Genome-wide Regression Analysis using Bayesian Mixed Linear Models

Md Mahedi Hasan
Swarnita Chakraborty

Abstract

Many modern genomic data analyses require implementing regressions where the number of parameters (p, e.g., the number of marker effects) exceeds the sample size (n). Implementing these large-p-with-small-n regressions poses several statistical and computational challenges, some of which can be confronted using Bayesian methods. This approach allows integrating various parametric and non-parametric shrinkage and variable selection procedures in a unified and consistent manner. In this study, we have used parametric Bayesian methods for estimating the effect of the genetic marker (e.g., SNPs) on the selected phenotype of a publicly available "mice" data. The application of different methods exhibits that for large sample the estimated effect for different markers is not much different, but the complexity of the model is different for different choice of priors. We also have estimated the effect of the genetic marker using the genome-wide regression model with both genetic and non-genetic predictors. The results showed that incorporation of the non-genetic factors improve the Bayesian prediction to estimate the effect of the markers, however; it is also revealed that for large data set, the regularization of the parameters require large chain to have a precise estimate.

Keywords: Genome-wide regression, Bayesian analysis, genotype, phenotype, LASSO

1.0 Introduction

Modern statistical learning problem requires dealing with high dimensional data. This is particularly common in genetic studies. For an instance, one may need to regress the phenotypes on large number of predictors (e.g., SNPs). In such cases, number of parameters is greater than the sample size. This may possess statistical computational challenges. The high dimensionality of data is also another matter of concern. Not all the predictors of the dataset are equally important, thus, the dimension reduction techniques could make the problem easier to deal without losing much information from the data. Here comes the issue of shrinkage and variable selection before running the regression analysis. Bayesian analysis could be a solution for all these situations. Because of the high dimension of the genetic data, the whole-genome regression is becoming popular day by day (Perez and Campos, 2014). The Bayesian setup of the genome-wide regression allows to incorporate with the prior information and update the posterior estimates. Another important advantage of Bayesian regression is that this allows to select the variables and to reduce the dimension of the data which is necessary while dealing with large dimensional data. This study deals with Bayesian models for estimating the genotypic information such as the effects of the markers on the

phenotypic traits of the mice. For analytical convenience, body mass index is chosen as the candidate phenotypic trait for the mice. The parametric Bayesian regression and the genome-wide regression were used to estimate the effect of the markers (e.g. SNPs) and also the combine effect of marker and non-genetic factors such as sex, litter size, case etc.

2.0 Objective of the study

- Implementing the genome-wide Bayesian mixed linear regression models to identify the association of mice SNPS (or markers) with its behavioral and physiological phenotype and infer about the marker effects.

3.0 Related Works

With the increase of volume and variety of data the application of whole genome-wide regression approaches has become popular for analysing and predicting the complex traits of the biological data (Meuwissen et al. 2001). In the recent past, it is found that different parameter and non-parametric techniques were applied, and the results are demonstrating that there is no such single estimation method exists which can uniformly perform to give the best results. However, the choice of the appropriate method depends on number of factors such as the sample size, the architecture of the traits, marker density, the span of the linkage etc (de los Campos et al. 2013). In a study by (Li et. al., 2011) for the heart rate data also used genome-wide regression analysis and they detected several significant genes that are associated with body mass index. Donnelly (2008) conducted a study in humans in order to see the progress and challenges in genome-wide association and after more than a decade long experiment, they reached to a conclusion on the genetic basis of many common human disease.

4.0 Data and Data Source

In this project we have used publicly available data called “mice” data. This dataset is available in the (http://gscan.well.ox.ac.uk) and this has been used for detection of quantitative trait loci (QTL) and for whole-genome regression analysis. It has genotype and phenotype information for 1814 mice. Each mouse was genotyped at 10346 SNPs. For convenience, we removed the SNPs those are with minor allele frequencies less than 0.05. In case of missing values, they are imputed with the expected values computed with the estimates of the allele frequencies derived from the same data.

