Package 'bWGR'

March 22, 2017

Type Package

Title Bagging Whole-Genome Regression

Version 1.4

Date 2017-03-21

Author Alencar Xavier, William Muir, Shizhong Xu, Katy Rainey.

Maintainer Alencar Xavier <alenxav@gmail.com>

Description Whole-genome regression methods on Bayesian framework fitted via EM or Gibbs sampling, with optional sampling techniques and kernel term.

License GPL-3

Imports Rcpp

LinkingTo Rcpp

Depends R (>= 3.2.0)

NeedsCompilation yes

Repository CRAN

Suggests BGLR,ranger,glmnet,kernlab

Date/Publication 2017-03-22 06:13:10 UTC

R topics documented:

bWGR-package 2
Dataset 2
WGR1 (MCMC) 3
WGR2 (EM) 5
WGR3 (MKR) 7

Index **10**

bWGR-package *Bagging Whole-Genome Regression*

Description

Whole-genome regression methods on Bayesian framework fitted via EM or Gibbs sampling, with optional sampling techniques and kernel term.

Details

Package:	bWGR
Type:	Package
Version:	1.4
Date:	2017-03-21
License:	GPL-3

Author(s)

Alencar Xavier, Shizhong Xu, William Muir, Katy Rainey.

Maintainer: Alencar Xavier <alenxav@gmail.com>

Dataset *Tetra-seed Pods*

Description

Two biparental crosses phenotyped for the percentage of pods containing four seeds

Usage

```
data(tpod)
```

Details

Soybean nested association panel with 2 families (fam) containing 196 indiviuals. Genotypic matrix (gen) have 376 SNP across 20 chromosome (chr). Phenotypic information (y) regards the proportion of tetra-seed pods. Data provided by Rainey Lab for Soybean Breeding and Genetics, Purdue University.

Author(s)

Alencar Xavier and Katy Rainey

WGR1 (MCMC) *Whole-genome Regression*

Description

Univariate model to find breeding values through regression with optional resampling techniques (SBMC) and polygenic term (Kernel).

Usage

```
wgr(y,X,it=1500,bi=500,th=1,bag=1,rp=FALSE,iv=FALSE,de=FALSE,
    pi=0,df=5,R2=0.5,eigK=NULL,VarK=0.95,verb=FALSE)
```

Arguments

y	Numeric vector of observations (n) describing the trait to be analyzed. NA is allowed.
X	Numeric matrix containing the genotypic data. A matrix with n rows of observations and (m) columns of molecular markers.
it	Integer. Number of iterations or samples to be generated.
bi	Integer. Burn-in, the number of iterations or samples to be discarted.
th	Integer. Thinning parameter, used to save memory by storing only one every 'th' samples.
bag	If different than one (= complete data), it indicates the proportion of data to be subsampled in each Markov chain. For datasets with moderate number of observations, values of bag from 0.30 to 0.60 may speed up computation without losses in predicion properties. This argument enable users to enhance MCMC through SBMC ($subssamplingbootstrapMarkovchain$).
rp	Logical. Use replacement for bootstrap samples when bag is different than one.
iv	Logical. Assign markers independent variance, hence T prior. If true, turns the default model BLUP into BayesA. For this model, the shape parameter is conjugated by a gamma with hyperpriors calculated based on the R2 rule.
de	Logical. Assign markers independent variance through double-exponential prior. If true, turns the default model BLUP into Bayesian LASSO. This argument overides iv.
pi	Value between 0 and 1. If greater than zero it activates variable selection, where markers have expected probability pi of having null effect (or 1-pi if pi>0.5). The model conjugates variable selection from a Beta-Binomial distribution.
df	Hyperprior degrees of freedom of variance components.
R2	Expected R2, used to calculate the prior shape as proposed by de los Campos et al. (2013).
eigK	Output of function 'eigen'. Spectral decomposition of the kernel used to compute the polygenic term.

VarK	Numeric between 0 and 1. For reduction of dimensionality. Indicates the proportion of variance explained by Eigenpairs used to fit the polygenic term.
verb	Logical. If verbose is TRUE, function displays MCMC progress bar.

