Package 'EBglmnet'

January 30, 2016

Type Package

Title Empirical Bayesian Lasso and Elastic Net Methods for Generalized Linear Models

Version 4.1

Date 2016-01-15

Author Anhui Huang, Dianting Liu

Maintainer Anhui Huang <a.huang1@umiami.edu>

Suggests knitr, glmnet

Description

Provides empirical Bayesian lasso and elastic net algorithms for variable selection and effect estimation. Key features include sparse variable selection and effect estimation via generalized linear regression models, high dimensionality with p>>n, and significance test for nonzero effects. This package outperforms other popular methods such as lasso and elastic net methods in terms of power of detection, false discovery rate, and power of detecting grouping effects.

License GPL

VignetteBuilder knitr

URL https://sites.google.com/site/anhuihng/

NeedsCompilation yes

Repository CRAN

Depends R (>= 2.10)

Date/Publication 2016-01-30 00:36:25

R topics documented:

EBglmnet-package *Empirical Bayesian Lasso (EBlasso) and Elastic Net (EBEN) Methods for Generalized Linear Models*

Description

Fast Empirical Bayesian Lasso (EBlasso) and Elastic Net (EBEN) are generalized linear regression methods for variable selections and effect estimations. Similar as `lasso` and `elastic net` implemented in the package **glmnet**, **EBglmnet** features the capabilities of handling $p >> n$ data, where `p` is the number of variables and `n` is the number of samples in the regression model, and inferring a sparse solution such that irrelevant variables will have exactly zero value on their regression coefficients. Additionally, there are several unique features in **EBglmnet**:

1) Both `EBlasso` and `EBEN` can select more than `n` nonzero effects.
2) EBglmnet also performs hypothesis testing for the significance of nonzero estimates.
3) EBglmnet includes built-in functions for epistasis analysis.

There are three sets of hierarchical prior distributions implemented in **EBglmnet**:

1) EBlasso-NE is a two-level prior with (normal + exponential) distributions for the regression coefficients.
2) EBlasso-NEG is a three-level hierarchical prior with (normal + exponential + gamma) distributions.
3) EBEN implements a normal and generalized gamma hierarchical prior.

While those sets of priors are all "peak zero and flat tails", `EBlasso-NE` assigns more probability mass to the tails, resulting in more nonzero estimates having large p-values. In contrast, `EBlasso-NEG` has a third level constraint on the `lasso` prior, which results in higher probability mass around zero, thus more sparse results in the final outcome. Meanwhile, `EBEN` encourages a grouping effect such that highly correlated variables can be selected as a group. Similar as the relationship between `elastic net` and `lasso`, there are two parameters (α, λ) required for `EBEN`, and it is reduced to `EBlasso-NE` when parameter $\alpha = 1$. We recommend using EBlasso-NEG when there are a large number of candidate effects, using EBlasso-NE when effect sizes are relatively small, and using EBEN when groups of highly correlated variables such as co-regulated gene expressions are of interest.

Two models are available for both methods: linear regression model and logistic regression model. Other features in this package includes:
* 1 * epistasis (two-way interactions) can be included for all models/priors;
* 2 * model implemented with memory efficient `C` code;
* 3 * LAPACK/BLAS are used for most linear algebra computations.

Several simulation and real data analysis in the reference papers demonstrated that **EBglmnet** enjoys better performance than `lasso` and `elastic net` methods in terms of power of detection, false

discover rate, as well as encouraging grouping effect when applicable.

Key Algorithms are described in the following paper:
1. EBlasso-NEG: (Cai X., Huang A., and Xu S., 2011), (Huang A., Xu S., and Cai X., 2013)
2. EBlasso-NE: (Huang A., Xu S., and Cai X., 2013)
3. group EBlasso: (Huang A., Martin E., et al. 2014)
4. EBEN: (Huang A., Xu S., and Cai X., 2015)
5. Whole-genome QTL mapping: (Huang A., Xu S., and Cai X., 2014)

Details

Package:	EBglmnet
Type:	Package
Version:	4.1
Date:	2016-01-15
License:	gpl

Author(s)

Anhui Huang, Dianting Liu
Maintainer: Anhui Huang <a.huang1@umiami.edu>

References

Huang, A., Xu, S., and Cai, X. (2015). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity 114(1): 107-115.

