Package 'bgsmtr'

January 7, 2019

Title Bayesian Group Sparse Multi-Task Regression

Version 0.5

Description Fits a Bayesian group-sparse multi-task regression model using Gibbs sampling. The hierarchical prior encourages shrinkage of the estimated regression coefficients at both the gene and SNP level. The model has been extended to a spatial model that allows for two type correlation in neuroimaging genetics data and been applied successfully to imaging phenotypes of dimension up to 100; it can be used more generally for multivariate (non-imaging) phenotypes.

Author Yin Song, Shufei Ge, Liangliang Wang, Farouk S. Nathoo, Keelin Greenlaw, Mary Lesperance

Maintainer Yin Song <yinsong@uvic.ca>

Depends R (>= 3.3.0), Matrix (>= 1.2.6), mvtnorm (>= 1.0.5),matrixcalc(>= 1.0.3), miscTools (>= 0.6.22)

Imports coda (>= 0.18.1), EDISON (>= 1.1.1),statmod (>= 1.4.26),methods (>= 3.3.3), sparseMVN (>= 0.2.0),inline (>= 0.3.15),LaplacesDemon (>= 16.1.0), CholWishart (>= 0.9.3),mnormt (>= 1.5.4),Rcpp(>= 0.12.14),

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

NeedsCompilation no

biocViews

Repository CRAN

Date/Publication 2019-01-07 00:00:09 UTC

R topics documented:

bgsmtr *Bayesian Group Sparse Multi-Task Regression for Imaging Genetics*

Description

Runs the the Gibbs sampling algorithm to fit a Bayesian group sparse multi-task regression model. Tuning parameters can be chosen using either the MCMC samples and the WAIC (multiple runs) or using an approximation to the posterior mode and five-fold cross-validation (single run).

Usage

```
bgsmtr(X, Y, group, tuning = "CV.mode", lam_1_fixed = NULL,
  lam_2_fixed = NULL, iter_num = 10000, burn_in = 5001)
```

Arguments

X	A d-by-n matrix; d is the number of SNPs and n is the number of subjects. Each row of X should correspond to a particular SNP and each column should correspond to a particular subject. Each element of X should give the number of minor alleles for the corresponding SNP and subject. The function will center each row of X to have mean zero prior to running the Gibbs sampling algorithm.
Y	A c-by-n matrix; c is the number of phenotypes (brain imaging measures) and n is the number of subjects. Each row of Y should correspond to a particular phenotype and each column should correspond to a particular subject. Each element of Y should give the measured value for the corresponding phentoype and subject. The function will center and scale each row of Y to have mean zero and unit variance prior to running the Gibbs sampling algorithm.
group	A vector of length d; d is the number of SNPs. Each element of this vector is a string representing a gene or group label associated with each SNP. The SNPs represented by this vector should be ordered according to the rows of X.
tuning	A string, either 'WAIC' or 'CV.mode'. If 'WAIC', the Gibbs sampler is run with fixed values of the tuning parameters specified by the arguments *lam_1_fixed* and *lam_2_fixed* and the WAIC is computed based on the sampling output. This can then be used to choose optimal values for *lam_1_fixed* and *lam_2_fixed* based on multiple runs with each run using different values of *lam_1_fixed* and *lam_2_fixed*. This option is best suited for either comparing a small set of tuning parameter values or for computation on a high performance computing cluster where different nodes can be used to run the function with different values of *lam_1_fixed* and *lam_2_fixed*. Posterior inference is then based on the run that produces the lowest value for the WAIC. The option 'CV.mode', which is the default, is best suited for computation using just a single processor. In this case the tuning parameters are chosen based on five-fold cross-validation over a grid of possible values with out-of-sample prediction based on an approximate posterior mode. The Gibbs sampler is then run using the chosen values of the tuning parameters. When tuning = 'CV.mode' the values for the arguments *lam_1_fixed* and *lam_2_fixed* are not required.

