Package ‘GSelection’

January 20, 2025

Type Package

Title Genomic Selection

Version 0.1.0

Author Sayanti Guha Majumdar, Anil Rai, Dwijesh Chandra Mishra

Maintainer Sayanti Guha Majumdar <sayanti23gm@gmail.com>

Description Genomic selection is a specialized form of marker assisted selection. The package contains functions to select important genetic markers and predict phenotype on the basis of fitted training data using integrated model framework (Guha Majumdar et. al. (2019) <doi:10.1089/cmb.2019.0223>) developed by combining one additive (sparse additive models by Ravikumar et. al. (2009) <doi:10.1111/j.1467-9868.2009.00718.x>) and one non-additive (hsic lasso by Yamada et. al. (2014) <doi:10.1162/NECO_a_00537>) model.

License GPL-3

Encoding UTF-8

LazyData true

Imports SAM, penalized, gdata, stats, utils

RoxygenNote 6.1.1

Depends R (>= 3.5)

NeedsCompilation no

Repository CRAN

Date/Publication 2019-11-04 16:30:27 UTC

Contents

`GSelection-package` *Genomic Selection*

Description

Genomic selection is a specialized form of marker assisted selection. The package contains functions to select important genetic markers and predict phenotype on the basis of fitted training data using integrated model framework (Guha Majumdar et. al. (2019) <doi:10.1089/cmb.2019.0223>) developed by combining one additive (sparse additive models by Ravikumar et. al. (2009) <doi:10.1111/j.1467-9868.2009.00718.x>) and one non-additive (hsic lasso by Yamada et. al. (2014) <doi:10.1162/NECO_a_00537>) model.

Details

The DESCRIPTION file:

Package:	GSelection
Type:	Package
Title:	Genomic Selection
Version:	0.1.0
Author:	Sayanti Guha Majumdar, Anil Rai, Dwijesh Chandra Mishra
Maintainer:	Sayanti Guha Majumdar <sayanti23gm@gmail.com>
Description:	Genomic selection is a specialized form of marker assisted selection. The package contains functions to
License:	GPL-3
Encoding:	UTF-8
LazyData:	true
Imports:	SAM, penalized, gdata, stats, utils
RoxygenNote:	6.1.1
Depends:	R (>= 3.5)
NeedsCompilation:	no
Packaged:	2019-10-26 10:25:25 UTC; user6

Index of help topics:

```
GS                      Genotypic and phenotypic simulated dataset
GSelection-package      Genomic Selection
RED                     Redundancy Rate
feature.selection       Genomic Feature Selection
genomic.prediction      Genomic Prediction
hsic.var.ensemble       Error Variance Estimation in Genomic Prediction
hsic.var.rcv            Error Variance Estimation in Genomic Prediction
spam.var.ensemble       Error Variance Estimation in Genomic Prediction
spam.var.rcv            Error Variance Estimation in Genomic Prediction
```

Author(s)

Sayanti Guha Majumdar, Anil Rai, Dwijesh Chandra Mishra

Maintainer: Sayanti Guha Majumdar <sayanti23gm@gmail.com>

References

Guha Majumdar, S., Rai, A. and Mishra, D. C. (2019). Integrated framework for selection of additive and non-additive genetic markers for genomic selection. *Journal of Computational Biology*. doi:10.1089/cmb.2019.0223
Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009-1030. doi:10.1111/j.1467-9868.2009.00718.x
Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2014). High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, 26(1):185-207. doi:10.1162/NECO_a_00537

`feature.selection` *Genomic Feature Selection*

Description

Feature (marker) selection in case of genomic prediction with integrated model framework using both additive (Sparse Additive Models) and non-additive (HSIC LASSO) statistical models.

Usage

```
feature.selection(x,y,d)
```

Arguments

`x`	a matrix of markers or explanatory variables, each column contains one marker and each row represents an individual.
`y`	a column vector of response variable.
`d`	number of variables to be selected from x.

Details

Integrated model framework was developed by combining one additive model (Sparse Additive Model) and one non-additive model (HSIC LASSO) for selection of important markers from whole genome marker data.

Value

Returns a LIST containing

`spam_selected_feature_index`
: returns index of selected markers from x using Sparse Additive Model

`coefficient.spam`
: returns coefficient values of selected markers using Sparse Additive Model.

`hsic_selected_feature_index`
: returns index of selected markers from x using HSIC LASSO.

