Package ‘vbsr’

February 20, 2015

Type Package

Title Variational Bayes Spike Regression Regularized Linear Models

Version 0.0.5

Date 2014-06-05

Author Benjamin Logsdon

Maintainer Benjamin Logsdon <ben.logsdon@sagebase.org>

Depends R (>= 3.0.0)

Description Efficient algorithm for solving ultra-sparse regularized regression models using a variational Bayes algorithm with a spike (l0) prior. Algorithm is solved on a path, with coordinate updates, and is capable of generating very sparse models. There are very general model diagnostics for controlling type-1 error included in this package.

License GPL-2

Copyright Benjamin Logsdon 2014

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-06-05 22:50:33

R topics documented:

compute_KL ......................................................... 2
vbsr ................................................................. 3

Index 7
compute_KL

Compute an empirical Kullback Leibler (KL) divergence for an observed distribution of Z-statistics

Description

This function computes the KL divergence between an observed distribution of Z-statistics and the expected distribution, when truncating at a given percentile of the reference normal distribution.

Usage

compute_KL(Zmat, alpha, pval)

Arguments

- **Zmat**: Matrix of Z-statistics outputted from vbsr, where columns are Z-statistics of covariates computed at different values of the penalty parameter $l0/path$, and rows are covariates in the model.
- **alpha**: The inner percentile of the reference normal distribution to compare to, e.g., if alpha=0.99, the KL divergence will only be computed for the inner 99% quantile of the reference distribution. Allows for deviations in the tails of the distribution to be ignored.
- **pval**: If marginal pre-screening was performed originally, the P-value threshold used for the marginal screening.

Details

This function is a vbsr internal function that computes the KL divergence for the Z-statistic distribution output by vbsr if run on a grid of $l0/path$, and takes as input the inner quantile to compute the KL statistic with (alpha), and if there was already marginal pre-screening performed to remove the central part of the Z-statistic distribution (pval).

Value

- **kl_vec**: This is the observed KL statistic computed along the specified path of $l0/path$.
- **min_kl**: This is the minimum value of observed KL statistic.
- **mean_kl**: Random permutations are performed to determine the expected KL statistic given the number of covariates being tested, and the setting of alpha, pval. Useful for determining if the observed distribution is well approximated by a normal distribution for a given setting of $l0/path$ based on the KL statistic.
- **se_kl**: The error in the KL statistics from the random permutations. Good for determining the range of KL values that is reasonable given the model fits.
Note

This function is an internal function, and this functionality is included primarily to include the model fit functions proposed by Logsdon et al. 2012. The regular `vbsr` function with `post=0.95`, produces very similar results to the KL statistic using a liberal cutoff, and `post=0.5` produces very similar results to the more conservative cutoff proposed in Logsdon et al. 2012, and the post approaches are much more computationally efficient, since the algorithm is fit based on just a single penalty parameter.

Author(s)

Benjamin A. Logsdon

References


See Also

`vbsr`

Examples

```r
n <- 100;
m <- 500;
ntrue <- 10;
e <- rnorm(n);
X <- matrix(rnorm(n*m),n,m);
tbeta <- sample(1:m,ntrue);
beta <- rep(0,m);
beta[tbeta]<- rnorm(ntrue,0,.3);
y <- X%*%beta;
y <- y+e;
res<- vbsr(y,X,family="normal",l0_path=seq(-15,-3,length.out=100),post=0.5);
klRes <- compute_KL(res$z,0.01,1);
```

Description

Fit a linear model via a fast coordinate variational Bayes algorithm. Applicable to linear and logistic regression, and solves the problem on either a path of the spike (l0) parameter or at a fixed value based on the data-dimensions.
Usage

vbsr(y, 
  X, 
  ordering_mat=NULL, 
  eps=1e-6, 
  exclude=NULL, 
  add.intercept=TRUE, 
  maxit = 1e4, 
  n_orderings = 10, 
  family = "normal", 
  scaling = TRUE, 
  return_kl = TRUE, 
  estimation_type = "BMA", 
  bma_approximation = TRUE, 
  screen = 1.0, 
  post=0.95, 
  already_screened = 1.0, 
  kl = 0.99, 
  l0_path=NULL, 
  cleanSolution=FALSE)

Arguments

y response variable. Normally distributed errors for family="normal". For family="binomial" should be coded as a vector of 0's and 1's.