5.0 Key Terminologies

Gene: A gene is the basic physical and functional unit of heredity. In biology, a gene is a sequence of nucleotides in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein

Genetic Marker: A genetic marker is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. For example, single-nucleotide polymorphism (SNP)

Single-nucleotide Polymorphism (SNP): A single-nucleotide polymorphism is a substitution of a single nucleotide at a specific position in the genome, that is present in a sufficiently large fraction of the population.

Phenotype: Phenotype is the term used in genetics for the composite observable characteristics or traits of an organism such as height, eye color, blood type etc.

Genotype: The genetic contribution to the phenotype is called the genotype.

Allele: one of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.

6.0 Methodology

6.1 Parametric Bayesian Regression

Parametric Bayesian regression can be implemented for both continuous and categorical response variables. In our case, the response variable is the body mass index of the mice, so, for a continuous regression, the data equation is:

$$y_i = \eta_i + \varepsilon_i$$

Where η_i is a linear predictor (the expected value of y_i given predictors). ε_i are independent normal model residuals with mean zero and variance $w_i^2\sigma_\varepsilon^2$. w_i' are the user defined weights and σ_ε^2 is the residual variance parameter. Now, the linear predictor that represents the conditional expectation function is:

$$\boldsymbol{\eta} = 1\mu + \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \sum_{l=1}^{L} \mathbf{u}_l,$$

Where, where μ is an intercept, $\mathbf{X}_j$ are design matrices for predictors, $\boldsymbol{X}_j = \{x_{ijk}\}$, $\boldsymbol{\beta}_j$ are vectors of effects associated to the columns of $\mathbf{X}_j$, and $\mathbf{u}_l = \{u_{l1}, \ldots, u_{ln}\}$ are vectors of random effects. The only element of the linear predictor included by default is the intercept. The other elements are user specified. And therefore, based on these assumptions, the conditional distribution of the data is

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} N\left(y_i|\mu + \sum_{j=1}^{J}\sum_{k=1}^{K_j} x_{ijk}\beta_{jk} + \sum_{l=1}^{L} u_{li}, \sigma_\varepsilon^2 w_i^2\right)$$

where θ represents the collection of unknowns, including the intercept, regression coefficients, other random effects, and the residual variance.

6.1.1 Regression Coefficients

Regression coefficients for this model can be assigned with uninformative or informative priors. The flat priors (uninformative) can only estimate the fixed effects based on the information contained in the likelihood function only. But for the coefficients with informative priors plays a vital role depending on the choice of the priors. Different prior uses different types of shrinkage of estimates

to find the effect in the model. In this project, we have dealt with both uninformative and informative priors, but the results of the informative priors are shown here.

6.1.2 Prior Density

The prior density is assumed to admit the following factorization

$$p(\boldsymbol{\theta}) = p(\mu)\, p\left(\sigma_{\varepsilon}^{2}\right) \prod_{j=1}^{J} p\left(\boldsymbol{\beta}_{j}\right) \prod_{l=1}^{L} p\left(\mathbf{u}_{l}\right)$$

Here, the intercept is assigned a flat prior and the residual variance is assigned a scaled-inverse χ^2 density. Table 1, list out the priors that are used for the Bayesian models with their hyperparameters.

Table 1: Prior densities used for the parametric Bayesian regression

Model (Prior Density)	**Hyperparameters**
Scaled-t (BayesA)	- Degrees of freedom (df) - Scale (S)
Scaled-t mixture (BayesB)	- Pi (prop. Of nonnull effects)
Gaussian mixture (BayesC)	- Pi (prop. Of nonnull effects) - df_{β}
Double Exponential (BayesLASSO)	- λ^2

Here, different priors can be specified for each of the set of coefficients of the linear predictor which gives a greater flexibility for fitting the model for analysis.

6.1.3 Basics Understanding of the selected Bayesian Models

According to the specifications of (Meuwissen, 2017), the basic purpose and understanding of the selected Bayesian models are given here:

BayesA: BayesA is a mixed model where it is assumed that each marker is a random effect, and this solved via MCMC.

