Documents for VIGoR

May, 2015

Akio Onogi
onogiakio@gmail.com

Contents

1. Introduction

1-1. Overview

VIGoR (Variational Bayesian Inference for Genome-wide Regression) conducts fast genome-wide regression using variational Bayesian inference (VB). VIGoR comprises two programs, *vigor* and *hyperpara* (Fig. 1-1). *vigor* has three main functions: 1) ***Model fitting***, 2) ***Model fitting after hyperparameter tuning***, and 3) ***Cross-validation*** (Fig. 1-1). Using the default function ***Model fitting***, users can conduct variable selection (association mapping) by fitting genome-wide regression models to data and estimating the marker effects. ***Model fitting after hyperparameter tuning*** estimates the marker effects with hyperparameters that are automatically tuned using cross-validation (CV). ***Cross-validation*** evaluates the predictive ability of the regression models using CV. Required inputs to *vigor* are phenotypic values, marker genotypes, and hyperparameter values. The *hyperpara* program calculates the values of hyperparameters that influence the inference, based on several assumptions about the genetic architecture and values of hyperparameters that affect the inference less.

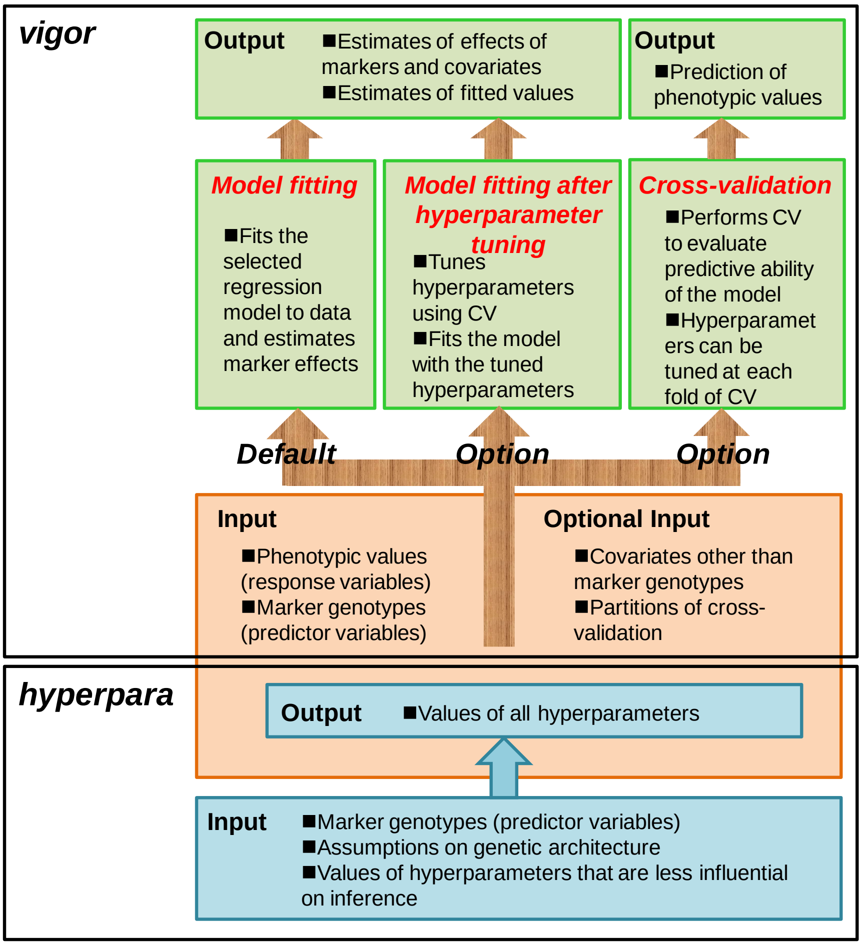
vigor
Output ■Estimates of effects of markers and covariates ■Estimates of fitted values
Output ■Prediction of phenotypic values
Model fitting
■Fits the selected regression model to data and estimates marker effects
Model fitting after hyperparameter tuning
■Tunes hyperparameters using CV ■Fits the model with the tuned hyperparameters
Cross-validation
■Performs CV to evaluate predictive ability of the model ■Hyperparameters can be tuned at each fold of CV
Default
Option
Option
Input
■Phenotypic values (response variables) ■Marker genotypes (predictor variables)
Optional Input
■Covariates other than marker genotypes ■Partitions of cross-validation
hyperpara
Output ■Values of all hyperparameters
Input ■Marker genotypes (predictor variables) ■Assumptions on genetic architecture ■Values of hyperparameters that are less influential on inference

Fig. 1-1 Overview of VIGoR.

1-2. Distribution

VIGoR is distributed as a command line program (CLP) package for Linux/Mac and as a cross-platform R package. The CLP package includes:

- Executable files (*vigor* and *hyperpara*)
- Sample files (sample.geno.txt, sample.pheno.txt, sample.covariate.txt, sample.ped, and sample.dose)

We provide distinct packages for Linux and Mac. The CLP package is available at https://github.com/Onogi/VIGoR. The R package includes:

- R functions (*vigor* and *hyperpara*)
- Sample data (comprising Geno, Pheno, and Covariates)
- R documentation

The R package is deposited at CRAN (http://cran.r-project.org/). The programs *vigor* and *hyperpara* are implemented similarly in the CLP and R packages (see **Section 1-3. Quick user guide**).

1-3. Quick user guide

1-3-1. Command line programs

Place the executable and sample files in the same directory. To execute ***Model fitting***, type

```
$./vigor   sample.pheno.txt   sample.geno.txt   BayesC   5 1 0.01
```

The files sample.pheno.txt and sample.geno.txt include the phenotypic values and marker genotypes, respectively. ***Model fitting*** uses a regression method called BayesC (see **Chapter 2. Regression methods**). The last three values (5, 1, and 0.01) are the hyperparameter values required by BayesC (see **Chapter 3. Hyperparameters**). The execution creates an output file, “BayesC_Height_set1.fitting” (see **Subsection 4-5-1. Fitting file**). To execute ***Model fitting after hyperparameter tuning***, type

```
$./vigor   sample.pheno.txt   sample.geno.txt   BayesB   5 1 0.01 -v 5 1 0.1 -t
```

Model fitting after hyperparameter tuning adopts another regression method called BayesB. The -v option is followed by an additional hyperparameter value set (5, 1, and 0.1). The -t option enables hyperparameter tuning which selects the best hyperparameter set among the given sets using CV. Selecting the first set, [5, 1, 0.01], produces an output file “BayesB_Height_set1.fitting”, and selecting the second set, [5, 1, 0.1], generates the file “BayesB_Height_set2.fitting”. To execute ***Cross-validation***, type

```
$./vigor   sample.pheno.txt   sample.geno.txt   BayesC   5 1 0.01 -c 5
```

The -c option signifies CV, and its argument (5 in this example) specifies the fold number. Thus, the above command executes a file-fold CV. The two output files, “BayesC_Height.crossvalidation” and “BayesC_Height.partition” contain the CV results and the partition of individuals in the CV, respectively.

