Package ‘bWGR’

February 14, 2022

Type Package
Title Bayesian Whole-Genome Regression
Version 2.0
Date 2022-02-13
Author Alencar Xavier, William Muir, David Habier, Kyle Kocak, Shizhong Xu, Katy Rainey.
Maintainer Alencar Xavier <alenxav@gmail.com>
Description Whole-genome regression methods on Bayesian framework fitted via EM
    or Gibbs sampling, univariate and multivariate, with optional kernel term and
    sampling techniques.
License GPL-3
Imports Matrix, Rcpp, RcppEigen
LinkingTo Rcpp, RcppEigen
Depends R (>= 3.6.0)
NeedsCompilation yes
Repository CRAN
Date/Publication 2022-02-14 15:40:11 UTC

R topics documented:

  bWGR-package .......................................................... 2
  Dataset ................................................................. 2
  WGR1 (MC) ............................................................. 3
  WGR2 (EM) ............................................................. 6
  WGR3 (MV) ............................................................. 8
  XTRA 1 ................................................................. 9
  XTRA 2 ................................................................. 11

Index 12
bWGR-package  

*Bayesian Whole-Genome Regression*

**Description**

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

**Details**

<table>
<thead>
<tr>
<th>Package:</th>
<th>bWGR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>2.0</td>
</tr>
<tr>
<td>Date:</td>
<td>2022-02-13</td>
</tr>
<tr>
<td>License:</td>
<td>GPL-3</td>
</tr>
</tbody>
</table>

**Author(s)**

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

Maintainer: Alencar Xavier <alenxav@gmail.com>

**Examples**

```r
data(tpod)
Fit = wgr(y, gen)
cor(y, Fit$hat)
```

**Dataset**

*Tetra-seed Pods*

**Description**

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

**Usage**

```r
data(tpod)
```
**Details**

Soybean nested association panel with 2 families (fam) containing 196 individuals. 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 (MC) MCMC Whole-genome Regression**

**Description**

Univariate model to find breeding values through regression with optional resampling techniques (Xavier et al. 2017) and polygenic term (Kernel). See “Details” for additional standalone functions written in C++.

**Usage**

```r
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 discarded.

- **th**
  Integer. Thinning parameter, used to save memory by storing only one every 'th' samples.

- **bag**
  If different than one, 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 prediction properties. This argument enable users to enhance MCMC through subsampling (Xavier et al. 2017).

- **rp**
  Logical. Use replacement for bootstrap samples when bag is different than one.

- **iv**
  Logical. Assign markers independent variance, a T prior from a mixture of normals. If true, turns the default model BLUP into BayesA.

- **de**
  Logical. Assign markers independent variance through double-exponential prior. If true, turns the default model BLUP into Bayesian LASSO. This argument overrides 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.

df  Prior degrees of freedom of variance components.

R2  Expected R2, used to calculate the prior shape.

eigK  Output of function eigen. Spectral decomposition of the kernel used as a second random effect (e.g. pedigree matrix).

VarK  Numeric between 0 and 1. For reduction of dimensionality. Indicates the proportion of variance explained by Eigenpairs used to fit second random effect.

verb  Logical. If verbose is TRUE, function displays MCMC progress bar.

Details

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

\[ y = \mu + Xb + u + e \]

where \( y \) is the response variable, \( \mu \) is the intercept, \( X \) is the genotypic matrix, \( b \) is the regression coefficient or effect of an allele substitution, with \( d \) probability of being included into the model, \( u \) is the polygenic term if a kernel is used, and \( e \) is the residual term.

Users can obtain four WGR methods out of this function: BR (pi=0, iv=F), BayesA (pi=0, iv=T), BayesB (pi=0.95, iv=T), BayesC (pi=0.95, iv=F) and Bayesian LASSO or BayesL (pi=0, de=T). 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 of functions wgr, BayesB and BayesC works through the unconditional prior algorithm proposed by Kuo and Mallick (1998), whereas BayesCpi and BayesDpi are based on Metropolis-Hastings. Prior shape estimates are computed as \( S_B = R^2 * df * \text{var}(y) / \text{MS}_x \) and \( S_e = (1-R^2) * df * \text{var}(y) \), with an exception for BayesC and BayesCpi where the prior shape is \( S_B = R^2 * df * \text{var}(y) / \text{MS}_x (1-pi) \). The polygenic term is solved by Bayesian algorithm of reproducing kernel Hilbert Spaces proposed by de los Campos et al. (2010).

In addition to wgr, standalone C++ functions available include:

01) BayesA(y,X, it=1500, bi=500, df=5, R2=0.5)
02) BayesB(y,X, it=1500, bi=500, pi=0.95, df=5, R2=0.5)
03) BayesC(y,X, it=1500, bi=500, pi=0.95, df=5, R2=0.5)
04) BayesCpi(y,X, it=1500, bi=500, df=5, R2=0.5)
05) BayesDpi(y,X, it=1500, bi=500, df=5, R2=0.5)
06) BayesL(y,X, it=1500, bi=500, df=5, R2=0.5)
07) BayesRR(y,X, it=1500, bi=500, df=5, R2=0.5)

The implementations that support two random effects include:

08) BayesA2(y,X1,X2, it=1500, bi=500, df=5, R2=0.5)
09) BayesB2(y,X1,X2, it=1500, bi=500, pi=0.95, df=5, R2=0.5)
10) BayesRR2(y,X1,X2, it=1500, bi=500, df=5, R2=0.5)

And the cross-validation for the C++ implementations, with arguments analogous to emCV.

mcmcCV(y, gen, k=5, n=5, it=1500, bi=500, pi=0.95, df=5, R2=0.5, avg=T, llo=NULL, tbv=NULL, ReturnGebv=FALSE)
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 ($V_b$), the intercept ($\mu$), the polygene ($u$) and polygenic variance ($V_k$), residual variance ($V_e$) and the fitted value ($\hat{y}$).

Author(s)

Alencar Xavier

References


Examples