`coefficient.hsic`
: returns coefficient values of selected markers using HSIC LASSO.

`integrated_selected_feature_index`
: returns index of selected markers from x using integrated model framework.

Author(s)

Sayanti Guha Majumdar <<sayanti23gm@gmail.com>>, Anil Rai, Dwijesh Chandra Mishra

References

Guha Majumdar, S., Rai, A. and Mishra, D. C. (2019). Integrated framework for selection of additive and non-additive genetic markers for genomic selection. *Journal of Computational Biology*. doi:10.1089/cmb.2019.0223
Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009-1030. doi:10.1111/j.1467-9868.2009.00718.x
Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2014). High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, 26(1):185-207. doi:10.1162/NECO_a_00537

Examples

```
library(GSelection)
data(GS)
x_trn <- GS[1:40,1:110]
y_trn <- GS[1:40,111]
x_tst <- GS[41:60,1:110]
y_tst <- GS[41:60,111]
fit <- feature.selection(x_trn,y_trn,d=10)
```

genomic.prediction *Genomic Prediction*

Description

Prediction of phenotypic values based on selected markers with integrated model framework using both additive (Sparse Additive Models) and non-additive (HSIC LASSO) statistical models.

Usage

```
genomic.prediction(x,spam_error_var,hsic_error_var,
spam_selected_feature_index,hsic_selected_feature_index,
coefficient.spam,coefficient.hsic)
```

Arguments

`x` a matrix of markers or explanatory variables for which phenotype will be predicted. Each column contains one marker and each row represents an individual.

`spam_error_var` estimated error variance of genomic prediction by Sparse Additive Model.

`hsic_error_var` estimated error variance of genomic prediction by HSIC LASSO.

`spam_selected_feature_index`
index of selected markers from x using Sparse Additive Model

`hsic_selected_feature_index`
index of selected markers from x using HSIC LASSO.

`coefficient.spam`
coefficient values of selected markers using Sparse Additive Model.

`coefficient.hsic`
coefficient values of selected markers using HSIC LASSO.

Details

Phenotypic values will be predicted for given genotype of markers by using previously fitted model object. Integrated model framework is used for this purpose which is developed by combining selected features from SpAm and HSIC LASSO.

Value

`Integrated_y` returns predicted phenotype

Author(s)

Sayanti Guha Majumdar <`<sayanti23gm@gmail.com>`>, Anil Rai, Dwijesh Chandra Mishra

References

Guha Majumdar, S., Rai, A. and Mishra, D. C. (2019). Integrated framework for selection of additive and non-additive genetic markers for genomic selection. *Journal of Computational Biology*. doi:10.1089/cmb.2019.0223
Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009-1030. doi:10.1111/j.1467-9868.2009.00718.x
Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2014). High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, 26(1):185-207. doi:10.1162/NECO_a_00537

Examples

```
library(GSelection)
data(GS)
x_trn <- GS[1:40,1:110]
y_trn <- GS[1:40,111]
x_tst <- GS[41:60,1:110]
y_tst <- GS[41:60,111]

## estimate spam_var from function spam.var.ensemble or spam.var.rcv
spam_var <- 2.681972
## estimate hsic_var from function hsic.var.ensemble or hsic.var.rcv
hsic_var <- 10.36974

fit <- feature.selection(x_trn,y_trn,d=10)
pred_y <- genomic.prediction(x_tst,spam_var,hsic_var,
fit$spam_selected_feature_index,fit$hsic_selected_feature_index,
fit$coefficient.spam,fit$coefficient.hsic)
```

GS *Genotypic and phenotypic simulated dataset*

Description

This dataset is simulated with the R package "qtlbim" where 10 are true features associated with the trait of study and remaining 100 are random markers. we consider 10 chromosomes each containing 10 markers. Each chromosome have 1 qtl which is the true feature.

Usage

```
data("GS")
```

Format

A data frame with 60 rows as genotypes with 111 columns (i.e. contains information of genotyped markers and phenotypic traits).

Details

It has total 60 rows which represents 200 individuals genotypes and a total of 111 of columns, in which first 110 columns contain information of genotyped markers and last column represents value of phenotypic trait associated with genotype under study.