X Design matrix, an n x m matrix, with rows as observations

ordering_mat Optionally specified coordinate update ordering matrix. Must be in matrix form with columns as permutation vectors of length m, and there must be n_orderings columns.

eps Tolerance used to determine convergence of the algorithm based on the lower bound.

exclude An optional indicator vector of length m of 0's and 1's indicating whether to penalize a particular variable or not (0=penalize, 1=unpenalized)

add.intercept A boolean variable indicating whether or not to include an unpenalized intercept variable.

maxit The maximum number of iterations to run the algorithm for a given solution to a penalized regression problem.

n_orderings The number of random starts used.

family The type of error model used. Currently supported modes are family="normal" and family="binomial"

scaling A boolean variable indicating whether or not to scale the columns of X to have mean zero and variance one.

return_kl A boolean variable indicating whether or not to return an analysis of the null distributed features in the data-set as a function of the penalty parameter.
estimation_type
The type of estimation to perform based on the number of unique solution identified to the penalized regression problem. Valid values are estimation_type="BMA" and estimation_type="MAXIMAL".

bma_approximation
A boolean variable indicating whether to compute a full correction to the z statistic. WARNING can make the algorithm very computationally intensive for highly multi-modal posterior surfaces.

screen
P-value to do marginal screening. Default is to not do marginal prescreening (e.g marginal p-value of 1.0)

post
Choice of penalty parameter such that a feature will have a posterior probability of 0.95 if it passes a Bonferroni correction in the multivariate model. Default is post=.95. More conservative approach would be post=0.5

already_screened
If features are already screened, the marginal p-value used for screening.

kl
The inner percentiles of the distribution to compute the Kullback-Leibler overfitting statistic. Only works for analysis when directly specifying a path of penalization parameter (e.g. 10_path). For default kl=0.99 the KL-statistic is used for the statistics between the 1%-99% of the distribution.

10_path
The path of penalty parameters to solve the spike regression problem. If post is specified, this is computed automatically.

cleanSolution
This parameter determines whether a given solution is further filtered using an unpenalized model. If cleanSolution=TRUE, then the features that are significant after a Bonferroni correction given the p-values from the vbsr regression model are then tested in an unpenalized linear regression model. The p-values and z-statistics are updated using the Wald test from the unpenalized linear regression model for the features that were selected.

Details
The solutions to the spike penalized regression model are fit with a coordinate variational Bayes algorithm based on the 10_path values of the spike hyper-parameter.

Value
A list with all the results of the vbsr analysis.

beta
The expected value of the penalized regression coefficients.

alpha
The estimated value of the unpenalized regression coefficients.

z
The Z-statistic for each penalized regression coefficient

pval
The p-values based on the asymptotic normal assumption of the Z-statistics

post
The posterior probabilities of each of the regression coefficients

l0
The penalty parameters used to solve the penalized regression problem

modelEntropy
The entropy of the identified approximate posterior probability distribution over model space.
modelProb  The approximate posterior probability distribution over the identified model space.

kl_index  If a path solution was run with the KL diagnostic statistic then the points in the path where the KL statistic is nearest the min, the mean, the min + 1 s.e., and the mean +1 s.e.

kl  The KL statistic computed across the path

kl_min  The minimum KL statistic identified along the path

kl_mean  The expected KL statistic given the number of features identified

Author(s)
Benjamin A. Logsdon

References


See Also
compute_KL

Examples
n <- 100;
m <- 500;
ntrue <- 10;
e <- rnorm(n);
X <- matrix(rnorm(n*m),n,m);
tbeta <- sample(1:m,ntrue);
beta <- rep(0,m);
beta[tbeta]<- rnorm(ntrue,0,.3);
y <- X%*%beta;
y <- y+e;

res<- vbsr(y,X,family="normal");
Index

*Topic model fit diagnostic
  compute_KL, 2
*Topic models
  vbsr, 3
*Topic regression
  vbsr, 3

compute_KL, 2

vbsr, 3