Huang, A., E. Martin, et al. (2014). "Detecting genetic interactions in pathway-based genome-wide association studies." Genet Epidemiol 38(4): 300-309.

Huang, A., S. Xu, et al. (2014). "Whole-genome quantitative trait locus mapping reveals major role of epistasis on yield of rice." PLoS ONE 9(1): e87330.

Huang, A. (2014). "Sparse model learning for inferring genotype and phenotype associations." Ph.D Dissertation. University of Miami(1186).

Huang A, Xu S, Cai X. (2013). Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. BMC genetics 14(1):5.

Cai, X., Huang, A., and Xu, S. (2011). Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. BMC Bioinformatics 12, 211.

BASIS *An Example Data*

Description

This is a 1000x481 sample feature matrix

Usage

```
data(BASIS)
```

Format

The format is: int [1:1000, 1:481] 0 -1 0 0 1 0 1 0 1 0 ...

Details

The data was simulated on a 2400 centimorgan (cM) chromosome of an F2 population from a cross of two inbred lines. The three genotype of AA, Aa and aa were coded as 1, 0, -1, respectively. Each column corresponds to an even spaced d= 5cM genetic marker, and each row represents a sample. The Haldane map function was assumed in the simulation, such that correlation between markers having distance d is $R = exp(-2d)$. Example of using this dataset for multiple QTL mapping is available in the EBglmnet Vignette.

Source

Huang, A., Xu, S., and Cai, X. (2014). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity 10.1038/hdy.2014.79

Examples

```
data(BASIS)
```

cv.EBglmnet *Cross Validation (CV) Function to Determine Hyperparameters of the EBglmnet Algorithms*

Description

The degree of shrinkage, or equivalently, the number of non-zero effects selected by EBglmnet are controlled by the hyperparameters in the prior distribution, which can be obtained via Cross Validation (CV). This function performs k-fold CV for hyperparameter selection, and outputs the model fit results using the optimal parameters. Therefore, this function runs `EBglmnet` for (`k x n_parameters + 1`) times. By default, EBlasso-NE tests 20 λs , EBEN tests an additional 10 αs (thus a total of 200 pair of hyperparameters), and EBlasso-NEG tests up to 25 pairs of (a,b).

Usage

```
cv.EBglmnet(x, y, family=c("gaussian","binomial"),
prior= c("lassoNEG","lasso","elastic net"), nfolds=5,
foldId, Epis = FALSE, group = FALSE, verbose = 0)
```

Arguments

x	input matrix of dimension n x p; each row is an observation vector, and each column is a candidate variable. When epistasis is considered, users do not need to create a giant matrix including both main and interaction terms. Instead, x should always be the matrix corresponding to the p main effects, and cv.EBglmnet will generate the interaction terms dynamically during running time.
y	response variable. Continuous for family="gaussian", and binary for family="binomial". For binary response variable, y can be a Boolean or numeric vector, or factor type array.
family	model type taking values of "gaussian" (default) or "binomial".
prior	prior distribution to be used. Taking values of "lassoNEG"(default), "lasso", and "elastic net". All priors will produce a sparse outcome of the regression coefficients; see Details for choosing priors.
nfolds	number of n-fold CV. nfolds typically >=3. Although nfolds can be as large as the sample size (leave-one-out CV), it will be computationally intensive for large datasets. Default value is nfolds=5.
foldId	an optional vector of values between 1 and nfolds identifying which fold each observation is assigned to. If not supplied, each of the n samples will be assigned to the nfolds randomly.
Epis	Boolean parameter for including two-way interactions. By default, Epis = FALSE. When Epis = TRUE, EBglmnet will take all pair-wise interaction effects into consideration. EBglmnet does not create a giant matrix for all the p(p+1)/2 effects. Instead, it dynamically allocates the memory for the nonzero effects identified in the model, and reads the corresponding variables from the original input matrix x.
group	Boolean parameter for group EBlasso (currently only available for the "lassoNEG" prior). This parameter is only valid when Epis = TRUE, and is set to FALSE by default. When Epis = TRUE and group = TRUE, the hyperparameter controlling degree of shrinkage will be further scaled such that the scale hyperparameter for interaction terms is different with that of main effects by a factor of $\sqrt{p(p-1)/2}$. When p is large, eg., several thousands of genetic markers, the total number of effects can easily be more than 10 millions, and group EBlasso helps to reduce the interference of spurious correlation and noise accumulation.
verbose	parameter that controls the level of message output from EBglment. It takes values from 0 to 5; larger verbose displays more messages. 0 is recommended for CV to avoid excessive outputs. Default value for verbose is minimum message output.

