Package 'BLR'

October 12, 2022

Version 1.6

Date 2020-01-04

Title Bayesian Linear Regression

Author Gustavo de los Campos, Paulino Perez Rodriguez,

Maintainer Paulino Perez Rodriguez <perpdgo@colpos.mx>

Depends R (>= 3.1.2)

Description Bayesian Linear Regression.

LazyLoad true

License GPL-2

NeedsCompilation yes

Repository CRAN

Date/Publication 2020-01-07 20:40:02 UTC

R topics documented:

A *Pedigree info for the wheat dataset*

Description

Is a numerator relationship matrix (599 x 599) computed from a pedigree that traced back many generations. This relationship matrix was derived using the Browse application of the International Crop Information System (ICIS), as described in http://repository.cimmyt.org/xmlui/bitstream/handle/10883/3488/72673.pdf (McLaren *et al.* 2005).

Source

International Maize and Wheat Improvement Center (CIMMYT), Mexico.

References

McLaren, C. G., R. Bruskiewich, A.M. Portugal, and A.B. Cosico. 2005. The International Rice Information System. A platform for meta-analysis of rice crop data. *Plant Physiology* **139**: 637-642.

BLR *Bayesian Linear Regression*

Description

The BLR ('Bayesian Linear Regression') function was designed to fit parametric regression models using different types of shrinkage methods. An earlier version of this program was presented in de los Campos *et al.* (2009).

Usage

```
BLR(y, XF, XR, XL, GF, prior, nIter, burnIn, thin,thin2,saveAt,
    minAbsBeta,weights)
```

Arguments

y	(numeric, n) the data-vector (NAs allowed).
XF	(numeric, $n \times pF$) incidence matrix for $\boldsymbol{\beta}_F$, may be NULL.
XR	(numeric, $n \times pR$) incidence matrix for $\boldsymbol{\beta}_R$, may be NULL.
XL	(numeric, $n \times pL$) incidence matrix for $\boldsymbol{\beta}_L$, may be NULL.
GF	(list) providing an \$ID (integer, n) linking observations to groups (e.g., lines or sires) and a (co)variance structure (\$A, numeric, $pU \times pU$) between effects of the grouping factor (e.g., line or sire effects). Note: ID must be an integer taking values from 1 to pU; ID[i]=q indicates that the ith observation in $\boldsymbol{y}$ belongs to cluster q whose (co)variance function is in the qth row (column) of $\boldsymbol{A}$. GF may be NULL.
weights	(numeric, n) a vector of weights, may be NULL.
nIter,burnIn, thin	(integer) the number of iterations, burn-in and thinning.
saveAt	(string) this may include a path and a pre-fix that will be added to the name of the files that are saved as the program runs.
prior	(list) containing the following elements,

- prior\$varE, prior\$varBR, prior\$varU: (list) each providing degree of freedom (\$df) and scale (\$S). These are the parameters of the scaled inverse-χ^2 distributions assigned to variance components, see Eq. (2) below. In the parameterization used by BLR() the prior expectation of variance parameters is $S/(df-2)$.

- prior$lambda: (list) providing $value (initial value for λ); $type ('random' or 'fixed') this argument specifies whether λ should be kept fixed at the value provided by $value or updated with samples from the posterior distribution; and, either $shape and $rate (this when a Gamma prior is desired on λ^2) or $shape1, $shape2 and $max, in this case $p(\lambda|\max,\alpha_1,\alpha_2) \propto Beta\left(\frac{\lambda}{\max}|\alpha_1,\alpha_2\right)$. For detailed description of these priors see de los Campos *et al.* (2009).

`thin2` This value controls wether the running means are saved to disk or not. If thin2 is greater than nIter the running means are not saved (default, thin2=1×10^{10}).

`minAbsBeta` The minimum absolute value of the components of $\boldsymbol{\beta}_L$ to avoid numeric problems when sampling from $\boldsymbol{\tau}^2$, default 1×10^{-9}

Details

The program runs a Gibbs sampler for the Bayesian regression model described below.