BayesB: BayesB is also a mixed model but it has variable selection approach which means the model can assign zero effect to a marker. This also run through MCMC and it will never be a true zero but something very close to zero.

BayesC: The BayesC model assumes a prior density that the markers have normally distributed effects with a certain probability (p) and no effect with the opposite of that probability (1 − p). Marker effects are estimated by using Singular Value Decomposition (SVD) and the posterior probability of the marker having a non-zero effect is also calculated.

Bayes LASSO: Bayesian LASSO is used to select a subset of significant SNPs. The Bayesian lasso is implemented with a hierarchical model, in which scale mixtures of normal are used as prior distributions for the genetic effects and exponential priors are considered for their variances, and then solved by using the Markov chain Monte Carlo (MCMC) algorithm.

6.2 Fitting models for genetic and nongenetic factors

In the setup of hyper-dimensional problems, when a model has more parameter than observations (i.e., p > n) then genome-wide regression can estimate the parameters without having to compute the large matrices (Xavier et al., 2016). This regression also can incorporate with both genotypical information and other covariates. For a linear model, the equation for a genome-wide regression is as follows:

$$y = X\beta + M\alpha + e$$

Where, the regression computes the additive values (α_i) for each marker (m_i) and the coefficient (b) measures the effect of the covariates (X).

For our case, we fitted the model with various set of predictors using the mice dataset. What we tried to point out here is that the case where the mice were housed had important effect on the psychological covariates. It also found in different literature (Legarra et al. (2008) and de los Campos et al. (2009)) that sex, litter size, familial relationships and markers have effect on the psychological covariates such as SNPs. Therefore, a possible linear model with the continuous response variable is:

$$y = \mu + X_1\beta_1 + X_2\beta_2 + M\alpha + \varepsilon$$

where y is the phenotype vector (body mass index, in the example), μ is an intercept, X_1 is a design matrix for the effects of sex and litter size, β_1 is the corresponding vector of effects, X_2 is the design matrix for the effects of cage, β_2 is a vector of cage effects, M is the matrix with marker genotypes, and α is the corresponding vector of marker effects. We treat β_1 as "FIXED" and the other two vectors of effects as random; β_2 is treated as Gaussian and marker effects, α, are assigned IID double-exponential priors, which corresponds to the prior used in the Bayesian LASSO model.

7.0 Results and Discussion

7.1 Parametric Bayesian Regression

First of all, we fitted parametric Bayesian regression models with different priors and estimated the effect of the markers (SNPs) on the phenotypic (BMI) trait of the mice. The estimated posterior means and standard deviation of the posterior mean are also obtained for fitted models (Appendix 1). The following figures are showing the results of the estimated effects for 4 Bayes models with 4 different priors.