Details

The three priors in EBglmnet all contain hyperparameters that control how heavy the tail probabilities are. Different values of the hyperparameters will yield different number of non-zero effects retained in the model. Appropriate selection of their values is required to obtain optimal results, and CV is the most oftenly used method. For Gaussian model, CV determines the optimal hyperparameter values that yield the minimum square error. In Binomial model, CV calculates the mean logLikelihood in each of the left out fold, and chooses the values that yield the maximum mean logLikelihood value of the k-folds. See `EBglmnet` for the details of hyperparameters in each prior distribution.

Value

`CrossValidation`
: matrix of CV result with columns of:
column 1: hyperparameter1
column 2: hyperparameter2
column 3: prediction metrics/Criteria
column 4: standard error in the k-fold CV.

Prediction metrics is the mean square error (MSE) for Gaussian model and mean log likelihood (logL) for the binomial model.

`optimal hyperparameter`
: the hyperparameters that yield the smallest MSE or the largest logL.

`fit`
: model fit using the optimal parameters computed by CV. See `EBglmnet` for values in this item.

`WaldScore`
: the Wald Score for the posterior distribution. See (Huang A., Martin E., et al., 2014b) for using Wald Score to identify significant effect set.

`Intercept`
: model intercept. This parameter is not shrunk (assumes uniform prior).

`residual variance`
: the residual variance if the Gaussian family is assumed in the GLM

`logLikelihood`
: the log Likelihood if the Binomial family is assumed in the GLM

`hyperparameters`
: the hyperparameter(s) used to fit the model

`family`
: the GLM family specified in this function call

`prior`
: the prior used in this function call

`call`
: the call that produced this object

`nobs`
: number of observations

`nfolds`
: number of folds in CV

Author(s)

Anhui Huang and Dianting Liu
Dept of Electrical and Computer Engineering, Univ of Miami, Coral Gables, FL

References

Cai, X., Huang, A., and Xu, S. (2011). Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. BMC Bioinformatics 12, 211.

Huang A, Xu S, Cai X. (2013). Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. BMC genetics 14(1):5.

Huang, A., Xu, S., and Cai, X. (2014a). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity 10.1038/hdy.2014.79

uang, A., E. Martin, et al. (2014b). Detecting genetic interactions in pathway-based genome-wide association studies. Genet Epidemiol 38(4): 300-309.

Examples

```
rm(list = ls())
library(EBglmnet)
#Use R built-in data set state.x77
y= state.x77[,"Life Exp"]
xNames = c("Population","Income","Illiteracy", "Murder","HS Grad","Frost","Area")
x = state.x77[,xNames]
#
#Gaussian Model
#lassoNEG prior as default
out = cv.EBglmnet(x,y)
out$fit
#lasso prior
out = cv.EBglmnet(x,y,prior= "lasso")
out$fit
#elastic net prior
out = cv.EBglmnet(x,y,prior= "elastic net")
out$fit
#
#Binomial Model
#create a binary response variable
yy = y>mean(y);
out = cv.EBglmnet(x,yy,family="binomial")
out$fit
#with epistatic effects
out = cv.EBglmnet(x,yy,family="binomial",prior= "elastic net",Epis =TRUE)
out$fit
```

`EBglmnet` *Main Function for the EBglmnet Algorithms*

Description

EBglmnet is the main function to fit a generalized linear model via the empirical Bayesian methods with lasso and elastic net hierarchical priors. It features with p>>n capability, produces a sparse outcome for the regression coefficients, and performs significance test for nonzero effects in both linear and logistic regression models.