`lam_1_fixed` Only required if tuning = 'WAIC'. A positive number giving the value for the gene-specific tuning parameter. Larger values lead to a larger degree of shrinkage to zero of estimated regression coefficients at the gene level (across all SNPs and phenotypes).

`lam_2_fixed` Only required if tuning = 'WAIC'. A positive number giving the value for the SNP-specific tuning parameter. Larger values lead to a larger degree of shrinkage to zero of estimated regression coefficients at the SNP level (across all phenotypes).

`iter_num` Positive integer representing the total number of iterations to run the Gibbs sampler. Defaults to 10,000.

`burn_in` Nonnegative integer representing the number of MCMC samples to discard as burn-in. Defaults to 5001.

Value

A list with the elements

`WAIC` If tuning = 'WAIC' this is the value of the WAIC computed from the MCMC output. If tuning = 'CV.mode' this component is excluded.

`Gibbs_setup` A list providing values for the input parameters of the function.

`Gibbs_W_summaries`
A list with five components, each component being a d-by-c matrix giving some posterior summary of the regression parameter matrix W, where the ij-th element of W represents the association between the i-th SNP and j-th phenotype.

-Gibbs_W_summaries$W_post_mean is a d-by-c matrix giving the posterior mean of W.

-Gibbs_W_summaries$W_post_mode is a d-by-c matrix giving the posterior mode of W.

-Gibbs_W_summaries$W_post_sd is a d-by-c matrix giving the posterior standard deviation for each element of W.

-Gibbs_W_summaries$W_2.5_quantile is a d-by-c matrix giving the posterior 2.5 percent quantile for each element of W.

-Gibbs_W_summaries$W_97.5_quantile is a d-by-c matrix giving the posterior 97.5 percent quantile for each element of W.'

Author(s)

Farouk S. Nathoo, <nathoo@uvic.ca>

Keelin Greenlaw <keelingreenlaw@gmail.com>

Mary Lesperance <mlespera@uvic.ca>

References

Greenlaw, Keelin, Elena Szefer, Jinko Graham, Mary Lesperance, and Farouk S. Nathoo. "A Bayesian Group Sparse Multi-Task Regression Model for Imaging Genetics." arXiv preprint arXiv:1605.02234 (2016).

Nathoo, Farouk S., Keelin Greenlaw, and Mary Lesperance. "Regularization Parameter Selection for a Bayesian Multi-Level Group Lasso Regression Model with Application to Imaging Genomics." arXiv preprint arXiv:1603.08163 (2016).

Examples

```
data(bgsmtr_example_data)
names(bgsmtr_example_data)


## Not run:
## test run the sampler for 100 iterations with fixed tunning parameters and compute WAIC
## we recomend at least 5,000 iterations for actual use
fit = bgsmtr(X = bgsmtr_example_data$SNP_data, Y = bgsmtr_example_data$BrainMeasures,
group = bgsmtr_example_data$SNP_groups, tuning = 'WAIC', lam_1_fixed = 2, lam_2_fixed = 2,
iter_num = 100, burn_in = 50)
## posterior mean for regression parameter relating 100th SNP to 14th phenotype
fit$Gibbs_W_summaries$W_post_mean[100,14]
## posterior mode for regression parameter relating 100th SNP to 14th phenotype
fit$Gibbs_W_summaries$W_post_mode[100,14]
## posterior standard deviation for regression parameter relating 100th SNP to 14th phenotype
fit$Gibbs_W_summaries$W_post_sd[100,14]
## 95% equal-tail credible interval for regression parameter relating 100th SNP to 14th phenotype
c(fit$Gibbs_W_summaries$W_2.5_quantile[100,14],fit$Gibbs_W_summaries$W_97.5_quantile[100,14])

## End(Not run)

## Not run:
## run the sampler for 10,000 iterations with tuning parameters set using cross-validation
## On a standard computer with a small numer of cores this is the recomended option
fit = bgsmtr(X = bgsmtr_example_data$SNP_data, Y = bgsmtr_example_data$BrainMeasures,
group = bgsmtr_example_data$SNP_groups, tuning = 'CV.mode',iter_num = 10000, burn_in = 5000)

## End(Not run)
```

---

bgsmtr_example_data *Example Structural Neuroimaging and Genetic Data*

Description

Simulated data with 632 subjects, 486 SNPs from 33 genes, 15 structural neuroimaging measures.