Likelihood. The equation for the data is:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}_F\boldsymbol{\beta}_F + \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{X}_L\boldsymbol{\beta}_L + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{y}$, the response is a $n \times 1$ vector (NAs allowed); μ is an intercept; $\boldsymbol{X}_F, \boldsymbol{X}_R, \boldsymbol{X}_L$ and $\boldsymbol{Z}$ are incidence matrices used to accommodate different types of effects (see below), and; $\boldsymbol{\varepsilon}$ is a vector of model residuals assumed to be distributed as $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, Diag(\sigma^2_{\boldsymbol{\varepsilon}}/w_i^2))$, here $\sigma^2_{\boldsymbol{\varepsilon}}$ is an (unknown) variance parameter and w_i are (known) weights that allow for heterogeneous-residual variances.

Any of the elements in the right-hand side of the linear predictor, except μ and $\boldsymbol{\varepsilon}$, can be omitted; by default the program runs an intercept model.

Prior. The residual variance is assigned a scaled inverse-χ^2 prior with degree of freedom and scale parameter provided by the user, that is, $\sigma^2_{\boldsymbol{\varepsilon}} \sim \chi^{-2}(\sigma^2_{\boldsymbol{\varepsilon}}|df_{\boldsymbol{\varepsilon}}, S_{\boldsymbol{\varepsilon}})$. The regression coefficients $\{\mu, \boldsymbol{\beta}_F, \boldsymbol{\beta}_R, \boldsymbol{\beta}_L, \boldsymbol{u}\}$ are assigned priors that yield different type of shrinkage. The intercept and the vector of regression coefficients $\boldsymbol{\beta}_F$ are assigned flat priors (i.e., estimates are not shrunk). The vector of regression coefficients $\boldsymbol{\beta}_R$ is assigned a Gaussian prior with variance common to all effects, that is, $\beta_{R,j} \overset{iid}{\sim} N(0, \sigma^2_{\boldsymbol{\beta}_R})$. This prior is the Bayesian counterpart of Ridge Regression. The variance parameter $\sigma^2_{\boldsymbol{\beta}_R}$, is treated as unknown and it is assigned a scaled inverse-χ^2 prior, that is, $\sigma^2_{\boldsymbol{\beta}_R} \sim \chi^{-2}(\sigma^2_{\boldsymbol{\beta}_R}|df_{\boldsymbol{\beta}_R}, S_{\boldsymbol{\beta}_R})$ with degrees of freedom $df_{\boldsymbol{\beta}_R}$, and scale $S_{\boldsymbol{\beta}_R}$ provided by the user.
The vector of regression coefficients $\boldsymbol{\beta}_L$ is treated as in the Bayesian LASSO of Park and Casella (2008). Specifically,

$$p(\boldsymbol{\beta}_L, \boldsymbol{\tau}^2, \lambda|\sigma^2_{\boldsymbol{\varepsilon}}) = \left\{\prod_k N(\beta_{L,k}|0, \sigma^2_{\boldsymbol{\varepsilon}}\tau^2_k) Exp\left(\tau^2_k|\lambda^2\right)\right\} p(\lambda),$$

where, $Exp(\cdot|\cdot)$ is an exponential prior and $p(\lambda)$ can either be: (a) a mass-point at some value (i.e., fixed λ); (b) $p(\lambda^2) \sim Gamma(r, \delta)$ this is the prior suggested by Park and Casella (2008); or, (c) $p(\lambda|\max,\alpha_1,\alpha_2) \propto Beta\left(\frac{\lambda}{\max}|\alpha_1,\alpha_2\right)$, see de los Campos *et al.* (2009) for details. It can be shown that the marginal prior of regression coefficients $\beta_{L,k}$, $\int N(\beta_{L,k}|0, \sigma^2_{\boldsymbol{\varepsilon}}\tau^2_k) Exp\left(\tau^2_k|\lambda^2\right) \partial\tau^2_k$, is Double-Exponential. This prior has thicker tails and higher peak of mass at zero than the Gaussian prior used for $\boldsymbol{\beta}_R$, inducing a different type of shrinkage.