Usage

```
EBglmnet(x, y, family=c("gaussian","binomial"),prior= c("lassoNEG","lasso","elastic net"),
hyperparameters,Epis = FALSE,group = FALSE, verbose = 0)
```

Arguments

x	input matrix of dimension n x p; each row is an observation vector, and each column is a variable. When epistasis is considered, users do not need to create a giant matrix including both main and interaction terms. Instead, x should always be the matrix corresponding to the p main effects, and EBglmnet will generate the interaction terms dynamically during running time.
y	response variable. Continuous for family="gaussian", and binary for family="binomial". For binary response variable, y can be a Boolean or numeric vector, or factor type array.
family	model type taking values of "gaussian" (default) or "binomial".
prior	prior distribution to be used. It takes values of "lassoNEG"(default), "lasso", and "elastic net". All priors will produce a sparse outcome of the regression coefficients; see Details for choosing priors.
hyperparameters	the optimal hyperparameters in the prior distribution. Similar as λ in lasso method, the hyperparameters control the number of nonzero elements in the regression coefficients. Hyperparameters are most oftenly determined by CV. See cv.EBglmnet for the method in determining their values. While cv.EBglmnet already provides the model fitting results using the hyperparameters determined in CV, users can use this function to obtain the results under other parameter selection criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC).
Epis	Boolean parameter for including two-way interactions. By default, Epis = FALSE. When Epis = TRUE, EBglmnet will take all pair-wise interaction effects into consideration. EBglmnet does not create a giant matrix for all the p(p+1)/2 effects. Instead, it dynamically allocates memory for the nonzero effects identified in the model, and reads the corresponding variables from the original input matrix x
group	Boolean parameter for group EBlasso (currently only available for the "lassoNEG" prior). This parameter is only valid when Epis = TRUE, and is set to FALSE by default. When Epis = TRUE and group = TRUE, the hyperparameter controlling degree of shrinkage will be further scaled such that the scale hyperparameter for interaction terms is different with that of main effects by a factor of $\sqrt{p(p-1)/2}$. When p is large, eg., several thousands of genetic markers, the

total number of effects can easily be more than 10 millions, and `group EBlasso` helps to reduce the interference of spurious correlation and noise accumulation.

`verbose` parameter that controls the level of message output from EBglment. It takes values from 0 to 5; larger verbose displays more messages. small values are recommended to avoid excessive outputs. Default value for `verbose` is minimum message output.

Details

EBglmnet implements three set of hierarchical prior distributions for the regression parameters β:

lasso prior:

$$\beta_j \sim N(0, \sigma_j^2),$$
$$\sigma_j^2 \sim exp(\lambda), j = 1, \ldots, p.$$

lasso-NEG prior:

$$\beta_j \sim N(0, \sigma_j^2),$$
$$\sigma_j^2 \sim exp(\lambda),$$
$$\lambda \sim gamma(a, b), j = 1, \ldots, p.$$

elastic net prior:

$$\beta_j \sim N[0, (\lambda_1 + \tilde{\sigma_j}^{-2})^{-2}],$$
$$\tilde{\sigma_j}^2 \sim generalized - gamma(\lambda_1, \lambda_2), j = 1, \ldots, p.$$

The prior distributions are peak zero and flat tail probability distributions that assign a high prior probability mass to zero and still allow heavy probability on the two tails, which reflect the prior belief that a sparse solution exists: most of the variables will have no effects on the response variable, and only some of the variables will have non-zero effects in contributing the outcome in `y`.