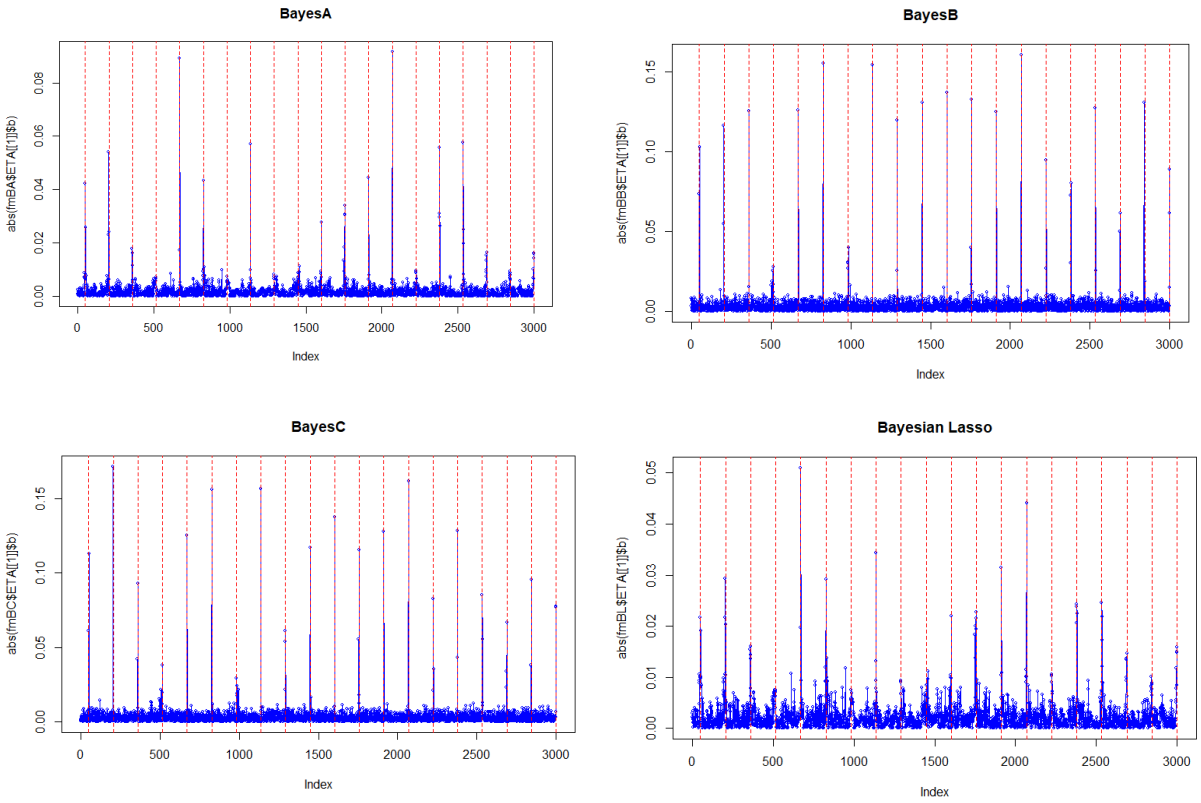

Figure 1: Absolute value of the estimated marker effects by different Bayesian models for the phenotypes of mice

Figure 1, is showing the results that obtained from different bayes models (BayesA, BayesB, BayesC, and Bayed LASSO). These plots are showing the estimated effect of the markers (e.g., SNPs) on the phenotypes. The vertical blue lines are showing the absolute effects of the markers. The lines with larger estimates are showing the prominent effect of the specific markers on the specified phenotypes (in our case, the body mass index of the mice). Larger the estimates or longer the vertical lines indicate the greater effect of the corresponding marker (e.g. SNPs). One interesting finding that is derived from these plots is that for the larger dimension of data, the effect of the markers is not much different in the different model set ups except the complexity of the model. If we look at the marker effects from these 4 types of bayes models we will see the marker effects are not much different in different model. So, the less complex model can be adopted to estimate the effect of the markers. This result is also supported by different literatures in the application of genomic studies (Perez and Campos, 2016). The following section will describe about the model selection criteria based on some model selection criteria.

7.1.1 Model Fit and complexity indices

Table 2, provides estimated residual variance, the deviance information criterion (DIC), and the effective number of parameters(pD). Effective number of parameters (pD) measure the complexity of the fitted model. So, higher the value of the pD greater the complexity of the model. In case of deviance information criterion, smaller is better.

Table 2: Measure of fit and Model complexity

Model	Residual Variance	Deviance Information Criterion (DIC)	Effective Number of Parameters (pD)
BayesA	0.914	5066	83.4
BayesB	0.4773	3858	56.4
BayesC	0.4789	3961	53.4
Bayes LASSO	0.4919	4076	218

Looking at the effective number of parameters, it is clear that the bayes LASSO is having the highest number of parameters (218), so, the most complex model whereas bayesC is having the least number of parameters amongst all 4 models, so, is the least complex model. Based on the deviance information criterion, bayesB is having the smallest value, so, DIC favoured bayesB over other models. However, if we also look at the DIC values of bayesB and bayesC, we will see a small difference of their DIC values So, the performance of the bayesB and bayesC is not much different to estimate the marker effects.

