be given in this vector.

- \texttt{tagged} A vector of integers indicating which columns (markers) from the original marker matrix is retained in \texttt{newmarker}.

**References**


**See Also**

\texttt{amltest}.

**Examples**

```r
## process the markers in the wheat data set.
data("wheat")
clmarker<- cleanclust(wheat$marker, nafrac=0.2, mafb=0.1, corbnd=0.5, method="complete")```
**epigen**  

*Generate Epistatic Effect Matrix*

**Description**

This function selects a specified number of markers using *amltest* and then forms a matrix including both main marker effects and two-way epistatic effects.

**Usage**

```r
epigen(response, marker, kin, numkeep=floor(length(response)*.5), selectvar, corbnd=0.5, mafb=0.04, method="complete")
```

**Arguments**

- **response**  
  A numerical vector of the trait (phenotype) to be analyzed. It is passed to *amltest*.

- **marker**  
  A matrix or data frame for the markers from which the main effects will be selected. The number of rows should equal the number of lines and the number of columns should equal the number of markers. The values of each element should be between 0 and 1 with the minor allele encoded as 1 and the majority allele as 0. If the minor allele is encoded as 1 instead for some markers, *cleanclust* can be used to re-encode it. The function *cleanclust* should also be used to preprocess the marker data to remove marker with a high proportion of missing values or a very low minor allele frequency as well as impute missing values with the sample mean. It is also recommended that *cleanclust* be used to filter the markers so that no markers are highly correlated. It is passed to *amltest*.

- **kin**  
  The kinship matrix representing relationships between lines. It should be symmetric and positive definite, and have the number of rows and columns equal to the number of rows of *marker*. It is passed to *amltest*.

- **numkeep**  
  The number of main marker effects that should be retained after the preliminary screening in *amltest*. It should be less than the number of lines. The default value is a half of the number of lines.

- **selectvar**  
  The number of main marker effects to be included in the model. Strictly speaking, it is the number of iterations for the fitting procedure of *amltest*. The number of main marker effects that are retained could be slightly less than *selectvar*. See the documentation for *amltest*.

- **mafb**  
  The minimum mean value of an effect. Effects with lower mean values (too many zeros) are removed. For a main marker effect, this is just the minimum value for minor allele frequency. The default is 0.04 and is passed to *cleanclust*.

- **corbnd**  
  The bound used for cutting the dendrogram after the hierarchical clustering, the default is 0.5. See the documentation for *cleanclust*.
method

The method of clustering passed to hclust. The values could be one of "complete", "average" or "single". The default is "complete". See the documentation for cleanclust.

Details

Since considering all two-way epistatic effects are not computationally feasible in most cases, amltest is called first to select a subset of markers with the most significant main effects. Then two-way epistatic effects are formed from these selected markers by taking the product of the two columns corresponding to each pair of markers. Subsequently, the cleanclust function is called to remove effects with very low mean values and also filter the effects such that no two effects are highly correlated. The resulted genetic effect matrix include both main effects and epistatic effects. It can then be used as input for amltest in the same manner as a marker matrix.

Value

A list containing the following:

- effects: A matrix of both selected main marker effects and two-way epistatic effects.
- marker1: A vector of names corresponding to the first marker in two-way epistatic effects given in effects, or the marker name for a main effect.
- marker2: A vector of names corresponding to the second marker in two-way epistatic effects given in effects, or the marker name for a main effect.

References


See Also

amltest, cleanclust.

Examples

```r
## process the markers in the wheat data set.
data("wheat")
clmarker<- cleanclust(wheat$marker, nafrac=0.2, mafb=0.1, corbnd=0.5, method="complete")
intermat <- epigen(wheat$y, clmarker$newmarker, wheat$a, numkeep=100, selectvar=30,
corbnd=0.5, mafb=0.04)
```
wheat

Genetic and Phenotype Data of a Wheat Breeding Population

Description
This data set gives the genetic marker information (DArT markers), phenotype, and the kinship matrix of a wheat breeding population. The list wheat contains three elements, ‘y’, ‘marker’ and ‘A’. The vector ‘y’ contains the phenotypic values for 282 wheat accessions, ‘A’ is a relationship (kinship) matrix of all the wheat accessions, and ‘marker’ is a 282 by 300 data frame for the genotype on 300 DArT markers for the wheat accessions.

Usage
data(wheat)

Format
A list of three elements.

Source
Nebraska Wheat Breeding

References
Index

*Topic adaptive mixed LASSO
   aml.estimate, 2
   aml.pred.outside, 3
   amltest, 5
   cleanclust, 7
   epigen, 9
*Topic datasets
   wheat, 11

aml.estimate, 2, 4
aml.pred.outside, 2, 3
amltest, 2, 4, 5, 8, 10

  cleanclust, 6, 7, 10

  epigen, 9

  wheat, 11