```r
## Not run:

# Load data
data(tpod)

# BLUP
fit_BRR = wgr(y,gen,iv=FALSE,pi=0)
cor(y,fit_BRR$hat)

# BayesA
fit_BayesA = wgr(y,gen,iv=TRUE,pi=0)
cor(y,fit_BayesA$hat)

# BayesB
fit_BayesB = wgr(y,gen,iv=TRUE,pi=.95)
cor(y,fit_BayesB$hat)

# BayesC
fit_BayesC = wgr(y,gen,iv=FALSE,pi=.95)
cor(y,fit_BayesC$hat)

# BayesCpi
fit_BayesCpi = BayesCpi(y,gen)
```
cor(y,fit_BayesCpi$hat)

# BayesDpi
fit_BayesDpi = BayesDpi(y,gen)
cor(y,fit_BayesDpi$hat)

# BayesL
fit_BayesL = wgr(y,gen,de=TRUE)
cor(y,fit_BayesL$hat)

# Bagging BLUP
fit_Bag = wgr(y,gen,bag=0.5)
cor(y,fit_Bag$hat)

## End(Not run)

---

**WGR2 (EM)**

*Expectation-Maximization WGR*

---

**Description**

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

**Usage**

```r
emRR(y, gen, df = 10, R2 = 0.5)
emBA(y, gen, df = 10, R2 = 0.5)
emBB(y, gen, df = 10, R2 = 0.5, Pi = 0.75)
emBC(y, gen, df = 10, R2 = 0.5, Pi = 0.75)
emBL(y, gen, R2 = 0.5, alpha = 0.02)
emEN(y, gen, R2 = 0.5, alpha = 0.02)
emDE(y, gen, R2 = 0.5)
emML(y, gen, D = NULL)
emCV(y, gen, k = 5, n = 5, Pi = 0.75, alpha = 0.02,
     df = 10, R2 = 0.5, avg=TRUE, llo=NULL, tbv=NULL, ReturnGebv = FALSE)
```  

**Arguments**

- **y**
  - Numeric vector of response variable (n). 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 (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).
\textbf{alpha} \quad \text{Value between 0 and 1. Intensity of L1 variable selection.} \\
\textbf{D} \quad \text{NULL or numeric vector with length p. Vector of weights for markers.} \\
\textbf{k} \quad \text{Integer. Folding of a k-fold cross-validation.} \\
\textbf{n} \quad \text{Integer. Number of cross-validation to perform.} \\
\textbf{avg} \quad \text{Logical. Return average across CV, or correlations within CV.} \\
\textbf{llo} \quad \text{NULL or a vector (numeric or factor) with the same length as y. If provided, the cross-validations are performed as Leave a Level Out (LLO). This argument allows the user to redefine the splits. This argument overrides k and n.} \\
\textbf{tbv} \quad \text{NULL or numeric vector of 'true breeding values' (n) to use to compare cross-validations to. If NULL, the cross-validations will have the phenotypes as prediction target.} \\
\textbf{ReturnGebv} \quad \text{Logical. If TRUE, it returns a list with the average marker values and fitted values across all cross-validations, in addition to the regular output.} \\

\textbf{Details} 

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. A k-fold cross-validation for model evaluation is provided by \textit{emCV}. 

\textbf{Value} 

The EM functions returns a list with the intercept (\( \mu \)), the regression coefficient (\( b \)), the fitted value (\( \hat{y} \)), and the estimated intraclass-correlation (\( h^2 \)). 

The function \textit{emCV} returns the predictive ability of each model, that is, the correlation between the predicted and observed values from \( k \)-fold cross-validations repeated \( n \) times.

\textbf{Author(s)} 

Alencar Xavier

\textbf{Examples} 