The *hyperpara* program determines the hyperparameters based on several assumptions about the genetic architecture. For example, in the command

```
$./hyperpara    sample.geno.txt    0.5    BayesC   0.01
```

the second argument (0.5) indicates the proportion of phenotypic variance (i.e., variance of response variables) explained by the markers. The last argument (0.01) indicates the proportion of markers with non-zero effects. Thus, in this example, 50% of the phenotypic variance is assumed to be explained by 1% of the markers. Based on these two assumptions, *hyperpara* outputs the following hyperparameter value set in standard output:

```
Genotype file     : sample.geno.txt
Mvar                     : 0.500000
Method                  : BayesC
Kappa                    : 0.010000
Nu                         : 5.000000
Inbreeding coef.: 0.000000
# individuals      : 100
# markers            : 1000

Hyperparameters
Nu    S2    Kappa
5.000000    0.071744    0.010000
```

The final row displays the hyperparameter values. These three values are arguments of *vigor*.

```
$./vigor    sample.pheno.txt    sample.geno.txt    BayesC    5    0.071744    0.01
```

1-3-2. R functions

After installing the package *vigor*, load the package and sample data by typing

```
>library(vigor)
>data(sampledata)
```

To execute ***Model fitting***, type

```
>Result <- vigor (Pheno$Height,   Geno,   "BayesC",   c(5, 1, 0.01) )
```

Pheno is a data frame including phenotypic values of three traits: “Height”, “Weight”, and “Length”. Geno is a matrix of marker genotypes. The regression method is BayesC. The three values in c(5, 1, 0.01) are the hyperparameter values required by BayesC. *Vigor* creates a list object (see **Subsection 6-4-1. The output list of “fitting” or “tuning”**). To execute ***Model fitting after hyperparameter tuning***, type

```
>Result<-vigor(Pheno$Height, Geno, "BayesC", matrix(c(5,1,0.01,5,1,0.1), nrow=2,byrow=T), "tuning" )
```

The fourth argument is a matrix of two hyperparameter sets, [5, 1, 0.01] and [5, 1, 0.1]. The fifth argument, “tuning”, denotes hyperparameter tuning. To plot the absolute values of the estimated marker effects (i.e., to create a Manhattan plot), type

```
>plot (abs (Result$Beta) )
```

To execute ***Cross-Validation***, type

```
>Result <- vigor (Pheno$Height,   Geno,   "BayesC",   c(5, 1, 0.01),   "cv",   5 )
```

In this command line, “cv” indicates CV, and the last argument (5) specifies the fold number. Therefore, this line executes a five-fold CV. To evaluate the prediction accuracy, type

```
>cor (Result$Prediction$Y,   Result$Prediction$Yhat)
```

Result$Prediction$Y and Result$Prediction$Yhat contain the observed (given) and predicted phenotypic values, respectively.

Based on the genetic architecture assumptions, the hyperparameters are determined by *hyperpara*. For example, in the command line

```
>hyperpara( Geno,   0.5,    "BayesC",   0.01)
```

the second argument (0.5) indicates the proportion of phenotypic variance (i.e., variance of response variables) explainable by the markers. The final argument (0.01) indicates the proportion of markers with non-zero effects. Thus, this example assumes that 50% of the phenotypic variance is explained by 1% of the markers. Based on these two assumptions, *hyperpara* outputs the following hyperparameter values:

```
> hyperpara (Geno,   0.5,   "BayesC",   0.01)
         Nu          S2        Kappa
5.00000000   0.07174377   0.01000000
```

The output vector containing these three values (5, 0.0717, 0.01) can be input into *vigor*. For example,

```
>Result <- vigor (Pheno$Height,   Geno,   "BayesC",   hyperpara   (Geno,   0.5,   "BayesC",   0.01)   )
```

1-4. Organization of the manual

Chapters 2 and 3 of this manual briefly explain the regression methods and hyperparameters, respectively. This information is common to the CLP and R packages. Chapters 4 and 5 describe CLPs *vigor* and *hyperpara*, respectively; the corresponding functions in the R package are described in Chapters 6 and 7. In Chapter 8 we provide citations for VIGoR, the regression methods, and variational Bayesian algorithms. The variational Bayesian algorithms and hyperparameter calculation are provided in Appendix A and B, respectively.

2. Regression methods

VIGoR provides seven regression methods: Bayesian lasso (BL), extended Bayesian lasso (EBL), weighted Bayesian shrinkage regression (wBSR), BayesB, BayesC, stochastic search variable selection (SSVS), and Bayesian mixture regression (MIX) (Table 2-1). These methods select the important variables (i.e., the variables related to response variables) among the given predictor variables in different ways (i.e., model structures). The linear regression model assumed in VIGoR is

$$y_i = \sum_{j=1}^{F} z_{ij}\alpha_j + \sum_{p=1}^{P} \gamma_p x_{ip}\beta_p + \varepsilon_i$$

where y_i is the phenotypic value of individual i, F is the number of covariates other than markers, z_{ij} is the covariate corresponding to effect α_j, P is the number of markers, and γ_p is a binary (0 or 1) indicator variable. Here x_{ip} and β_p denote the genotype and effect of marker p, respectively, and ε_i is the residual. Except in wBSR, all indicator variables are fixed to 1. **Note that the regression methods select the important variables from *x*, but not from *z* (that is, all *z* are included in the model).** The residual, ε_i, is assumed to follow a normal distribution with 0 mean and variance $1/\tau_0^2$. More details of the regression methods are provided in **Appendix A**.

All of the regression methods standardize the phenotypic values (response variables) to a mean and standard deviation of 0 and 1, respectively. Two aspects of the standardization should be noted:

1) in the CV, **only the phenotypic values used for training are standardized at each fold**. Testing individuals are excluded from the standardization.
2) VIGoR outputs most of the estimated parameter values on the original scale. **However, some estimates are output in the standardized scale. See Sections 4-5. Output files and 6-4. Output lists for the CLP and R packages, respectively.**

In the default setting, VIGoR starts analyses from the same initial values but randomizes the update order of marker effects. Thus, the results might change across runs. The initial values can also be randomized (see **Sections 4-2. Arguments and options** and **6-2. Optional arguments** for the CLP and R packages, respectively).