Usage

```
data(bgsmtr_example_data)
```

Format

A list with three components: "SNP_data", "SNP_groups", "BrainMeasures". SNP_data is a 486-by-632 matrix containing minor allele counts for 632 subjects and 486 SNPs. SNP_groups is a vector of length 486 with labels partitioning the 486 SNPs into 33 genes. BrainMeasures is a 15-by-632 matrix containing simulated volumetric and cortical thickness measures for 15 regions of interest.

Examples

```
data(bgsmtr_example_data)
names(bgsmtr_example_data)
dim(bgsmtr_example_data$SNP_data)
dim(bgsmtr_example_data$BrainMeasures)
unique(bgsmtr_example_data$SNP_groups)
```

---

sp_bgsmtr *Spatial Bayesian Group Sparse Multi-Task Regression for Imaging Genetics*

Description

Bayesian Group Sparse Multi-Task Regression that allows for two types of correlation typically seen in structural brain imaging data. First, the spatial correlation in the imaging phenotypes obtained from neighbouring regions of the brain. Second, the correlation between corresponding measures on opposite hemispheres.

Usage

```
sp_bgsmtr(X, Y, method = "MCMC", rho = NULL, lambdasq = NULL,
  alpha = NULL, A = NULL, c.star = NULL, FDR_opt = TRUE,
  WAIC_opt = TRUE, iter_num = 10000, burn_in = 5001)
```

Arguments

X	A d-by-n matrix; d is the number of SNPs and n is the number of subjects. Each row of X should correspond to a particular SNP and each column should correspond to a particular subject. Each element of X should give the number of minor alleles for the corresponding SNP and subject. The function will center each row of X to have mean zero prior to running the Gibbs sampling algorithm.
Y	A c-by-n matrix; c is the number of phenotypes (brain imaging measures) and n is the number of subjects. Each row of Y should correspond to a particular phenotype and each column should correspond to a particular subject. Each element of Y should give the measured value for the corresponding phentoype and subject. The function will center and scale each row of Y to have mean zero and unit variance prior to running the Gibbs sampling algorithm.
method	A string, either 'MCMC' or 'MFVB'. If 'MCMC', the Gibbs sampling method will be used. If 'MFVB', mean field variational bayes method will be used.

rho	spatial cohesion paramter. If no value has been assigned to it, it takes 0.8 by default.
lambdasq	A tuning paratmeter. If no value has beeen assigned to it, it takes 1000 by default.
alpha	Bayesian False Discovery Rate (FDR) level. Default level is 0.05.
A	A c/2 by c/2 neighborhood structure matrix for different brain regions.
c.star	The threshold for computing posterior tail probabilities p_ij for Bayesian FDR as defined in Section 3.2 of Song et al. (2018). If not specified the default is to set this threshold as the minimum posterior standard deviation, where the minimum is taken over all regression coefficients in the model.
FDR_opt	A logical operator for computing Bayesian FDR. By default, it's TRUE.
WAIC_opt	A logical operator for computing WAIC from MCMC method. By default, it's TRUE.
iter_num	Positive integer representing the total number of iterations to run the Gibbs sampler. Defaults to 10,000.
burn_in	Nonnegative integer representing the number of MCMC samples to discard as burn-in. Defaults to 5001.