Source

Yandell, B. S., Mehta, T., Banerjee, S., Shriner, D., Venkataraman, R. et al. (2007). R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses. *Bioinformatics*, 23, 641-643.
Yandell, B. S., Nengjun, Y., Mehta, T., Banerjee, S., Shriner, D. et al. (2012). qtlbim: QTL Bayesian Interval Mapping. R package version 2.0.5. http://CRAN.R-project.org/package=qtlbim

Examples

```
library(GSelection)
data(GS)
X<-GS[,1:110]   ## Extracting Genotype
Y<-GS[,111] ## Extracting Phenotype
```

`hsic.var.ensemble` *Error Variance Estimation in Genomic Prediction*

Description

Estimation of error variance using Ensemble method which combines bootstraping and sampling with srswor in HSIC LASSO.

Usage

```
hsic.var.ensemble(x,y,b,d)
```

Arguments

`x` a matrix of markers or explanatory variables, each column contains one marker and each row represents an individual.

`y` a column vector of response variable.

`b` number of bootstrap samples.

`d` number of variables to be selected from x.

Details

In this method, both bootstrapping and simple random sampling without replacement are combined to estimate error variance. Variables are selected using HSIC LASSO from the original datasets and all possible samples of a particular size are taken from the selected variables set with simple random sampling without replacement. With these selected samples error variance is estimated from bootstrap samples of the original datasets using least squared regression method. Finally the average of all the estimated variances is considered as the final estimate of the error variance.

Value

`Error variance`

Author(s)

Sayanti Guha Majumdar <`<sayanti23gm@gmail.com>`>, Anil Rai, Dwijesh Chandra Mishra

References

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2014). High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, 26(1):185-207. doi:10.1162/NECO_a_00537

Examples

```
library(GSelection)
data(GS)
x_trn <- GS[1:40,1:110]
y_trn <- GS[1:40,111]
x_tst <- GS[41:60,1:110]
y_tst <- GS[41:60,111]
hsic_var <- hsic.var.ensemble(x_trn,y_trn,2,10)
```

---

hsic.var.rcv *Error Variance Estimation in Genomic Prediction*

Description

Estimation of error variance using Refitted Cross Validation in HSIC LASSO.

Usage

```
hsic.var.rcv(x,y,d)
```

Arguments

x	a matrix of markers or explanatory variables, each column contains one marker and each row represents an individual.
y	a column vector of response variable.
d	number of variables to be selected from x.

Details

Refitted cross validation method (RCV) which is a two step method, is used to get the estimate of the error variance. In first step, dataset is divided into two sub-datasets and with the help of HSIC LASSO most significant markers(variables) are selected from the two sub-datasets. This results in two small sets of selected variables. Then using the set selected from 1st sub-dataset error variance is estimated from the 2nd sub-dataset with ordinary least square method and using the set selected from the 2nd sub-dataset error variance is estimated from the 1st sub-dataset with ordinary least square method. Finally the average of those two error variances are taken as the final estimator of error variance with RCV method.

Value

`Error variance`

Author(s)

Sayanti Guha Majumdar <<`sayanti23gm@gmail.com`>>, Anil Rai, Dwijesh Chandra Mishra

References

Fan, J., Guo, S., Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society*, 74(1), 37-65.
Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2014). High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, 26(1):185-207. doi:10.1162/NECO_a_00537

Examples

```
library(GSelection)
data(GS)
x_trn <- GS[1:40,1:110]
y_trn <- GS[1:40,111]
x_tst <- GS[41:60,1:110]
y_tst <- GS[41:60,111]
hsic_var <- hsic.var.rcv(x_trn,y_trn,10)
```

RED *Redundancy Rate*

Description

Calculate the redundancy rate of the selected features(markers). Value will be high if many redundant features are selected.

Usage

```
RED(x,spam_selected_feature_index,hsic_selected_feature_index,
integrated_selected_feature_index)
```

Arguments

`x` a matrix of markers or explanatory variables, each column contains one marker and each row represents an individual.

`spam_selected_feature_index`
index of selected markers from x using Sparse Additive Model.

`hsic_selected_feature_index`
index of selected markers from x using HSIC LASSO.

`integrated_selected_feature_index`
index of selected markers from x using integrated model framework

Details

The RED score (Zhao et al., 2010) is determined by average of the correlation between each pair of selected markers. A large RED score signifies that selected features are more strongly correlated to each other which means many redundant features are selected. Thus, a small redundancy rate is preferable for feature selection.

Value

Returns a LIST containing

`RED_spam` returns redundancy rate of features selected by using Sparse Additive Model.