Table 2-1 Structures of regression methods[a]

Hierarchical	1st	2nd	3rd
level	Marker effect and indicator	Effect variance and indicator	Shrinkage magnitude
BL	$\beta_p \sim N\left(0, \frac{1}{\tau_0^2 \tau_p^2}\right)$	$\tau_p^2 \sim Inv-G\left(1, \frac{\lambda^2}{2}\right)$	$\lambda^2 \sim G\left(\varphi, \varpi\right)$
EBL	$\beta_p \sim N\left(0, \frac{1}{\tau_0^2 \tau_p^2}\right)$	$\tau_p^2 \sim Inv-G\left(1, \frac{\delta^2 \eta_p^2}{2}\right)$	$\delta^2 \sim G\left(\varphi, \varpi\right)$ $\eta_p^2 \sim G\left(\psi, \theta\right)$
wBSR	$\beta_p \sim N\left(0, \sigma_p^2\right)$ $\gamma_p \sim Bernoulli\left(\kappa\right)$	$\sigma_p^2 \sim \chi^{-2}\left(\nu, S^2\right)$	
BayesB	$\beta_p \sim N\left(0, \sigma_p^2\right)$ if $\rho_p = 1$ $\beta_p = 0$ if $\rho_p = 0$	$\sigma_p^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	
BayesC	$\beta_p \sim N\left(0, \sigma^2\right)$ if $\rho_p = 1$ $\beta_p = 0$ if $\rho_p = 0$	$\sigma^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	
SSVS	$\beta_p \sim N\left(0, \sigma^2\right)$ if $\rho_p = 1$ $\beta_p \sim N\left(0, c\sigma^2\right)$ if $\rho_p = 0$	$\sigma^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	
MIX	$\beta_p \sim N\left(0, \sigma_A^2\right)$ if $\rho_p = 1$ $\beta_p \sim N\left(0, \sigma_B^2\right)$ if $\rho_p = 0$	$\sigma_A^2 \sim \chi^{-2}\left(\nu, S^2\right)$ $\sigma_B^2 \sim \chi^{-2}\left(\nu, cS^2\right)$ $\rho_p \sim Bernoulli\left(\kappa\right)$	

[a] Hyperparameters are highlighted with red.

BL, Bayesian lasso; EBL, extended Bayesian lasso; wBSR, weighted Bayesian shrinkage regression; SSVS, stochastic search variable selection; MIX, Bayesian mixture regression; *N*, normal distribution; *Inv-G*, inverse-gamma distribution; *G*, gamma distribution; *Bernoulli*, Bernoulli distribution; χ^{-2}, scaled inverse-chi-square distribution.

All of the regression methods require hyperparameter values in addition to phenotypic values and marker genotypes. The user-specified hyperparameters are listed in Table 2-2. Hyperparameter specification is described in **Chapter 3. Hyperparameters** and **Appendix B**.

Table 2-2 Hyperparameters

Methods	Hyperparameters
BL	φ, ω
EBL	φ, ω, ψ, θ
wBSR	ν (Nu), S^2, κ
BayesB	ν (Nu), S^2, κ
BayesC	ν (Nu), S^2, κ
SSVS	c, ν (Nu), S^2, κ
MIX	c, ν (Nu), S^2, κ

References of the regression methods and VB algorithms are presented in Table 2-3. The VB algorithms are provided in **Appendix A**.

Table 2-3 References of regression methods and VB algorithms

	Regression methods	VB algorithm
BL	Park and Casella (2008)	Li and Sillanpaa (2012)
EBL	Mutshinda and Sillanpaa (2010)	Li and Sillanpaa (2012)
wBSR	Hayashi and Iwata (2010)	Hayashi and Iwata (2013)
BayesB	Meuwissen et al. (2001)	Onogi and Iwata (2015)
BayesC	Habier et al. (2011)	Carbonetto and Stephens (2012)
SSVS	George and McCulloch (1993)	Onogi and Iwata (2015)
MIX	Luan et al. (2009)	Onogi and Iwata (2015)

3. Hyperparameters

Choosing the hyperparameter values is often problematic in Bayesian computation. Users might need to test different values and find a case-dependent solution. The function "***Model fitting after hyperparameter tuning***" determines the hyperparameter values by CV, which tends to select redundant markers; therefore, this approach is suitable for prediction but not for association mapping. Alternatively, hyperparameter values can be specified based on several assumptions, as proposed by Habier et al. (2011). In this approach, the more influential hyperparameters are determined from assumptions about the genetic architecture and values of less influential hyperparameters (Table 3-1). This approach is implemented by the *hyperpara* function, which is provided in both CLP and R (see **Chapters 5. Command line program *hyperpara*** and **7. R function *hyperpara***). The calculation is explained in **Appendix B**.

Table 3-1 Hyperparameters determined by *hyperpara*

Regression	Less influential hyperparameters (values given as default)	Assumption	Influential hyperparameters determined by *hyperpara*
BL	φ (1.0)	κ, *Mvar*	ω
EBL	φ (0.1), ω (0.1), ψ (1.0)	κ, *Mvar*	θ
wBSR	ν (5.0)	κ, *Mvar*	S^2
BayesB	ν (5.0)	κ, *Mvar*	S^2
BayesC	ν (5.0)	κ, *Mvar*	S^2
SSVS	ν (5.0)	κ, *Mvar*, *A*	c, S^2
MIX	ν (5.0)	κ, *Mvar*, *A*	c, S^2

The required assumptions (κ, *Mvar*, *A*) are as follows:

- κ : proportion of markers with non-zero effects.
- *Mvar* : proportion of variance of phenotypic values (response variables) explainable by markers.
- *A* : proportion of *Mvar* explainable by markers assigned to the prior normal distribution with larger variance (σ^2 for SSVS and σ_A^2 for MIX). The default is 0.9.

For example, when the regression method is BL, and κ and *Mvar* are respectively set to 0.01 and 0.5, a half of the phenotypic variance (i.e., variance of response variables) is assumed to be explained by 1 % of the markers. When the regression method is SSVS, and κ, *Mvar*, and *A* are respectively set to 0.01, 0.5, and 0.9, 45% (0.5 × 0.9) of the phenotypic variance is assumed to be explained by 1% of the markers, and 5% (0.5 × 0.1) of the variance is explained by 99% of the markers.

4. Command line program *vigor*

4-1. Input files

The CLP *vigor* requires phenotype and genotype input files, which respectively include the phenotypic values (response variables) and marker genotypes (predictor variables). Optional input files are a covariate file of covariates other than markers and a partition file including the CV partitions. In the current version, **allowable input files are tab- or space-delimited text files**. **The number of individuals and their orders should be consistent among the genotype, phenotype, and covariate files.**

***Vigor* accepts PED files (.ped) of PLINK (Purcell et al. 2007).** Because PED files contain both phenotypic values and marker genotypes, *vigor* requires only a single PED file (see **Subsection 4-1-5. PED file** and **Section 4-6. Examples of usage**). ***Vigor* also accepts genotype dosage files (.dose) output by Beagle (Browning SR and Browning BL 2007) as the Genotype file.** See **Section 4-6. Examples of usage**. *Vigor* automatically recognizes these files by their extensions (.ped and .dose).