The vector $\boldsymbol{u}$ is used to model the so called 'infinitesimal effects', and is assigned a prior $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{A}\sigma^2_{\boldsymbol{u}})$, where, $\boldsymbol{A}$ is a positive-definite matrix (usually a relationship matrix computed from a pedigree) and $\sigma^2_{\boldsymbol{u}}$ is an unknow variance, whose prior is $\sigma^2_{\boldsymbol{u}} \sim \chi^{-2}(\sigma^2_{\boldsymbol{u}}|df_{\boldsymbol{u}}, S_{\boldsymbol{u}})$.

Collecting the above mentioned assumptions, the posterior distribution of model unknowns, $\boldsymbol{\theta} = \left\{\mu, \boldsymbol{\beta}_F, \boldsymbol{\beta}_R, \sigma^2_{\boldsymbol{\beta}_R}, \boldsymbol{\beta}_L, \boldsymbol{\tau}^2, \lambda, \boldsymbol{u}, \sigma^2_{\boldsymbol{u}}, \sigma^2_{\boldsymbol{\varepsilon}}, \right\}$, is,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) \quad \propto \quad & N\left(\boldsymbol{y}|\mathbf{1}\mu + \boldsymbol{X}_F\boldsymbol{\beta}_F + \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{X}_L\boldsymbol{\beta}_L + \boldsymbol{Z}\boldsymbol{u}; Diag\left\{\frac{\sigma^2_{\varepsilon}}{w^2_i}\right\}\right) \\ & \times \left\{\prod_j N\left(\beta_{R,j}|0, \sigma^2_{\boldsymbol{\beta}_R}\right)\right\} \chi^{-2}\left(\sigma^2_{\boldsymbol{\beta}_R}|df_{\boldsymbol{\beta}_R}, S_{\boldsymbol{\beta}_R}\right) \\ & \times \left\{\prod_k N\left(\beta_{L,k}|0, \sigma^2_{\boldsymbol{\varepsilon}}\tau^2_k\right) Exp\left(\tau^2_k|\lambda^2\right)\right\} p(\lambda) \\ & \times N(\boldsymbol{u}|\mathbf{0}, \boldsymbol{A}\sigma^2_{\boldsymbol{u}})\chi^{-2}(\sigma^2_{\boldsymbol{u}}|df_{\boldsymbol{u}}, S_{\boldsymbol{u}})\chi^{-2}(\sigma^2_{\boldsymbol{\varepsilon}}|df_{\boldsymbol{\varepsilon}}, S_{\boldsymbol{\varepsilon}}) \end{aligned} \tag{2}$$

Value

A list with posterior means, posterior standard deviations, and the parameters used to fit the model:

`$yHat`	the posterior mean of $\mathbf{1}\mu + \boldsymbol{X}_F\boldsymbol{\beta}_F + \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{X}_L\boldsymbol{\beta}_L + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}$.
`$SD.yHat`	the corresponding posterior standard deviation.
`$mu`	the posterior mean of the intercept.
`$varE`	the posterior mean of $\sigma^2_{\boldsymbol{\varepsilon}}$.
`$bR`	the posterior mean of $\boldsymbol{\beta}_R$.
`$SD.bR`	the corresponding posterior standard deviation.
`$varBr`	the posterior mean of $\sigma^2_{\boldsymbol{\beta}_R}$.
`$bL`	the posterior mean of $\boldsymbol{\beta}_L$.
`$SD.bL`	the corresponding posterior standard deviation.
`$tau2`	the posterior mean of $\boldsymbol{\tau}^2$.
`$lambda`	the posterior mean of λ.
`$u`	the posterior mean of $\boldsymbol{u}$.
`$SD.u`	the corresponding posterior standard deviation.
`$varU`	the posterior mean of $\sigma^2_{\boldsymbol{u}}$.
`$fit`	a list with evaluations of effective number of parameters and DIC (Spiegelhalter *et al.*, 2002).
`$whichNa`	a vector indicating which entries in $\boldsymbol{y}$ were missing.
`$prior`	a list containig the priors used during the analysis.
`$weights`	vector of weights.
`$fit`	list containing the following elements, • $logLikAtPostMean: log-likelihood evaluated at posterior mean. • $postMeanLogLik: the posterior mean of the Log-Likelihood. • $pD: estimated effective number of parameters, Spiegelhalter *et al.* (2002). • $DIC: the deviance information criterion, Spiegelhalter *et al.* (2002).

