Package 'BVSSemi'

October 20, 2023

Type Package

Title BVSSemi: A Bayesian variable selection for semicontinuous response

Version 1.0

Date 2023-06-20

Author Thierry Chekouo

Maintainer Thierry Chekouo <tchekouo@umn.edu>

Description The BVSSemi package implements an MCMC algorithm for a Bayesian variable selection for semicontinuous response. The models use a two-part model approach and propose to link the models via MRF priors.

License GPL (>= 2.0)

Imports truncnorm, gear, invgamma

Encoding UTF-8

R topics documented:

`GenDataSemiContinous` *Generation of simulated data as explained in the reference manuscript.*

Description

This function generates data described in the manuscript.

Usage

```
GenDataSemiContinous(Xreal=FALSE,n=n,p=p,X=NULL,sd=1,impf=20,beta=1,
                                percentOverlap="Full",seed=1)
```

Arguments

`Xreal`	If TRUE, then provide the matrix of covariates X to simulate the semicontinuous response. If FALSE, then X is simulated as well.
`n`	Number of individuals
`p`	Number of features in X
`X`	matrix of the set of features
`sd`	Standard deviation of the error of the continuous response
`impf`	Number of known important features
`beta`	Beta regression effect
`percentOverlap`	Data type. If percentOverlap="Full", then the set of important features is the same between the continous and binary models. If percentOverlap="Medium", then 50 percent of important features are the same between the two models. If percentOverlap="NoOverlap", then there is no commeon important features between the two models.
`seed`	Seed to generate random numbers

Details

The function will generate data as explained in the manuscript. To see the results, use the "$" operator.

Value

`Y`	A semicontinuous response variable of dimension n
`X`	A matrix of features of dimension $n \times p$
`Z.cont`	A binary vector that indicates whether a feature is important in the continous model (vector of dimension p)
`Z.bin`	A binary vector that indicates whether a feature is important in the binary model (vector of dimension p)

References

Samuel Babatunde, Tolulope Sajobi and Thierry Chekouo (2023), *A Bayesian Variable Selection for Semicontinuous Response data: Application to cardiovascular disease*, *submitted*.

See Also

`GenDataSemiContinous`

Examples

```
library(BVSSemi);
Dat=GenDataSemiContinous(n=500,p=500,sd=1,impf=20,beta=0.3,percentOverlap="Full",seed=1)
str(Dat)
```

MainBVSSemi — *An MCMC algorithm to perform a Bayesian variable selection for semicontinuous response via a two-part model: continous model (i.e. linear model with continuous response) and binary model (Probit model with binary response).*

Description

The algorithm is implemented on the three methods: i) BVSSemiMRF: this method encourages the common selection of important features between the two models (continuous and binary model). ii) BVSSemiComb: this method selects the same set of features associated for both the continuous and binary models such that each selected feature is related to our semicontinuous response (both zeros and positive values) and iii) BVSSemiIndep: it assumes that the set of selected features for the log-linear model is not necessarily the same with the logistic model, and the two models are fitted independently. The algorithm computes mainly the marginal posterior probabilities of inclusion of each feature for each model.

Usage

```
MainBVSSemi(Method="BVSSemiMRF",Y=Y,X=X,Xcov=NULL,seed=1,atheta=1,btheta=1,tau2cont=1,
                tau2bin=0.5,nu1cont=-3,nu2bin=-3,varpropTheta=.25,Bigtau2=100,
                     mcmcsample=10000,burnin=5000)
```

Arguments

Method	It's one of the three methods: "BVSSemiMRF","BVSSemiComb" or "BVSSemi-Indep". The default value is "BVSSemiMRF".
Y	A semicontinuous response.
X	A matrix of features of dimension $n \times p$
Xcov	A set of covariates that are not subject of varible selection (e.g. clinical/demographic variables such as sex, age, race, etc..)
seed	Set a seed number to generate distributions in the MCMC algorithm.
atheta	It's the shape (hyper)parameter of the gamma prior distribution of theta from the BVSSemiMRF method. The parameter theta measures the borrows strength between the two model and encourages the common selection between the two models.
btheta	It's the scale (hyper)parameter of the gamma prior distribution of theta from the BVSSemiMRF method.
tau2cont	It's the variance (hyper)parameter of the normal prior distribution of regression effects on the continuous model
tau2bin	It's the variance (hyper)parameter of the normal prior distribution of regression effects on the binary model
nu1cont	log-odds of prior prob. of feature inclusion in the continous model
nu2bin	log-odds of prior prob. of feature inclusion in the binary model
varpropTheta	Variance of the proposal distribution (in the Metro. Hasting step) of theta in our MCMC algorithm. It should be chosen to have an acceptance rate of theta around 20-60 percent. The default value is .25.

Bigtau2	It's the variance (hyper)parameter of the normal prior distribution of regression effects of features that are not subject of variable selection (e.g., intercept and other clinical covariates)
mcmcsample	Total number of MCMC draws. It must be larger than burnin.
burnin	Number of draws to discard for burn-in.

Details

The function will return several R objects, which can be assigned to a variable. To see the results, use the "$" operator.

Value

prob.Z.Cont	Marginal posterior probabilities of each feauture in the continuous model.
prob.Z.Bin	Marginal posterior probabilities of each feauture in the binary model.
thetasample	Posterior samples of theta from the BVSSemiMRF method.
AcceptanceRateTheta	MCMC Acceptance rate of theta from the BVSSemiMRF.
sigma2Sample	Posterior samples of the variance sigma2 from the continous model.

References

Samuel Babatunde, Tolulope Sajobi and Thierry Chekouo (2023), *A Bayesian Variable Selection for Semicontinuous Response data: Application to cardiovascular disease*, *submitted*.

See Also

GenDataSemiContinous

Examples

```
library(BVSSemi);
Dat=GenDataSemiContinous(n=500,p=200,sd=1,impf=20,beta=0.3,percentOverlap="Full",seed=1)
str(Dat)
result<-MainBVSSemi(Method="BVSSemiMRF",Y=Dat$Y,X=Dat$X,Xcov=NULL,seed=1,atheta=1,btheta=1,tau2cont=1,
tau2bin=1,nu1cont=-4,nu2bin=-4,varpropTheta=.25,Bigtau2=100,mcmcsample=50000,burnin=10000);
str(result)
library(pROC)
AUC1=as.numeric(auc(roc(as.factor(Dat$Z.cont),result$prob.Z.Cont)))
AUC2=as.numeric(auc(roc(as.factor(Dat$Z.bin),result$prob.Z.Bin)))
print(AUC1);print(AUC2)
```

Index