4-1-1. Phenotype file

The phenotypic values contained in the phenotype file are used as response variables. A single phenotype file can contain multiple traits. **The first row contains the trait names.** The maximum length of trait names is 100. Missing values are specified as NA.

Ex.) The following phenotype file contains two trait records of three individuals.

Height	Weight
40.5	50
20.9	NA
NA	102

The third individual lacks the record for “Height” and the second lacks the “Weight” record.

4-1-2. Genotype file

This file includes the marker genotypes. *Vigor* assumes bi-allelic markers, but multi-allelic markers can be handled as shown in Example 2 below. **Marker genotypes should be coded in an additive manner, such as −1, 0, and 1, or 0, 1, and 2, which respectively correspond to AA, AB, and BB. We recommend using the 0, 1, and 2 coding, because this coding alone is acceptable by the CLP *hyperpara* program (see Section 5-1. Input files).** The rows of the genotype file contain the marker genotypes of the individuals. Individual IDs or Marker IDs are not allowed. Consequently, the file is an ($N \times P$) matrix, where N and P denote the numbers of individuals and markers, respectively. The -o option enables using a ($P \times N$) matrix (see **Section 4-2. Arguments and options**). Because **no missing values are allowed**, the marker genotypes should be imputed before analysis. Real numbers (e.g., 1.24 or 0.92) are accepted.

Ex. 1). The following genotype files list four markers (columns) for three individuals (rows). Genotypes are coded as 0 (AA), 1 (AB), and 2 (BB).

0	2	1	0
1	1.2	1	1.8
2	0	2	2

For the second individual, the second (1.2) and fourth (1.8) marker genotypes are imputed, and the dosages of the B alleles are presented.

Ex. 2). The following genotype file lists one multi-allelic marker for three individuals. The marker consists of three alleles, A, B, and C. The genotypes of the first, second, and third individuals (rows) are AA, AB, and CC, respectively. Note that the first, second, and third columns indicate the number of A, B, and C alleles, respectively.

2	0	0
1	1	0
0	0	2

Ex. 3). By specifying the -o option, an ($P \times N$) matrix can be used as the genotype file. The following genotype file contains four markers (rows) of three individuals (columns), which is same as Ex. 1.

0	1	2
2	1.2	0
1	1	2
0	1.8	2

***Vigor* accepts genotype dosage files (.dose) created by Beagle. The .dose.gz files need to be extracted before use**. See the Beagle manual for the format of dose files.

4-1-3. Covariate file (optional)

This file contains the covariates included in the regression models besides the marker genotypes. Covariates included in this file are treated as "fixed effects" (i.e., non-informative prior distributions are assigned). **No missing values are allowed**. Note that

1) **when no covariate file is provided by the user, *vigor* automatically adds the intercept (overall mean) to the regression models, and**
2) **when a covariate file is provided, *vigor* regards the first column of the covariate file as the intercept.**

Ex. 1). The following is the covariate file of three individuals. The first column is the intercept. The second and third columns are the covariates. The total number of covariates is three.

1	0.2	11
1	0.4	9
1	1.2	18

Ex. 2). Again consider three individuals. Suppose that the first, second, and third individuals are respectively cultivated in fields "A", "B", and "C". Three field effects are represented in the following covariate matrix.

1	0	0
1	1	0
1	0	1

Herein, the second and third covariates (columns) indicate the relative effects of fields "B" and "C" on field "A". The intercept (first column) can be regarded as the mean of field "A". The number of covariates is three.

Ex. 3). Consider a matrix whose covariates are the probabilities that individuals belong to sub-populations (a so-called Q matrix). In this case, the Q matrix can be used as the covariates without the intercept. For example, a Q matrix of four sub-populations can be expressed as the following covariate file.

0.6	0.1	0.1	0.2
0.05	0.45	0.5	0.0
0.2	0.6	0.1	0.1

Here, the first individual belongs to sub-populations one, two, three, and four with probabilities of 0.6, 0.1, 0.1, and 0.2, respectively. The number of covariates is four.

4-1-4. Partition file (optional)

This file specifies the partitions of individuals in cross-validation or random sampling validation. *Vigor* can execute CV without this file by randomly partitioning individuals. In this case, *vigor* outputs the partition as the partition file, which can be used as an input file in subsequent analyses. **Partition files specify the individuals used in the prediction (i.e., individuals that are not used for training) at each fold**.

Ex. 1). The following file partitions 19 individuals in a five-fold CV.

16	5	17	13	9
12	18	3	14	6
8	7	11	15	19
1	10	2	4	−9

This matrix specifies the tested individuals at each fold. **The elements correspond to the row numbers of the genotype/phenotype file**. In the first fold, individuals 16, 12, 8, and 1 are excluded from training and predicted. In the second fold, individuals 5, 18, 7, and 10 are excluded and predicted. Spaces in the matrix are filled with "−9" (in this example, the fourth individual is missing in the fifth fold).

The partition file is applicable to random sampling validation, in which individuals are not used for testing exactly once.

Ex. 2). The following file shows five splits of 19 individuals. Four individuals are tested in each split.

18	3	11	16	13
17	8	13	13	18
7	15	14	19	7
1	13	12	7	2

In this example, individuals 18, 13 and 7 are repeatedly used for testing.

4-1-5. PED file (optional)

PLINK can create multiple PED files by specifying various options. *Vigor* accepts the most basic file type, demonstrated at http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml. The first six columns of the file (Family ID, Individual ID, Paternal ID, Maternal ID, Sex, and Phenotype) are mandatory. Among these, only “Phenotype” is read by *vigor*; the others are ignored. **Missing phenotypic records should be coded as −9**. Subsequent columns contain the marker genotypes. Each row stores the information of one individual. Alleles can be coded by any **single character**, such as A, C, G, T or 1, 2, 3, 4. **Note that 0 cannot be used, because 0 is regarded as a missing value by *vigor*, and missing genotypes (alleles) are not allowed.** Only bi-allelic markers are allowed.