$nIter	the number of iterations made in the Gibbs sampler.
$burnIn	the nuber of iteratios used as burn-in.
$thin	the thin used.
$y	original data-vector.

The posterior means returned by BLR are calculated after burnIn is passed and at a thin as specified by the user.

Save. The routine will save samples of μ, variance components and λ and running means (rm*.dat). Running means are computed using the thinning specified by the user (see argument thin above); however these running means are saved at a thinning specified by argument thin2 (by default, thin2=1×10^{10} so that running means are computed as the sampler runs but not saved to the disc).

Author(s)

Gustavo de los Campos, Paulino Perez Rodriguez,

References

de los Campos G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel and J. Cotes. 2009. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. *Genetics* **182**: 375-385.

Park T. and G. Casella. 2008. The Bayesian LASSO. *Journal of the American Statistical Association* **103**: 681-686.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin and A. van der Linde. 2002. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, Series B (Statistical Methodology) **64** (4): 583-639.

Examples

```
#############################################################################
##Example 1:
#############################################################################
if(FALSE){
rm(list=ls())
library(BLR)
data(wheat)     #Loads the wheat dataset

y=Y[,1]
### Creates a testing set with 100 observations
whichNa<-sample(1:length(y),size=100,replace=FALSE)
yNa<-y
yNa[whichNa]<-NA

### Runs the Gibbs sampler
fm<-BLR(y=yNa,XL=X,GF=list(ID=1:nrow(A),A=A),
                           prior=list(varE=list(df=3,S=0.25),
                           varU=list(df=3,S=0.63),
                           lambda=list(shape=0.52,rate=1e-4,
                           type='random',value=30)),
```

```
                          nIter=5500,burnIn=500,thin=1,
                          saveAt="example_")

MSE.tst<-mean((fm$yHat[whichNa]-y[whichNa])^2)
MSE.tst
MSE.trn<-mean((fm$yHat[-whichNa]-y[-whichNa])^2)
MSE.trn
COR.tst<-cor(fm$yHat[whichNa],y[whichNa])
COR.tst
COR.trn<-cor(fm$yHat[-whichNa],y[-whichNa])
COR.trn

plot(fm$yHat~y,xlab="Phenotype",
     ylab="Pred. Gen. Value" ,cex=.8)
points(x=y[whichNa],y=fm$yHat[whichNa],col=2,cex=.8,pch=19)

x11()
plot(scan('example_varE.dat'),type="o",
        ylab=expression(paste(sigma[epsilon]^2)))
}
#############################################################################
#Example 2: Ten fold, Cross validation, environment 1,
#############################################################################
if(FALSE){
rm(list=ls())
library(BLR)
data(wheat)     #Loads the wheat dataset
nIter<-1500     #For real data sets more samples are needed
burnIn<-500
thin<-10
folds<-10
y<-Y[,1]
priorBL<-list(
               varE=list(df=3,S=2.5),
               varU=list(df=3,S=0.63),
               lambda = list(shape=0.52,rate=1e-5,value=20,type='random')
             )

set.seed(123)  #Set seed for the random number generator
sets<-rep(1:10,60)[-1]
sets<-sets[order(runif(nrow(A)))]
COR.CV<-rep(NA,times=(folds+1))
names(COR.CV)<-c(paste('fold=',1:folds,sep=''),'Pooled')
w<-rep(1/nrow(A),folds) ## weights for pooled correlations and MSE
yHatCV<-numeric()

for(fold in 1:folds)
{
   yNa<-y
   whichNa<-which(sets==fold)
   yNa[whichNa]<-NA
   prefix<-paste('PM_BL','_fold_',fold,'_',sep='')
   fm<-BLR(y=yNa,XL=X,GF=list(ID=(1:nrow(A)),A=A),prior=priorBL,
```

```
                nIter=nIter,burnIn=burnIn,thin=thin)
   yHatCV[whichNa]<-fm$yHat[fm$whichNa]
   w[fold]<-w[fold]*length(fm$whichNa)
   COR.CV[fold]<-cor(fm$yHat[fm$whichNa],y[whichNa])
}

COR.CV[11]<-mean(COR.CV[1:10])
COR.CV
}
#############################################################################
```

