Package 'varEst'

October 12, 2022

Type Package

Title Variance Estimation

Version 0.1.0

Author Sayanti Guha Majumdar, Anil Rai, Dwijesh Chandra Mishra

Maintainer Sayanti Guha Majumdar <sayanti23gm@gmail.com>

Description Error variance estimation in ultrahigh dimensional datasets with four different methods, viz. Refitted cross validation, k-fold refitted cross validation, Bootstrap-refitted cross validation, Ensemble method.

License GPL-3

Encoding UTF-8

LazyData TRUE

Imports SAM, caret, lm.beta, glmnet

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-09-23 16:10:02 UTC

R topics documented:

---

varEst-package	*Variance Estimation*

Description

Error variance estimation in ultrahigh dimensional datasets with four different methods, viz. Refitted cross validation, k-fold refitted cross validation, Bootstrap-refitted cross validation, Ensemble method.

Details

The DESCRIPTION file:

Package:	varEst
Type:	Package
Title:	Variance Estimation
Version:	0.1.0
Author:	Sayanti Guha Majumdar, Anil Rai, Dwijesh Chandra Mishra
Maintainer:	Sayanti Guha Majumdar <sayanti23gm@gmail.com>
Description:	Error variance estimation in ultrahigh dimensional datasets with four different methods, viz. Refitted cross va
License:	GPL-3
Encoding:	UTF-8
LazyData:	TRUE
Imports:	SAM, caret, lm.beta, glmnet
RoxygenNote:	6.1.1

Index of help topics:

```
bsrcv              Variance Estimation with Bootstrap-RCV
ensemble           Variance Estimation with Ensemble method
krcv               Variance Estimation with kfold-RCV
rcv                Variance Estimation with Refitted Cross
                   Validation(RCV)
varEst-package     Variance Estimation
```

Author(s)

Sayanti Guha Majumdar, Anil Rai, Dwijesh Chandra Mishra

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

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society*, 58, 267-288

---

bsrcv	*Variance Estimation with Bootstrap-RCV*

Description

Estimation of error variance using Bootstrap-refitted cross validation method in ultrahigh dimensional dataset.

Usage

```
bsrcv(x,y,a,b,d,method=c("spam","lasso","lsr"))
```

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.
a	value of alpha, range is 0<=a<=1 where, a=1 is LASSO penalty and a=0 is Ridge penalty.If variable selection method is LASSO then providing value to a is compulsory. For other methods a should be NULL.
b	number of bootstrap samples.
d	number of variables to be selected from x.
method	variable selection method, user can choose any method among "spam", "lasso", "lsr"

Details

In this method, bootstrap samples are taken from the original datasets and then RCV (Fan et al., 2012) method is applied to each of these bootstrap samples.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society*, 58, 267-288

Examples

```
## data simulation
marker <- as.data.frame(matrix(NA, ncol =500, nrow = 200))
for(i in 1:500){
marker[i] <- sample(1:3, 200, replace = TRUE, prob = c(1, 2, 1))
}
pheno <- marker[,1]*1.41+marker[,2]*1.41+marker[,3]*1.41+marker[,4]*1.41+marker[,5]*1.41

pheno <- as.matrix(pheno)
marker<- as.matrix(marker)

## estimation of error variance
var <- bsrcv(marker,pheno,1,10,5,"lasso")
```

---

ensemble   *Variance Estimation with Ensemble method*

Description

Estimation of error variance using ensemble method which combines bootstraping and sampling with srswor in ultrahigh dimensional dataset.

Usage

```
ensemble(x,y,a,b,d,method=c("spam","lasso","lsr"))
```

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.
a	value of alpha, range is 0<=a<=1 where, a=1 is LASSO penalty and a=0 is Ridge penalty.If variable selection method is LASSO then providing value to a is compulsory. For other methods a should be NULL.
b	number of bootstrap samples.
d	number of variables to be selected from x.
method	variable selection method, user can choose any method among "spam", "lasso", "lsr"

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) or LASSO or least squared regression (lsr) 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
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society*, 58, 267-288

Examples

```
## data simulation
marker <- as.data.frame(matrix(NA, ncol =500, nrow = 200))
for(i in 1:500){
marker[i] <- sample(1:3, 200, replace = TRUE, prob = c(1, 2, 1))
}
pheno <- marker[,1]*1.41+marker[,2]*1.41+marker[,3]*1.41+marker[,4]*1.41+marker[,5]*1.41

pheno <- as.matrix(pheno)
marker<- as.matrix(marker)

## estimation of error variance
var <- ensemble(marker,pheno,1,10,10,"spam")
```

---

krcv	*Variance Estimation with kfold-RCV*

Description

Estimation of error variance using k-fold refitted cross validation in ultrahigh dimensional dataset.

Usage

```
krcv(x,y,a,k,d,method=c("spam","lasso","lsr"))
```

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.
a	value of alpha, range is 0<=a<=1 where, a=1 is LASSO penalty and a=0 is Ridge penalty.If variable selection method is LASSO then providing value to a is compulsory. For other methods a should be NULL.

k	dataset is divided into this many numbers of sub-datasets.
d	number of variables to be selected from x.
method	variable selection method, user can choose any method among "spam", "lasso", "lsr"

Details

The error variance is estimated from a high dimensional datasets where number of parameters are more than number of individuals, i.e. $p > n$. k-fold RCV is an extended version of original RCV method (Fan et al., 2012). In this case the datasets are divided into k equal size groups instead of 2 groups. Variables are selected using Sparse Additive Models (SpAM) or LASSO or least squared regression (lsr) from one group and variance is estimated using selected variables with ordinary least squared estimation from rest of the k-1 groups. Likewise, all the groups are covered and in the end, average value of all the variances from each group is the final error variance.

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
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society*, 58, 267-288

Examples

```
## data simulation
marker <- as.data.frame(matrix(NA, ncol =500, nrow = 200))
for(i in 1:500){
marker[i] <- sample(1:3, 200, replace = TRUE, prob = c(1, 2, 1))
}
pheno <- marker[,1]*1.41+marker[,2]*1.41+marker[,3]*1.41+marker[,4]*1.41+marker[,5]*1.41

pheno <- as.matrix(pheno)
marker<- as.matrix(marker)

## estimation of error variance
var <- krcv(marker,pheno,1,4,5,"spam")
```

---

rcv   *Variance Estimation with Refitted Cross Validation(RCV)*

Description

Estimation of error variance using Refitted cross validation in ultrahigh dimensional dataset.

Usage

```
rcv(x,y,a,d,method=c("spam","lasso","lsr"))
```

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.
a	value of alpha, range is 0<=a<=1 where, a=1 is LASSO penalty and a=0 is Ridge penalty. If variable selection method is LASSO then providing value to a is compulsory. For other methods a should be NULL.
d	number of variables to be selected from x.
method	variable selection method, user can choose any method among "spam", "lasso", "lsr"

Details

The error variance is estimated from a high dimensional datasets where number of parameters are more than number of individuals, i.e. p > n. 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) or LASSO or least squared regression (lsr) 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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society*, 58, 267-288

Examples

```
## data simulation
marker <- as.data.frame(matrix(NA, ncol =500, nrow = 200))
for(i in 1:500){
marker[i] <- sample(1:3, 200, replace = TRUE, prob = c(1, 2, 1))
}
pheno <- marker[,1]*1.41+marker[,2]*1.41+marker[,3]*1.41+marker[,4]*1.41+marker[,5]*1.41

pheno <- as.matrix(pheno)
marker<- as.matrix(marker)

## estimation of error variance
var <- rcv(marker,pheno,1,5,"spam")
```

Index