Ex.) The following PED file contains the information of three individuals (number of markers = 2).

```
FAM1   ID001   0   0   0   1.24   A   A   1   2
FAM1   ID002   0   0   0    −9    A   B   2   2
FAM1   ID003   0   0   0   4.12   A   A   1   1
```

The first five columns are ignored. The sixth column, “Phenotype”, is read as the phenotypic values (response variables). The second individual (ID002) lacks a phenotypic record, so is assigned a value of −9. The alleles of the first marker are encoded as “A” or “B”, and those of the second marker are encoded as “1” or “2”.

4-2. Arguments and options

Vigor requires four arguments: phenotype file name, genotype file name, selected regression method, and hyperparameter values. *Vigor* offers several options.

Arguments

Phenotype file

Phenotype file name. A path can be given. Maximum length is 400.

Genotype file

Genotype file name. A path can be given. Maximum length is 400.

Method

Seven regression methods are available, and abbreviated as follows:

- BL : Bayesian Lasso
- EBL : Extended Bayesian Lasso
- wBSR : weighted Bayesian shrinkage regression
- BayesB : BayesB
- BayesC : BayesC
- SSVS : Stochastic search variable selection
- MIX : Bayesian mixture model

Hyperparameter values

The regression methods require hyperparameter values. **The number of hyperparameters differs among the methods as described in Chapter 2. Regression models and 3. Hyperparameters. The hyperparameter values must be input in the order in Table 2-2.** For example, when the regression method is BayesB, typing

```
$./vigor   sample.pheno.txt   sample.geno.txt   BayesB   5 1 0.01
```

denotes that ν = 5, S^2 = 1, and κ = 0.01.

Options that take arguments

-a Covariate file name. A path can be given. Maximum length is 400.

-c Implement ***Cross-validation***. Allowed arguments are −1, −9 and *n* (*n* > 1).

- -c −1 : Leave-one-out validation
- -c −9 : CV defined by a given partition file
- -c *n* : *n*-fold CV

-p Partition file name. A path can be given. Maximum length is 400. Used when option -c is −9.

-u Fold number of hyperparameter tuning by CV. Used when executing ***Model fitting after hyperparameter tuning*** or ***Cross-validation*** with multiple hyperparameter sets. Default is 5.

-n Number of permutations, specified when executing ***Model fitting***. Default is 0 (no permutations).

-k Analyzed trait (column) of the Phenotype file. When k = 0, all traits are analyzed in turn. Default is 1.

-s Convergence threshold. Larger values denote stricter thresholds. See **Appendix A-3** for the convergence criterion. Default is 2 + log10(*P*), where *P* is the number of markers.

-i Maximum number of iterations. Default is 1000.

-v Additional hyperparameter value sets. For example, when the regression model is EBL,

-v 0.1 0.1 1 0.5

indicates that φ = 0.1, ω = 0.1, ψ = 1, and θ = 0.5 (see Table 2-2). Multiple hyperparameter value sets can be specified by repeated use of this option. The maximum number of hyperparameter value sets is 1000.

Other options

-t Execute ***Model fitting after hyperparameter tuning***. For a single hyperparameter value set, ***Model fitting*** is executed.

-o Use the *P* (number of markers) × *N* (number of individuals) matrix as the genotype file. Recommended when the *N* × *P* matrix generates a buffer error.

-q Quiet.

-r Randomize initial values. The default initial values are described in **Appendix A-3**.

-h Help.

4-3. Progress of run

Vigor outputs the run progress to the standard output. This output can be masked by the -q option.

Information	Descriptions
==Progress *t*/*T* trait … set … fold …==	*T* and t indicate the total number of runs and current run number, respectively. The trait name and hyperparameter set used in the current run are displayed. During a CV, the current fold is also displayed.
−−Progress *t*/*T* trait … fold … set … fold2…−−	Displayed when tuning the hyperparameters. When executing ***Model fitting after hyperparameter tuning***, “fold” is always 1. Here, “fold2” indicates the current fold number of the CV during the tuning.
Progress *t*/*T* trait … set … permutation *n*/*Np*	Diplayed when conducting permutation tests. Here, *n* and *Np* indicates the current run number and total number of permutations, respectively.
Re2: … Conv: …	Every 100 iterations, the residual variance (*Re2*) and metric for convergence assessment (*Conv*) are displayed. *Re2* is displayed in the standardized scale.

4-4. Error messages

Vigor outputs error messages to the standard error output.

Messages	Descriptions
Buffer overflow	The maximum length of rows is 1.0 Mb (32^4), corresponding to about 500,000 markers when genotypes are coded as integers. A buffer overflow error occurs when the length exceeds this maximum. To correct buffer overflow, use the matrix of *P* (number of markers) × *N* (number of individuals) as the genotype file, and select option -o.
Cannot open …	The file specified cannot be opened. Please check the file name or directory.
c should be > 0	When using SSVS or MIX, c should be > 0.
Hyperparameters should be positive	Hyperparameter values input to BL or EBL are negative. Please input positive values.
Incorrect dosage file format	Dosage file format is incorrect. Please check the file format in the Beagle manual.
Incorrect ped file format	PED file format is incorrect. Please check the file format in the PLINK and VIGoR manual (see **Subsection 4-1-5. PED file**). Note that the first six columns are mandatory and that missing alleles (0) are not allowed.
Kappa should be 0<Kappa<=1	of When using wBSR, BayesB, BayesC, SSVS, or MIX, ensure that $0 < \kappa \leq 1$.
K is larger than the number of traits in the phenotype file	The argument of option -k is larger than the number of traits.
Misspecification in -c	The argument of option -c is −9, −1, or *n* (> 1).
Misspecification in -u	The argument of option -u should be > 1.
Misspecification in -n	The argument of option -n should be ≥ 0.
Misspecification in -k	The argument of option -k should be ≥ 0.
Misspecification in -i	The argument of option -i should be > 0.
Negative value in Digamma	Digamma function is evaluated in the BayesB, BayesC, and MIX models (**Appendix A-3**). Please check all arguments, options, and input files and re-run the program.

continued

Number of arguments or method specification	Either the number of arguments is incorrect or the regression method is incorrectly abbreviated.
Number of elements in the … file	The number of elements in the file is inconsistent with the expected number. This error arises when row lengths (i.e., number of individuals) in the input files are different. This error also arises when length of each row in an input file is different from each other.
Nu should be > 2	v (Nu) of wBSR, BayesB, BayesC, SSVS, and MIX should be > 2.
S2 should be >= 0	S^2 of wBSR, BayesB, BayesC, SSVS, and MIX should be ≥ 0
Specify Partition file	Although -c is −9, the partition file is not given by the –p option.
The number of individuals differs between the genotype and the phenotype files	The number of individuals differs between these files. Note also that the order of individuals should be the same in both files.
The number of individuals differs between the covariate and the phenotype/genotype files.	The number of individuals differs between these files. Note also that the order of individuals should be the same in both files.