`sets` *Sets for cross validation (CV)*

Description

Is a vector (599 x 1) that assigns observations to 10 disjoint sets; the assignment was generated at random. This is used later to conduct a 10-fold CV.

Source

International Maize and Wheat Improvement Center (CIMMYT), Mexico.

`wheat` *wheat dataset*

Description

Information from a collection of 599 historical CIMMYT wheat lines. The wheat data set is from CIMMYT's Global Wheat Program. Historically, this program has conducted numerous international trials across a wide variety of wheat-producing environments. The environments represented in these trials were grouped into four basic target sets of environments comprising four main agroclimatic regions previously defined and widely used by CIMMYT's Global Wheat Breeding Program. The phenotypic trait considered here was the average grain yield (GY) of the 599 wheat lines evaluated in each of these four mega-environments.

A pedigree tracing back many generations was available, and the Browse application of the International Crop Information System (ICIS), as described in http://repository.cimmyt.org/xmlui/bitstream/handle/10883/3488/72673.pdf (McLaren *et al.* 2005), was used for deriving the relationship matrix A among the 599 lines; it accounts for selection and inbreeding.

Wheat lines were recently genotyped using 1447 Diversity Array Technology (DArT) generated by Triticarte Pty. Ltd. (Canberra, Australia). The DArT markers may take on two values, denoted by their presence or absence. Markers with a minor allele frequency lower than 0.05 were removed, and missing genotypes were imputed with samples from the marginal distribution of marker genotypes, that is, $x_{ij} = Bernoulli(\hat{p}_j)$, where $\hat{p}_j$ is the estimated allele frequency computed from the non-missing genotypes. The number of DArT MMs after edition was 1279.

Usage

```
data(wheat)
```

Format

Matrix Y contains the average grain yield, column 1: Grain yield for environment 1 and so on. The matrix A contains additive relationship computed from the pedigree and matrix X contains the markers information.

Source

International Maize and Wheat Improvement Center (CIMMYT), Mexico.

References

McLaren, C. G., R. Bruskiewich, A.M. Portugal, and A.B. Cosico. 2005. The International Rice Information System. A platform for meta-analysis of rice crop data. *Plant Physiology* **139**: 637-642.

X *Molecular markers*

Description

Is a matrix (599 x 1279) with DArT genotypes; data are from pure lines and genotypes were coded as 0/1 denoting the absence/presence of the DArT. Markers with a minor allele frequency lower than 0.05 were removed, and missing genotypes were imputed with samples from the marginal distribution of marker genotypes, that is, $x_{ij} = Bernoulli(\hat{p}_j)$, where $\hat{p}_j$ is the estimated allele frequency computed from the non-missing genotypes. The number of DArT MMs after edition was 1279.

Source

International Maize and Wheat Improvement Center (CIMMYT), Mexico.

Y *Grain yield*

Description

A matrix (599 x 4) containing the 2-yr average grain yield of each of these lines in each of the four environments (phenotypes were standardized to a unit variance within each environment).

Source

International Maize and Wheat Improvement Center (CIMMYT), Mexico.

Index