7.2 Genome-wide regression model for genetic and nongenetic factors

After fitting the Bayesian models for estimating the marker effects in the previous section, in this section we have discussed about the results for the genome-wide regression estimates for both genetic and non-genetic factors. Alongside the genetic factors, the non-genetic factors that are considered in this model is sex, litter size, cage where the mice were housed. This Bayesian model allows us to study the effect of the markers and these non-genetic covariates together. The fitted model provides us the posterior means and the estimated posterior standard deviations as well as the goodness of fit statistics and also the information on the model complexity.

In the model, the phenotypes were standardized to a unit sample variance and the estimated residual variance is found 0.537 which suggesting that the model could explain about 46% of the phenotypic variance (Table 3). From figure2 (next page), the top-left plot is giving the absolute values of the estimated effect of the marker. From the vertical line we can identify which of the markers are having larger effects on the phenotypes of the mice. Top-right scatter plot is giving idea about the observed phenotype vs the predicted genomic values. This also gives a rough understanding about the fit of the model. Now, the bottom-left trace plot of the residual variance is giving idea about the distribution of the residual for the fitted model and the residual variance had a very good mixing which indicates the good fit of the model. However, the mixing of the regularization parameter is not as good (bottom-right plot). The reason may be that we have large number of markers in the model and that requires a long chain to infer the regularization parameters precisely.