4-5. Output files

Vigor has three output file formats; fitting, crossvalidation, and partition. Fitting files are created during execution of ***Model fitting*** or ***Model fitting after hyperparameter tuning***. Crossvalidation files are created while executing ***Cross-validation***. Partition files are output by *vigor* as the result of random individual partitions in CV.

4-5-1. Fitting file (.fitting)

The file name is

*Method_ Traitname*_set*X*.fitting

where *Method* is the selected regression method (BL, EBL, wBSR, BayesB, BayesC, SSVS, or MIX), *Traitname* is the name of the analyzed trait extracted from the phenotype file, and *X* is the set number of hyperparameters. Different regression methods output different information. Below we demonstrate the file format of each method. Blue-font terms are the outputs of hyperparameter tuning. Green-font terms are the outputs of permutation tests.

4-5-1-1. BL

```
====Markers====================================================
Effect          SD              Tau2            Lambda2         Sig(SD)         Sig(Perm)
====Fitted values==============================================
Yhat            BV
====Covariates=================================================
Effect          SD
====logP(Y)====================================================
====Iterations=================================================
Iteration       ResidualVariance
====MSE under each hyperparameter set==========================
Set             Phi             Omega           MSE
Set X was chosen
====Significance level=========================================
1%
5%
====Parameters=================================================
====Version====================================================
```

==Markers==

Effect

Posterior means of marker effects. Markers are sorted by their order in the genotype file.

SD

Posterior standard deviations of marker effects.

Tau2

Posterior means of τ_p^2.

Lambda2

Posterior means of λ^2. Although this parameter is common to all markers, its estimated value is output for each marker. **Reported in the standardized scale.**

Sig(SD)

Significance of marker effects, judged by Effect and SD. An output of “1” or “5” indicates that the marker effect significantly deviates from 0 (that is, P < 0.01 or P < 0.05 respectively). An insignificant effect returns “0”. A marker effect is judged as significant if the interval [Effect − f(P) × SD, Effect + f(P) × SD], where f(0.05) = 1.96 and f(0.01) = 2.58 excludes 0.

Sig(Perm)

Significance of marker effects judged from permutation tests. An output of “1” or “5” indicates that the marker effect significantly deviates from 0 (that is, P < 0.01 or P < 0.05, respectively). An insignificant effect returns 0. An output of −9 indicates that permutation tests are not conducted.

==Fitted values==

Yhat

Summation over covariates and marker effects.

(i.e., $\sum_{f=1}^{F} z_{if}\alpha_f + \sum_{p=1}^{P} x_{ip}\beta_p$ for individual *i*).

BV

Summation over marker effects (i.e., $\sum_{p=1}^{P} x_{ip}\beta_p$ for individual *i*).

==Covariates==

Effect

Posterior means of covariate effects (α_j). The intercept effect is output even if no covariate file is provided. When a covariate file is input, the covariate effects are output in the order of their appearance in the covariate file.

SD

Posterior standard deviations of covariate effects.

==logP(Y)==

Marginal log likelihood of data approximated by VB (see **Appendix A-3**).

==Iterations==

Iteration

Iteration number.

ResidualVariance

Residual variance in each iteration. **Reported in the standardized scale.**

==MSE under each hyperparameter set==

Set

Hyperparameter set number.

Phi

φ value of the set.

Omega

ω value of the set.

MSE

MSE obtained in the CV during hyperparameter tuning.

Set X was chosen

X specifies the hyperparameter set number with the smallest MSE. Set X is used in the model fitting.

==Significance level==

1%

1% threshold obtained in the permutation tests.

5 %

5% threshold.

==Parameters==

Parameter values used.

==Version==

Version number of *vigor.*

4-5-1-2. EBL

```
====Markers==============================================================
Effect      SD          Tau2        Delta2      Eta2        Sig(SD)     Sig(Perm)
====Fitted values==========================================================
Yhat        BV
====Covariates============================================================
Effect      SD
====logP(Y)===============================================================
====Iterations=============================================================
Iteration     ResidualVariance
====MSE under each hyperparameter set========================================
Set         Phi         Omega       Psi         Theta       MSE
Set X was chosen
====Significance level=======================================================
1%
5%
====Parameters============================================================
====Version===============================================================
```

==Markers==

Delta2

Posterior means of δ^2. Although this parameter is common to all markers, its estimated value is output for each marker. **Reported in the standardized scale.**

Eta2

Posterior means of ρ_p^2. **Reported in the standardized scale.**

The other terms are explained in the BL captions.

4-5-1-3. wBSR

```
====Markers===========================================================
EffectxGamma   SD(EffectxGamm)   Effect   SD        Sigma2   Gamma   Sig(SD)   Sig(Perm)
====Fitted values=====================================================
Yhat          BV
====Covariates========================================================
Effect        SD
====logP(Y)===========================================================
====Iterations========================================================
Iteration        ResidualVariance
====MSE under each hyperparameter set=================================
Set           Nu           S2           Kappa        Theta        MSE
Set X was chosen
====Significance level================================================
1%
5%
====Parameters========================================================
====Version===========================================================
```

==Markers==

EffectxGamma

Posterior means of $\gamma_p\beta_p$.

SD(EffectxGamma)

Posterior standard deviations of $\gamma_p\beta_p$.

Effect

Posterior means of the marker effect (β_p).

SD

Posterior standard deviations of the marker effect (β_p).

Sigma2

Posterior means of the marker effect variance (σ_p^2).

Gamma

Posterior means of γ_p.

Sig(SD)

Significance of marker effects judged from EffectxGamma and SD(EffectxGamma).

Sig(Perm)

Significance of marker effects judged from permutation tests. Null distributions are generated using EffectxGamma.

The other terms are explained in the BL captions.

4-5-1-4. BayesB

```
====Markers=====================================================
Effect          SD              Rho             Sigma2          Sig(SD)         Sig(Perm)
====Fitted values===============================================
Yhat            BV
====Covariates==================================================
Effect          SD
====logP(Y)=====================================================
====Marker effect variance======================================
Sigma2
====Iterations==================================================
Iteration       ResidualVariance
====MSE under each hyperparameter set===========================
Set             Nu              S2              Kappa           MSE
Set X was chosen
====Significance level==========================================
1%
5%
====Parameters==================================================
====Version=====================================================
```

==Markers==

Effect

Posterior means of β_p, i.e., $E[\beta_p \mid \rho_p = 1, Data]E[\rho_p = 1 \mid Data]$.

Rho

Posterior means of ρ_p.

Sigma2

Posterior means of σ_p^2.

The other terms are explained in the BL captions.