The three priors all contains hyperparameters that control how heavy the tail probability is, and different values of them will yield different number of non-zero effects retained in the model. Appropriate selection of their values is required for obtaining optimal results, and CV is the most oftenly used method. See `cv.EBglmnet` for details for determining the optimal hyperparameters in each priors under different GLM families.

lassoNEG prior
`"lassoNEG"` prior has two hyperparameters (a,b), with $a \geq -1$ and `b>0`. Although `a` is allowed to be greater than -1.5, it is not encouraged to choose values in (-1.5, -1) unless the signal-to-noise ratio in the explanatory variables are very small.

lasso prior
`"lasso"` prior has one hyperparameter λ, with $\lambda \geq 0$. λ is similar as the shrinkage parameter in `lasso` except that even for `p>>n`, λ is allowed to be zero, and `EBlasso` can still provide a sparse solution thanks to the implicit constraint that $\sigma^2 \geq 0$.

elastic net prior
Similar as the elastic net in package **glmnet**, EBglmnet transforms the two hyperparameters λ_1 and λ_2 in the `"elastic net"` prior in terms of other two parameters $\alpha(0 \leq \alpha \leq 1)$ and $\lambda(\lambda > 0)$. Therefore, users are asked to specify `hyperparameters=c`(α, λ).

Value

`fit` the model fit using the hyperparameters provided. EBglmnet selects the variables having nonzero regression coefficients and estimates their posterior distributions. With the posterior mean and variance, a `t-test` is performed and the `p-value` is calculated. Result in fit is a matrix with rows corresponding to the variables having nonzero effects, and columns having the following values:

column1-2: (locus1, locus2) denoting the column number in the input matrix `x`. When locus1 equals to locus2, this effect is from one of the `p` main effects, otherwise, it is the interaction effect between `x[,locus1]` and `x[,locus2]`. When `Epis = FALSE`, which is the default setting, locus1 always equals locus2. If `Epis =TRUE`, fit always puts the main effects in the beginning, and epistatic effects after that.

column3: beta. It is the posterior mean of the nonzero regression coefficients.

column4: posterior variance. It is the diagonal element of the posterior covariance matrix among the nonzero regression coefficients.

column5: t-value calculated using column 3-4.

column6: p-value from t-test.

`WaldScore` the Wald Score for the posterior distribution. It is computed as $\beta^T \Sigma^{-1} \beta$. See (Huang A, 2014b) for using Wald Score to identify significant effect set.

`Intercept` the intercept in the linear regression model. This parameter is not shrunk.

`residual variance` the residual variance if the Gaussian family is assumed in the GLM

`logLikelihood` the log Likelihood if the Binomial family is assumed in the GLM

`hyperparameters` the hyperparameter used to fit the model

`family` the GLM family specified in this function call

`prior` the prior used in this function call

`call` the call that produced this object

`nobs` number of observations

Author(s)

Anhui Huang and Dianting Liu
Dept of Electrical and Computer Engineering, Univ of Miami, Coral Gables, FL

References

Cai, X., Huang, A., and Xu, S. (2011). Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. BMC Bioinformatics 12, 211.

Huang A, Xu S, Cai X. (2013). Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. BMC genetics 14(1):5.

Huang, A., Xu, S., and Cai, X. (2014a). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity 10.1038/hdy.2014.79

Examples

```
rm(list = ls())
library(EBglmnet)
#Use R built-in data set state.x77
y= state.x77[,"Life Exp"]
xNames = c("Population","Income","Illiteracy", "Murder","HS Grad","Frost","Area")
x = state.x77[,xNames]
#
#Gaussian Model
#lassoNEG prior as default
out = EBglmnet(x,y,hyperparameters=c(0.5,0.5))
out$fit
#lasso prior
out = EBglmnet(x,y,prior= "lasso",hyperparameters=0.5)
out$fit
#elastic net prior
out = EBglmnet(x,y,prior= "elastic net",hyperparameters=c(0.5,0.5))
out$fit
#residual variance
out$res
#intercept
out$Intercept
#
#Binomial Model
#create a binary response variable
yy = y>mean(y);
out = EBglmnet(x,yy,family="binomial",hyperparameters=c(0.5,0.5))
out$fit
#with epistatic effects
out = EBglmnet(x,yy,family="binomial",hyperparameters=c(0.5,0.5),Epis =TRUE)
out$fit
```

Index