```r
## Not run:

data(tpod)
emCV(y, gen, 3, 3)

## End(Not run)
```
Description

Multivariate model to find breeding values.

Usage

\[
mrr(Y, X) \quad \text{mkr}(Y, K)
\]

Arguments

\[
Y \quad \text{Numeric matrix of observations x trait. NA is allowed.}
K \quad \text{Numeric matrix containing the relationship matrix.}
X \quad \text{Numeric matrix containing the genotyping matrix.}
\]

Details

The model for the ridge regression (mrr) is as follows:

\[
Y = Mu + XB + E
\]

where \(Y\) is a matrix of response variables, \(Mu\) represents the intercepts, \(X\) is the matrix of genotypic information, \(B\) is the matrix of marker effects, and \(E\) is the residual matrix.

The model for the kernel regression (mkr) is as follows:

\[
Y = Mu + UB + E
\]

where \(Y\) is a matrix of response variables, \(Mu\) represents the intercepts, \(U\) is the matrix of Eigenvector of \(K\), \(b\) is a vector of regression coefficients and \(E\) is the residual matrix.

Algorithm: Residuals are assumed to be independent among traits. Regression coefficients are solved via a multivariate adaptation of Gauss-Seidel Residual Update. Since version 2.0, the solver of mrr is based on the Randomized Gauss-Seidel algorithm. Variance and covariance components are solved with an EM-REML like approach proposed by Schaeffer called Pseudo-Expectation.

Other related implementations:

01) mkr2X(Y,K1,K2): Solves multi-trait kernel regressions with two random effects.
02) mrr2X(Y,X1,X2): Solves multi-trait ridge regressions with two random effects.
03) MRR3(Y,X,...): Similar to mrr.

Value

Returns a list with the random effect covariances (\(V_b\)), residual variances (\(V_e\)), genetic correlations (\(G\)), matrix with marker effects (\(b\)) or eigenvector effects (if mkr), intercepts (\(Mu\)), heritabilities (\(h^2\)), and a matrix with fitted values (\(hat\)).
Author(s)
Alencar Xavier, David Habier

Examples

# Load genomic data
data(tpod)
X = CNT(gen)

# Simulate phenotyp
sim = SimY(X)
Y = sim$Y
TBV = sim$tbv

# Fit regression model
test = mrr(Y,X)

# Genetic correlation
test$GC

# Heritabilies
test$h2

# Accuracy
diag(cor(TBV,test$hat))

---

**XTRA 1  Mixed model solver**

**Description**

Function to solve univariate mixed models with or without the usage of omic information. This function allows single-step modeling of replicated observations with marker information available through the usage of a linkage function to connect to a whole-genome regression method. Genomic estimated values can be optionally deregressed (no shrinkage) while fitting the model.

**Usage**

mixed(y,random=NULL,fixed=NULL,data=NULL,X=list(),
alg=emML,maxit=10,Deregress=FALSE,...)

**Arguments**

- **y** Response variable from the data frame containg the dataset.
- **random** Formula. Right-hand side formula of random effects.
- **fixed** Formula. Right-hand side formula of fixed effects.
**data**
Data frame containing the response variable, random and fixed terms.

**X**
List of omic incidence matrix. Row names of these matrices connect the omic information to the levels of the indicated random terms (e.g. \(X=\text{list("ID"} = \text{gen})\)).

**alg**
Function. Whole-genome regression algorithm utilized to solve link functions. These include MCMC (wgr, BayesB, etc) and EM (emEN, emDE, etc) algorithms. By default, it runs maximum likelihood emML.

**maxit**
Integer. Maximum number of iterations.

**Deregress**
Logical. Deregress (unshrink) coefficients while fitting the model?

... Additional arguments to be passed to the whole-genome regression algorithms specified on alg.

**Details**

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

\[ y = Xb + Zu + Wa + e \]

where \(y\) is the response variable, \(Xb\) corresponds to the fixed effect term, \(Zu\) corresponds to one or more random effect terms, \(W\) is the incidence matrix of terms with omic information and \(a\) is omic values by \(a = Mg\), where \(M\) is the genotypic matrix and \(g\) are marker effects. Here, \(e\) is the residual term. An example is provided using the data from the NAM package with: demo(mixedmodel).

Other related implementations (untested):
1) \(\text{mm}(y, \text{random=\text{NULL}}, \text{fixed=\text{NULL}}, \text{data=\text{NULL}}, \text{M=\text{NULL}}, \text{bin=FALSE}, \text{AM=\text{NULL}}, \text{it=10}, \text{verb=TRUE}, \text{FLM=TRUE}, \text{wgtM=TRUE}, \text{cntM=TRUE}, \text{nPc=3})\)
2) \(\text{mtmixed = function}(\text{resp, random=\text{NULL}}, \text{fixed=\text{NULL}}, \text{data, X=\text{list}()}, \text{maxit=10}, \text{init=10}, \text{regVC=FALSE})\)

**Value**

The function wgr returns a list with Fitness values (Fitness) containing observation obs, fitted values hat, residuals res, and fitted values by model term fits; Estimated variance components (VarComp) containing the variance components per se (VarComponents) and variance explained by each model term (VarExplained), regression coefficients by model term (Coefficients), and the effects of structured terms (Structure) containing the marker effects of each model term where markers were provided.

**Author(s)**

Alencar Xavier

**References**


**Examples**