4-5-1-5. BayesC

```
====Markers==================================================
Effect          SD              Rho             Sig(SD)         Sig(Perm)
====Fitted values============================================
Yhat            BV
====Covariates===============================================
Effect          SD
====logP(Y)==================================================
====Marker effect variance===================================
Sigma2
====Iterations===============================================
Iteration       ResidualVariance
====MSE under each hyperparameter set========================
Set             Nu              S2              Kappa           MSE
Set X was chosen
====Significance level=======================================
1%
5%
====Parameters===============================================
====Version==================================================
```

==Markers==

Effect

Posterior means of β_p, i.e., $E[\beta_p \mid \rho_p = 1, Data]E[\rho_p = 1 \mid Data]$.

Rho

Posterior means of ρ_p.

==Marker effect variance==

Posterior mean of σ^2

The other terms are explained in the BL captions.

4-5-1-6. SSVS

```
====Markers=====================================================
Effect          SD              Rho             Sig(SD)         Sig(Perm)
====Fitted values===============================================
Yhat            BV
====Covariates==================================================
Effect          SD
====logP(Y)=====================================================
====Marker effect variance======================================
Sigma2
====Iterations==================================================
Iteration       ResidualVariance
====MSE under each hyperparameter set===========================
Set             Nu              S2              Kappa           MSE
Set X was chosen
====Significance level==========================================
1%
5%
====Parameters==================================================
====Version=====================================================
```

The terms are explained in the captions of BL and BayesC.

4-5-1-7. MIX

```
====Markers====================================================
Effect          SD              Rho             Sig(SD)         Sig(Perm)
====Fitted values==============================================
Yhat            BV
====Covariates=================================================
Effect          SD
====logP(Y)====================================================
====Marker effect variances=====================================
Sigma2A
Sigma2B
====Iterations=================================================
Iteration       ResidualVariance
====MSE under each hyperparameter set==========================
Set             Nu              S2              Kappa           MSE
Set X was chosen
====Significance level=========================================
1%
5%
====Parameters=================================================
====Version====================================================
```

==Marker effect variances==

Posterior means of σ_A^2 and $\sigma_B^{2.}$

The other terms are explained in the captions of BL and BayesC.

4-5-2. Crossvalidation file (.crossvalidation)

The file name is

Method_Traitname.crossvalidation

where *Method* is the selected regression method (BL, EBL, wBSR, BayesB, BayesC, SSVS, or MIX), and *Traitname* is the name of the analyzed trait extracted from the phenotype file. Blue-font terms are the outputs of hyperparameter tuning (i.e., when multiple hyperparameter sets are specified).

4-5-2-1. BL

```
====Predicted values==============================================
Test              Y                 Yhat              BV
====MSE===========================================================
Fold              ChosenSet         Phi               Omega             MSE
====Parameters====================================================
====Version=======================================================
```

==Predicted values==

Test

Row numbers of the predicted individuals in the genotype/phenotype file. When a partition file is given, individuals are sorted by their order in the file; otherwise, the order is randomly determined.

Y

Phenotypic values recorded in the phenotype file (i.e., true values).

Yhat

Values predicted as $\sum_{f=1}^{F} z_{if}\hat{\alpha}_f + \sum_{p=1}^{P} x_{ip}\hat{\beta}_p$ where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated effects.

BV

Values predicted as $\sum_{p=1}^{P} x_{ip}\hat{\beta}_p$.

The prediction accuracy is calculated as the Pearson correlation between Y and Yhat or between Y and BV.

==MSE==

Fold

Fold of the CV.

ChosenSet

Hyperparameter set number chosen by CV during hyperparameter tuning at the fold number.

Phi

φ value of the chosen hyperparameter set.

Omega

ω value of the chosen hyperparameter set.

MSE

MSE obtained in the CV during hyperparameter tuning.

==Parameters==

Parameter values used

==Version==

Version number of VIGoR

4-5-2-2. EBL

```
====Predicted values=====================================================
Test          Y            Yhat         BV
====MSE=================================================================
Fold          ChosenSet    Phi          Omega        Psi          Theta        MSE
====Parameters=========================================================
====Version============================================================
```

The terms are explained in the BL captions.

4-5-2-3. wBSR, BayesB, and BayesC

```
====Predicted values==================================================
Test          Y            Yhat         BV
====MSE==============================================================
Fold          ChosenSet    Nu           S2           Kappa        MSE
====Parameters======================================================
====Version=========================================================
```

The terms are explained in the BL captions.

4-5-2-4. SSVS and MIX

```
====Predicted values========================================================
Test          Y             Yhat          BV
====MSE=================================================================
Fold          ChosenSet     c             Nu            S2            Kappa         MSE
====Parameters===========================================================
====Version=============================================================
```

The terms are explained in the BL captions.

4-5-3. Partition file (.partition)

The file name is

Method_Traitname.partition

where *Method* is the selected regression method (BL, EBL, wBSR, BayesB, BayesC, SSVS, or MIX), and *Traitname* is the name of the analyzed trait extracted from the phenotype file. The file format is described in **Subsection 4-1-4. Partition file (optional)**.

4-6. Examples of usage

This section analyzes the sample data provided with the package. If the PATH variable is set, “./” is not needed.

Ex. 1) ***Model fitting*** using BayesC with three hyperparameter value sets.

```
$./vigor    sample.pheno.txt    sample.geno.txt    BayesC    5 2 0.1    -v 5 2 0.01    -v 5 2 0.001
```

In the first set, $\nu = 5$, $S^2 = 2$, and $\kappa = 0.1$, in the second, $\nu = 5$, $S^2 = 2$, and $\kappa = 0.01$, and in the third, $\nu = 5$, $S^2 = 2$, and $\kappa = 0.001$. *Vigor* sequentially analyses the data by model fitting using these sets, and outputs three files: “BayesC_Height_set1.fitting”, “BayesC_Height_set2.fitting”, and “BayesC_Height_set3.fitting”. The output files are explained in **Subsection 4-5-1. Fitting file**.

Ex. 2) ***Model fitting after hyperparameter tuning*** using BayesB.

```
$./vigor    sample.pheno.txt    sample.geno.txt    BayesB    5 2 0.1    -v 5 2 0.01    -v 5 2 0.001    -t
```

The above command line executes a five-fold CV (the default value of 5 can be changed by specifying the –u option) with each set of hyperparameter values. The model is fitted using the hyperparameter set with the lowest MSE score. If the first set ($\nu = 5$, $S^2 = 2$, and $\kappa = 0.1$) is used, the output file is “BayesB_Height_set1.fitting”.