`RED_hsic` returns redundancy rate of features selected by using HSIC LASSO.

`RED_I` returns redundancy rate of features selected by using integrated model framework.

Author(s)

Sayanti Guha Majumdar <<`sayanti23gm@gmail.com`>>, Anil Rai, Dwijesh Chandra Mishra

References

Guha Majumdar, S., Rai, A. and Mishra, D. C. (2019). Integrated framework for selection of additive and non-additive genetic markers for genomic selection. *Journal of Computational Biology*. doi:10.1089/cmb.2019.0223
Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009-1030. doi:10.1111/j.1467-9868.2009.00718.x
Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2014). High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. *Neural Computation*, 26(1):185-207. doi:10.1162/NECO_a_00537
Zhao, Z., Wang, L. and Li, H. (2010). Efficient spectral feature selection with minimum redundancy. *In AAAI Conference on Artificial Intelligence (AAAI)*, pp 673-678.

Examples

```
library(GSelection)
data(GS)
x_trn <- GS[1:40,1:110]
y_trn <- GS[1:40,111]
x_tst <- GS[41:60,1:110]
y_tst <- GS[41:60,111]
fit <- feature.selection(x_trn,y_trn,d=10)
red <- RED(x_trn,fit$spam_selected_feature_index,fit$hsic_selected_feature_index,
fit$integrated_selected_feature_index)
```

`spam.var.ensemble` *Error Variance Estimation in Genomic Prediction*

Description

Estimation of error variance using Ensemble method which combines bootstraping and sampling with srswor in Sparse Additive Models.

Usage

```
spam.var.ensemble(x,y,b,d)
```

Arguments

x	a matrix of markers or explanatory variables, each column contains one marker and each row represents an individual.
y	a column vector of response variable.
b	number of bootstrap samples
d	number of variables to be selected from x.

Details

In this method, both bootstrapping and simple random sampling without replacement are combined to estimate error variance. Variables are selected using Sparse Additive Models (SpAM) from the original datasets and all possible samples of a particular size are taken from the selected variables set with simple random sampling without replacement. With these selected samples error variance is estimated from bootstrap samples of the original datasets using least squared regression method. Finally the average of all the estimated variances is considered as the final estimate of the error variance.

Value

Error variance

Author(s)

Sayanti Guha Majumdar <<sayanti23gm@gmail.com>>, Anil Rai, Dwijesh Chandra Mishra

References

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009-1030. doi:10.1111/j.1467-9868.2009.00718.x

Examples

```
library(GSelection)
data(GS)
x_trn <- GS[1:40,1:110]
y_trn <- GS[1:40,111]
x_tst <- GS[41:60,1:110]
y_tst <- GS[41:60,111]
spam_var <- spam.var.ensemble(x_trn,y_trn,2,10)
```

`spam.var.rcv` *Error Variance Estimation in Genomic Prediction*

Description

Estimation of error variance using Refitted cross validation in Sparse Additive Models.

Usage

```
spam.var.rcv(x,y,d)
```

Arguments

`x` a matrix of markers or explanatory variables, each column contains one marker and each row represents an individual.

`y` a column vector of response variable.

`d` number of variables to be selected from x.

Details

Refitted cross validation method (RCV) which is a two step method, is used to get the estimate of the error variance. In first step, dataset is divided into two sub-datasets and with the help of Sparse Additive Models (SpAM) most significant markers(variables) are selected from the two sub-datasets. This results in two small sets of selected variables. Then using the set selected from 1st sub-dataset error variance is estimated from the 2nd sub-dataset with ordinary least square method and using the set selected from the 2nd sub-dataset error variance is estimated from the 1st sub-dataset with ordinary least square method. Finally the average of those two error variances are taken as the final estimator of error variance with RCV method.

Value

`Error variance`

Author(s)

Sayanti Guha Majumdar <<`sayanti23gm@gmail.com`>>, Anil Rai, Dwijesh Chandra Mishra

References

Fan, J., Guo, S., Hao, N. (2012).Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society*, 74(1), 37-65.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5), 1009-1030. doi:10.1111/j.1467-9868.2009.00718.x

Examples

```
library(GSelection)
data(GS)
x_trn <- GS[1:40,1:110]
y_trn <- GS[1:40,111]
x_tst <- GS[41:60,1:110]
y_tst <- GS[41:60,111]
spam_var <- spam.var.rcv(x_trn,y_trn,10)
```

Index