```r
# Not run:
demo(mixedmodel)
```

```r
# End(Not run)```
Description

Complementary functions that may help with handling parameters and routine operations.

Details

```r
emGWA(y, gen) # Simple MLM for association analysis
markov(gen, chr=NULL) # Markovian imputation of genotypes coded as 012
IMP(X) # Imputes genotypes with SNP expectation (column average)
CNT(X) # Recodes SNPs by centralizing columns in a matrix
GAU(X) # Creates Gaussian kernel as exp(-Dist2/mean(Dist2))
GRM(X, Code012=FALSE) # Creates additive kinship matrix VanRaden (2008)
SPC(y, blk, row, col, rN=3, cN=1) # Spatial covariate
SPM(blk, row, col, rN=3, cN=1) # Spatial design matrix
SibZ(id, p1, p2) # Pedigree design matrix compatible to regression methods
Hmat(ped, gen=NULL) # Kinship combining pedigree and genomics
```

Author(s)

Alencar Xavier
Index

BayesA (WGR1 (MC)), 3
BayesA2 (WGR1 (MC)), 3
BayesB (WGR1 (MC)), 3
BayesB2 (WGR1 (MC)), 3
BayesC (WGR1 (MC)), 3
BayesCpi (WGR1 (MC)), 3
BayesDpi (WGR1 (MC)), 3
BayesL (WGR1 (MC)), 3
BayesRR (WGR1 (MC)), 3
BayesRR2 (WGR1 (MC)), 3
bWGR (bWGR-package), 2
bWGR-package, 2
chr (Dataset), 2
CNT (XTRA 2), 11

Dataset, 2

eMBA (WGR2 (EM)), 6
eMBB (WGR2 (EM)), 6
eMBC (WGR2 (EM)), 6
eMBL (WGR2 (EM)), 6
eMCV (WGR2 (EM)), 6
eMDE (WGR2 (EM)), 6
eMEN (WGR2 (EM)), 6
eMGWA (XTRA 2), 11
eMML (WGR2 (EM)), 6
eMML2 (WGR2 (EM)), 6
eMRR (WGR2 (EM)), 6

fam (Dataset), 2

GAU (XTRA 2), 11
gen (Dataset), 2
GRM (XTRA 2), 11
GS2EIGEN (XTRA 1), 9
GSEN (WGR1 (MC)), 3
GSFLM (XTRA 1), 9
GSRR (XTRA 1), 9
Hmat (XTRA 2), 11

IMP (XTRA 2), 11
KMUP (WGR1 (MC)), 3
KMUP2 (WGR1 (MC)), 3
markov (XTRA 2), 11
mcmcCV (WGR1 (MC)), 3
mixed (XTRA 1), 9
mkr (WGR3 (MV)), 8
mkr2X (WGR3 (MV)), 8
mm (XTRA 1), 9
mrr (WGR3 (MV)), 8
mrr2X (WGR3 (MV)), 8
MRR3 (WGR3 (MV)), 8
mtgsru (XTRA 1), 9
mtmixed (XTRA 1), 9
NNS (XTRA 1), 9
NNSEARCH (XTRA 1), 9
SibZ (XTRA 2), 11
SimY (WGR3 (MV)), 8
SPC (XTRA 2), 11
SPM (XTRA 2), 11
tpod (Dataset), 2
wgr (WGR1 (MC)), 3
WGR1 (MC), 3
WGR2 (EM), 6
WGR3 (MV), 8

XTRA 1, 9
XTRA 2, 11

y (Dataset), 2