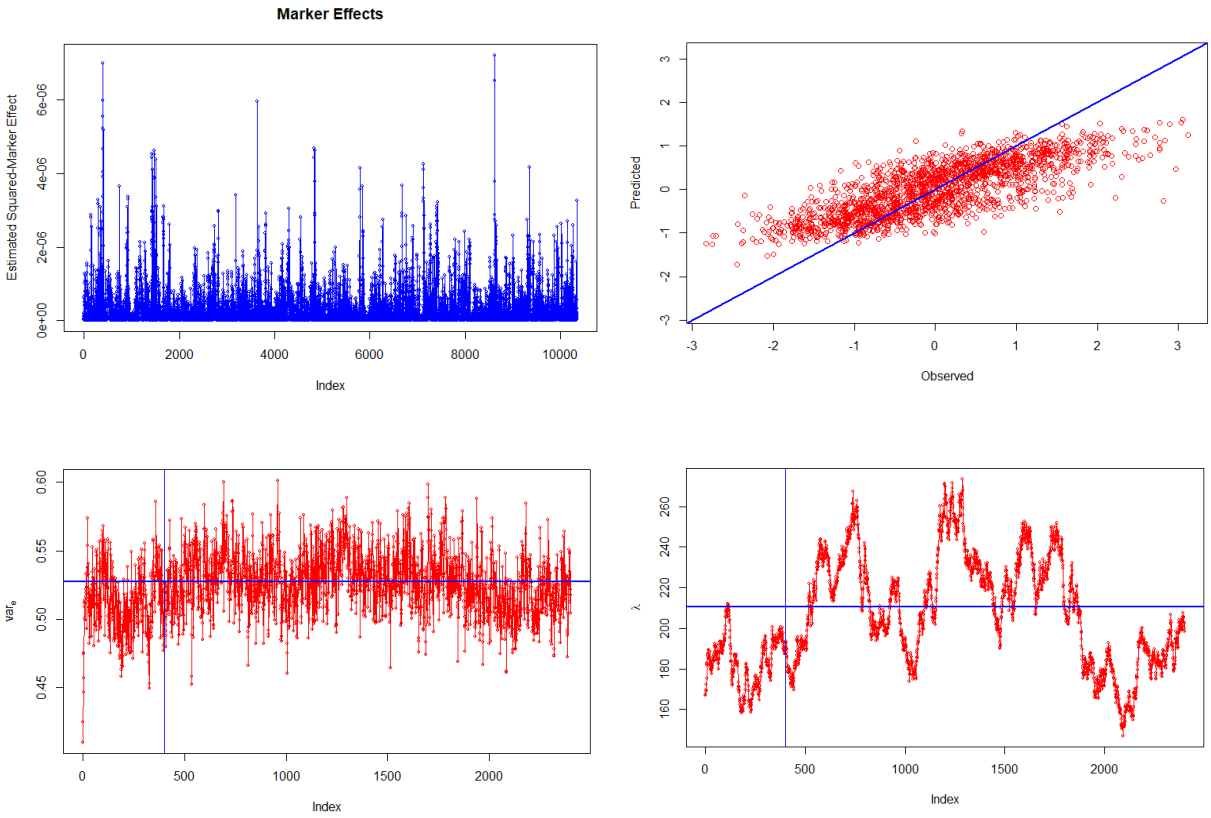

Figure 2: squared estimated marker effects (top-left), phenotype vs predicted genomic values (top-right), trace plot for residual variance (bottom-left), and plot of regularization parameter of the Bayesian model (bottom-right)

The following section will describe about the model fit and the complexity parameters.

7.2.1 Model Fit and Complexity Indices

The model fit statistics and the complexity indices are giving an understanding about the overall fit of the model. However, these estimates will make more sense if we compare this model with further models, like, a model with different set of non-genetic factors and so.

Table 3: Model fit and model complexity indices for the genetic and non-genetic factor model

Model Fit Indices	**Estimates**
Posterior Mean	-0.427
Posterior Standard Deviation	0.289
Residual Variance	0.537
Effective number of Parameters (pD)	345
Deviance Information Criterion (DIC)	4341

8.0 Conclusion

In this project, we implemented several methods of unified Bayesian framework for genome-wide regression predictions. The results showed that the effects of the markers are not much different for different priors, but the complexity of the models are different for different priors. For large data, the effects are not changing suddenly with the change of the prior. Bayesian estimates are good for choosing the subset of the large data and to implement with the genomic data to find the effects of the genomic traits. Moreover, genome-wide regression allows to find the effect of the markers along with the other covariates on the phenotype of the mice.

9.0 Acknowledgement

We are grateful to the authors of the research papers those we reviewed and assisted by. We are also thankful to the open data repository for the dataset “mice” that we have used in our study. We would also like to express our gratitude to the authors of R package “BGLR” which we have used for data analysis purpose. Last but not the least, we are grateful to our course instructor Dr. Yuan Wang for her excellent demonstration of the course content which taught us throughout the semester and helped to conduct this study.

10.0 Appendix

Appendix Table 1: Posterior mean and Standard deviation for the bayes models

Model	Posterior Mean	Posterior Standard Deviation of the mean
BayesA	-0.0076	0.016
BayesB	-0.0070	0.0167
BayesC	-0.0069	0.0172
Bayes LASSO	- 0.0079	0.0186