Details

The model for the whole-genome regression is as follows:

$$y = mu + Xg + u + e$$

where y is the response variable, mu is the intercept, X is the genotypic matrix, g is the regression coefficient as the product of bxd, b is the effect of an allele substitution, d is an indicator variable that define whether or not the marker should be included into the model, u is the polygenic term and e is the residual term.

Users can obtain four WGR methods out of this function: BRR (pi=0,iv=F), BayesA (pi=0,iv=T), BayesB (pi=0.01,iv=T), BayesC (pi=0.01,iv=F) and BayesL (pi=0,de=T). The full theoretical basis of each model is described by de los Campos et al. (2013).

Gibbs sampler that updates regression coefficients is adapted from GSRU algorithm (Legarra and Misztal 2008). The variable selection works through the unconditional prior algorithm proposed by Kuo and Mallick (1998). The polygenic term is solved by Bayesian algorithm of reproducing kernel Hilbert Spaces proposed by de los Campos et al. (2010).

Value

The function wgr returns a list with expected value from the marker effect (b), probability of marker being in the model (d), regression coefficient (g), variance of each marker (Vb), the intercept (mu), the polygene (u) and polygenic variance (Vk), residual variance (Ve) and the fitted value (hat).

Author(s)

Alencar Xavier

References

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2), 327-345.

de los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., & Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research, 92(04), 295-308.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. Sankhya: The Indian Journal of Statistics, Series B, 65-81.

Legarra, A., & Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. Journal of dairy science, 91(1), 360-366.

Examples

```
# Load data
data(tpod)

# BLUP
BRR = wgr(y,gen,iv=FALSE,pi=0,it=200,bi=50)
cor(y,BRR$hat)

# BayesA
BayesA = wgr(y,gen,iv=TRUE,pi=0,it=200,bi=50)
cor(y,BayesA$hat)

# BayesB
BayesB = wgr(y,gen,iv=TRUE,pi=.01,it=200,bi=50)
cor(y,BayesB$hat)

# BayesC
BayesC = wgr(y,gen,iv=FALSE,pi=.01,it=200,bi=50)
cor(y,BayesC$hat)

# BayesL
BayesL = wgr(y,gen,de=TRUE,it=200,bi=50)
cor(y,BayesL$hat)

# Bagging BLUP
Bag = wgr(y,gen,bag=0.5,it=200,bi=50)
cor(y,Bag$hat)
```

`WGR2 (EM)` *Expectation-Maximization WGR*

Description

Univariate models to find breeding values through regression fitted via expectation-maximization implemented in C++.

Usage

```
emRR(y, gen, df = 4, R2 = 0.5)
emBA(y, gen, df = 4, R2 = 0.5)
emBB(y, gen, df = 4, R2 = 0.5, Pi = 0.7)
emBC(y, gen, df = 4, R2 = 0.5, Pi = 0.7)
emBL(y, gen, R2 = 0.5, alpha = 0.02)
emEN(y, gen, R2 = 0.5, alpha = 0.02)
emDE(y, gen, R2 = 0.5)
```

Arguments

y	Numeric vector of observations (n) describing the trait to be analyzed. NA is not allowed.
gen	Numeric matrix containing the genotypic data. A matrix with n rows of observations and (m) columns of molecular markers.
df	Hyperprior degrees of freedom of variance components.
R2	Expected R2, used to calculate the prior shape as proposed by de los Campos et al. (2013).
Pi	Value between 0 and 1. Expected probability pi of having null effect (or 1-Pi if Pi>0.5).
alpha	Value between 0 and 1. Intensity of L1 variable selection.

Details

These regressions are still under development. The model for the whole-genome regression is as follows:

$$y = mu + Xb + e$$

where y is the response variable, mu is the intercept, X is the genotypic matrix, b is the effect of an allele substitution (or regression coefficient) and e is the residual term.

Value

The EM functions returns a list with the intercept (mu), the regression coefficient (b), the fitted value (hat), and the estimated intraclass-correlation ($h2$).