Value

A list with the elements

WAIC	WAIC is computed from the MCMC output if "MCMC" is chosen for method.
lower_boud	Lower bound from MFVB output if "MFVB is choosen for method.
Gibbs_setup	A list providing values for the input parameters of the function.
Gibbs_W_summaries	A list with five components, each component being a d-by-c matrix giving some posterior summary of the regression parameter matrix W, where the ij-th element of W represents the association between the i-th SNP and j-th phenotype. -Gibbs_W_summaries$W_post_mean is a d-by-c matrix giving the posterior mean of W. -Gibbs_W_summaries$W_post_mode is a d-by-c matrix giving the posterior mode of W. -Gibbs_W_summaries$W_post_sd is a d-by-c matrix giving the posterior standard deviation for each element of W. -Gibbs_W_summaries$W_2.5_quantile is a d-by-c matrix giving the posterior 2.5 percent quantile for each element of W. -Gibbs_W_summaries$W_97.5_quantile is a d-by-c matrix giving the posterior 97.5 percent quantile for each element of W.'
FDR_summaries	A list with three components providing the summaries for estimated Bayesian FDR results for both MCMC and MFVB methods. Details for Bayesian FDR computation could be found at Morris et al.(2008). -sensitivity_rate is the estimated sensitivity rate for each region. -specificity_rate is the estimated specificity rate for each region. -significant_snp_idx is the index of estimated significant/important SNPs for each region.

MFVB_summaries A list with four components, each component is the mean field variational bayes approximation summary of model paramters.
-Number of Iteration is how many iterations it takes for convergence.
-W_post_mean is MFVB approximation of W.
-Sigma_post_mean is MFVB approximation of Sigma.
-omega_post_mean is MFVB approximation of Omega.

Author(s)

Yin Song, <yinsong@uvic.ca>

Shufei Ge <shufeig@sfu.ca>

Farouk S. Nathoo, <nathoo@uvic.ca>

Liangliang Wang <lwa68@sfu.ca>

Examples

```
data(sp_bgsmtr_example_data)
names(sp_bgsmtr_example_data)


## Not run:

# Run the example data with Gibbs sampling and compute Bayesian FDR as follow:

fit_mcmc = sp_bgsmtr(X = sp_bgsmtr_example_data$SNP_data,
Y = sp_bgsmtr_example_data$BrainMeasures, method = "MCMC",
A = bgsmtr_example_data$neighborhood_structure, rho = 0.8,
FDR_opt = TRUE, WAIC_opt = TRUE,lambdasq = 1000, iter_num = 10000.)

# MCMC estimation results for regression parameter W and estimated Bayesian FDR summaries

fit_mcmc$Gibbs_W_summaries
fit_mcmc$FDR_summaries

# The WAIC could be also obtained as:

fit_mcmc$WAIC

# Run the example data with mean field variational Bayes and compute Bayesian FDR as follow:

fit_mfvb = sp_bgsmtr(X = sp_bgsmtr_example_data$SNP_data,
Y = sp_bgsmtr_example_data$BrainMeasures, method = "MFVB",
A = bgsmtr_example_data$neighborhood_structure, rho = 0.8,FDR_opt = TRUE,
lambdasq = 1000, iter_num = 10000.)

# MFVB estimated results for regression parameter W and estimated Bayesian FDR summaries
fit_mfvb$MFVB_summaries
fit_mfvb$FDR_summaries

# The corresponding lower bound of MFVB method after convergence is obtained as:
```

```
fit_mfvb$lower_boud


## End(Not run)
```

`sp_bgsmtr_example_data` *Example Structural Neuroimaging and Genetic Data for Spatial Model.*

Description

Simulated data with 632 subjects, 486 SNPs from 24 structural neuroimaging measures.

Usage

```
data(sp_bgsmtr_example_data)
```

Format

A list with three components: "SNP_data", "SNP_groups", "BrainMeasures". SNP_data is a 486-by-632 matrix containing minor allele counts for 632 subjects and 486 SNPs. neighbourhood_structure is a 12 by 12 first order neighbourhood structure matrix. BrainMeasures is a 24-by-632 matrix containing simulated volumetric and cortical thickness measures for 24 regions of interest.

Examples

```
data(sp_bgsmtr_example_data)
names(sp_bgsmtr_example_data)
dim(sp_bgsmtr_example_data$SNP_data)
dim(sp_bgsmtr_example_data$BrainMeasures)
dim(sp_bgsmtr_example_data$neighbourhood_structure)
```

Index