Appendix 2: R Codes for running the models

```
################################################################
## Installing the required libraries
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

library(survival)
library(Matrix)
library(BGData)
library(snpStats)
library(foreach)
library(iterators)
library(parallel)
library(doParallel)
library(gdsfmt)
library(SNPRelate)
library(GenABEL.data)
library(GenABEL)
```

```
library(dplyr)


#####################################################################
## Parametric Bayesian Models
library(BGLR)
data(mice)
X=scale(mice.X[,1:3000])
pheno=mice.pheno

# Scaled-t (BayesA)
fmBA=BGLR(y=y,ETA=list( list(X=X,model='BayesA')),
          nIter=nIter,burnIn=burnIn,saveAt='ba_')
summary(fmBA)
str(fmBA)
plot(abs(fmBA$ETA[[1]]$b),col=4,cex=.5,
type='o',main='BayesA');abline(v=QTL,col=2,lty=2)
str(fmBA)

# Point of mass at zero + t-Slab (BayesB)
fmBB=BGLR(y=y,ETA=list( list(X=X,model='BayesB')),
          nIter=nIter,burnIn=burnIn,saveAt='bb_')
summary(fmBB)
str(fmBB)
plot(abs(fmBB$ETA[[1]]$b),col=4,cex=.5,
type='o',main='BayesB');abline(v=QTL,col=2,lty=2)

# Point of mass at zero + Gaussian Slab (BayesC)
fmBC=BGLR(y=y,ETA=list( list(X=X,model='BayesC')),
          nIter=nIter,burnIn=burnIn,saveAt='bc_')
summary(fmBC)
str(fmBC)
plot(abs(fmBC$ETA[[1]]$b),col=4,cex=.5,
type='o',main='BayesC');abline(v=QTL,col=2,lty=2)


#  Double-Exponential (Bayesian Lasso)
fmBL=BGLR(y=y,ETA=list( list(X=X,model='BL')),
          nIter=nIter,burnIn=burnIn,saveAt='bl_')
summary(fmBL)
str(fmBL)
plot(abs(fmBL$ETA[[1]]$b),col=4,cex=.5, type='o',main='Bayesian
Lasso');abline(v=QTL,col=2,lty=2)

#####################################################################
## Fitting a model to markers and non-genetic effects in BGLR

#1# Loading and preparing the input data
library(BGLR)
data(mice)
Y<-mice.pheno
X<-mice.X
A=mice.A
y<-Y$Obesity.BMI
y<-(y-mean(y))/sd(y)

# Setting the linear predictor
ETA<-list( list(~factor(GENDER)+factor(Litter),
                data=Y,model='FIXED'),
           list(~factor(cage),data=Y, model='BRR'),
           list(X=X, model='BL')
```

```
)

# Fitting the model
fm<-BGLR(y=y,ETA=ETA, nIter=12000, burnIn=2000)
summary(fm)
ls(fm)
str(fm)

# Estimated Marker Effects & posterior SDs
bHat<- fm$ETA[[3]]$b
SD.bHat<- fm$ETA[[3]]$SD.b
plot(bHat^2, ylab='Estimated Squared-Marker Effect',
     type='o',cex=.5,col=4,main='Marker Effects')


# Predictions
yHat<-fm$yHat
tmp<-range(c(y,yHat))
plot(yHat~y,xlab='Observed',ylab='Predicted',col=2,
     xlim=tmp,ylim=tmp); abline(a=0,b=1,col=4,lwd=2)


# Genomic part of the model
gHat<-X%*%fm$ETA[[3]]$b
plot(gHat~y,xlab='Phenotype',
     ylab='Predicted Genomic Value',col=2,
     xlim=tmp,ylim=tmp); abline(a=0,b=1,col=4,lwd=2)


#3# Godness of fit and related statistics
fm$fit
fm$varE

#Trace plots
list.files()
# Residual variance
varE<-scan('varE.dat')
plot(varE,type='o',col=2,cex=.5,ylab=expression(var[e]));
abline(h=fm$varE,col=4,lwd=2);
abline(v=fm$burnIn/fm$thin,col=4)

# lambda (regularization parameter of the Bayesian Lasso)
lambda<-scan('ETA_3_lambda.dat')
plot(lambda,type='o',col=2,cex=.5,ylab=expression(lambda));
abline(h=fm$ETA[[3]]$lambda,col=4,lwd=2);
abline(v=fm$burnIn/fm$thin,col=4)
#####################################################################
```

11.0 Reference

de los Campos, Gustavo, et al. "Whole-genome regression and prediction methods applied to plant and animal breeding." *Genetics* 193.2 (2013): 327-345.

Donnelly, Peter. "Progress and challenges in genome-wide association studies in humans." *Nature* 456.7223 (2008): 728-731.

Li, Jiahan, et al. "The Bayesian lasso for genome-wide association studies." *Bioinformatics* 27.4 (2011): 516-523.

Meuwissen, Theo HE, Ulf G. Indahl, and Jørgen Ødegård. "Variable selection models for genomic selection using whole-genome sequence data and singular value decomposition." *Genetics Selection Evolution* 49.1 (2017): 1-9.

Pérez, Paulino, and Gustavo de Los Campos. "Genome-wide regression and prediction with the BGLR statistical package." *Genetics* 198.2 (2014): 483-495.

Xavier, Alencar, et al. "Walking through the statistical black boxes of plant breeding." *Theoretical and applied genetics* 129.10 (2016): 1933-1949.