Author(s)

Alencar Xavier

Examples

```
data(tpod)

# BLUP
BRR = emRR(y,gen)
cor(y,BRR$hat)

# BayesA
BayesA = emBA(y,gen)
cor(y,BayesA$hat)

# BayesB
BayesB = emBB(y,gen)
cor(y,BayesB$hat)
```

```
# BayesC
BayesC = emBC(y,gen)
cor(y,BayesC$hat)

# BayesL 1
BayesL1 = emBL(y,gen)
cor(y,BayesL1$hat)

# BayesL 2
BayesL2 = emDE(y,gen)
cor(y,BayesL2$hat)

# Elastic-Net
ElasticNet = emEN(y,gen)
cor(y,ElasticNet$hat)
```

`WGR3 (MKR)` *Multivariate Kernel Regression*

Description

Multivariate model to find breeding values through kernel regression.

Usage

```
mkr(Y,K=NULL,eK=NULL,it=500,bu=200,th=3,
    df=5,R2=0.5,EigT=0.05,verb=FALSE)
```

Arguments

`Y`	Numeric matrix of observations (n, k) describing the trait to be analyzed. `NA` is allowed.
`K`	Numeric matrix containing the genotypic relationship matrix. A matrix with n rows and columns.
`eK`	Output of eigen. Eigendecomposition of K. If eK is specified, the is no need for the argument K.
`it`	Integer. Number of iterations or samples to be generated.
`bu`	Integer. Burn-in, the number of iterations or samples to be discarted.
`th`	Integer. Thinning parameter, used to save memory by storing only one every 'th' samples.
`df`	Prior degrees of freedom for covariance components.
`R2`	Expected R2, used to calculate the prior shape as proposed by de los Campos et al. (2013).
`EigT`	Null or numeric. If provided, the model uses just Eigenpairs with Eigenvalues above the specified theshold.
`verb`	Logical. If verbose is TRUE, function displays MCMC progress bar.

Details

The model for the kernel regression is as follows:

$Y = mu + Z(UB) + E$,

where Y is a matrix of response variables, mu represents the intercepts, Z is the design matrix, U is the matrix of Eigenvector of K, b is a vector of regression coefficients and E is the residual matrix. Variance components are sampled from a inverse Wishart distribution (Sorensen and Gianola 2002). Regression coefficients are solved with an adaptation of the algorithm proposed by de los Campos et al. (2010).

Value

The function mkr returns a list with the random effect covariance matrix (VA), residual covariance matrix (VE) and a matrix with breeding values (BV).

Author(s)

Alencar Xavier

References

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2), 327-345.

de los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., & Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research, 92(04), 295-308.

Sorensen D., and Gianola D. (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer.

Examples

```
# G matrix
data(tpod)
G = tcrossprod(gen)
G = G/mean(diag(G))

# Phenotypes
Y1 = rnorm(196,y,.1)
Y2 = rnorm(196,y,.2)
Y3 = rnorm(196,y,.3)
Phe = cbind(Y1,Y2,Y3)

# Fit model
test = mkr(Phe,G)

# Genetic correlation
cov2cor(test$VA)

# Environmental correlation
```

```
cov2cor(test$VE)

# Heritabilies
diag(test$VA/(test$VA+test$VE))

# Goodness of fit
diag(cor(Phe,test$BV))
```

Index

.Random.seed *(Dataset)*, 2

bWGR *(bWGR-package)*, 2
bWGR-package, 2

chr *(Dataset)*, 2

Dataset, 2

emBA *(WGR2 (EM))*, 5
emBB *(WGR2 (EM))*, 5
emBC *(WGR2 (EM))*, 5
emBL *(WGR2 (EM))*, 5
emDE *(WGR2 (EM))*, 5
emEN *(WGR2 (EM))*, 5
emRR *(WGR2 (EM))*, 5

fam *(Dataset)*, 2

gen *(Dataset)*, 2
GSEN *(WGR1 (MCMC))*, 3

Hmat *(WGR3 (MKR))*, 7

KMUP *(WGR1 (MCMC))*, 3
KMUP2 *(WGR1 (MCMC))*, 3

mkr *(WGR3 (MKR))*, 7

tpod *(Dataset)*, 2

wgr *(WGR1 (MCMC))*, 3
WGR1 (MCMC), 3
WGR2 (EM), 5
WGR3 (MKR), 7

y *(Dataset)*, 2