Ex. 3) Cross-validation of the second trait (weight) included in sample.pheno.txt. The regression method is EBL.

```
$./vigor    sample.pheno.txt    sample.geno.txt    EBL    0.1 0.1 1 0.01    -c 10   -k 2
```

Here, “-c 10” indicates 10-fold cross-validation and “-k 2” indicates that the analyzed trait is the second column of the phenotype file (in this case, the “Weight” trait). The cross-validation outputs two files: “EBL_Weight.crossvalidation” and “EBL_Weight.partition”. The latter file includes the CV partition, which can be used as an input file to another 10-fold CV with the same partition. In this way, we can compare the prediction accuracy among methods. For example, we can input “EBL_Weight.partition” to a 10-fold CV of the BL model as follows:

```
$./vigor    sample.pheno.txt    sample.geno.txt    BL    1 0.01    -c -9   -k 2   -p EBL_Weight.partition
```

Ex. 4) ***Model fitting*** with covariates. All traits are analyzed in sequence. The regression model is SSVS.

```
$./vigor    sample.pheno.txt    sample.geno.txt     SSVS     0.01 4 2 0.01     -k 0     -a sample.covariate.txt
```

With the setting -k 0, all traits included in the file sample.pheno.txt are analyzed. *Vigor* outputs three files: “SSVS_Height_set1.fitting”, “SSVS_Weight_set1.fitting”, and “SSVS_Length_set1.fitting”.

Ex. 5) The input file is a PED file.

```
$./vigor     sample.ped     BayesC     5 2 0.1     -v 5 2 0.01     -v 5 2 0.001
```

Note that *vigor* requires only one PED file, because this file type includes both the phenotypic values and marker genotypes.

Ex. 6) The genotype file is a genotype dosage file created by Beagle.

```
$./vigor     sample.pheno.txt     sample.dose     BL     1 0.1     -k 0
```

5. Command line program *hyperpara*

Hyperpara calculates the values of hyperparameters that influence the inference, based on from several assumptions of the genetic architecture. The equations of the calculation are given in **Appendix B**.

5-1. Input files

Hyperpara requires the genotype file, which is introduced in **Subsection 4-1-2. Genotype file. Note that, for feasible calculation of the allele frequency, the genotypes should be coded as 0 (AA), 1 (AB), and 2 (BB)**. Other acceptable inputs are PED files of PLINK and Beagle dosage files. *Hyperpara* automatically recognizes these files by their extensions (.ped and .dose).

5-2. Arguments and options

Hyperpara requires four arguments: the genotype file name, *Mvar*, the selected regression method, and κ. *Hyperpara* offers several options.

Arguments

Genotype file

Genotype file name. A path can be given. Maximum length is 400.

Mvar

The assumed proportion of phenotypic variance (i.e., variance of response variables) that can be explained by markers. In BL and EBL regression, 0 < *Mvar* < 1.0; and in the other models, 0 < *Mvar* ≤ 1.

Method

Seven regression methods are available, and are listed with their abbreviations below.

■	BL	: Bayesian Lasso
■	EBL	: Extended Bayesian Lasso
■	wBSR	: weighted Bayesian shrinkage regression
■	BayesB	: BayesB
■	BayesC	: BayesC
■	SSVS	: Stochastic search variable selection
■	MIX	: Bayesian mixture model

κ

The assumed proportion of markers with non-zero effects. In MIX and SSVS regression, 0 < κ < 1; in the other methods, 0 < κ ≤ 1.

For further discussion on *Mvar* and κ, see **Chapter 3. Hyperparameters**.

Options that take arguments

-k Additional κ values. Multiple κ values can be specified by repeatedly declaring this option.

-a Specifies the A value. In SSVS and MIX regression, A represents the proportion of *Mvar* that can be explained by markers assigned to the normal prior distribution with the larger variance. The default is 0.9. A should satisfy $0 < A < 1$.

-f Inbreeding coefficient. Enter 1 for inbred species. The default is 0.

-b φ value of BL. Default is 1.

-p φ value of EBL. Default is 0.1.

-g ω value of EBL. Default is 0.1

-s ψ value of EBL. Default is 1.

-n ν (Nu) value, used in wBSR, BayesB, BayesC, SSVS, and MIX regression. The default is 5.

By repeated use of the options -k, -a, -b, -p, -g, -s, and –n, the user can construct multiple combinations (sets) of hyperparameter values.

Options

-t Treats variables in the genotype file as general variables. This option is recommended when the predictor variables are not marker genotypes. See **Appendix B**.

-o Use the P (number of markers) × N (number of individuals) matrix as the genotype file. Recommended when the N c P matrix invokes a buffer error.

-h Help.

5-3. Error messages

hyperpara outputs error messages to the standard error output.

Messages	Descriptions
A should be 0<A<1	*A* should satisfy 0 < *A* < 1.
Buffer overflow	The maximum row length is 1.0 Mb (32^4), corresponding to about 500,000 markers when the genotypes are coded as integers. A buffer overflow error occurs when the row length exceeds this limit. To correct buffer overflow, use *P* (number of markers) × *N* (number of individuals) matrix as the genotype file, and set the -o option.
Cannot open …	The file specified cannot be opened. Please check the file name or directory.
Genotypes should be coded as 0(AA), 1(AB), and 2(BB). Doubles between 0 and 2 are also allowed	In *hyperpara*, marker genotypes should be coded as 0, 1, and 2. Doubles (imputed genotypes) between 0 and 2 are allowed.
Inbreeding coefficient should be 0<=f<=1	Inbreeding coefficient should satisfy 0 ≤ *f* ≤ 1
Incorrect dosage file format	Dosage file format is incorrect. Please check the file format in the Beagle manual.
Incorrect ped file format	PED file format is incorrect. Please check the file format in the PLINK and VIGoR manual (see **Subsection 4-1-5. PED file**). Note that the first six columns are mandatory and that missing alleles (0) are not allowed.
Kappa should be 0<Kappa<1	In SSVS and MIX regression, ensure that 0 < κ < 1.
Kappa should be 0<Kappa<=1	For methods other than SSVS and MIX, ensure that 0 < κ ≤ 1.
Number of arguments or method specification	Either the number of arguments is incorrect or the regression method has been incorrectly abbreviated.
Number of elements in the genotype file	The number of elements in the genotype file is not the expected number. This error probably results from inconsistent row lengths in the file.
Nu should be positive	In wBSR, BayesB, BayesC, SSVS, and MIX regression, ν should be positive.
Omega of EBL should be positive	ω should be positive.
Phi of EBL should be positive	φ should be positive
Phi of BL should be positive	φ should be positive.
Psi of EBL should be positive	ψ should be positive.

5-4. Examples of usage

If the PATH variable is set, "./" is not needed.

Ex. 1) Calculate the hyperparameter values of BL with the following settings: *Mvar* = 0.5; *κ* = 0.01.

```
$./hyperpara    sample.geno.txt    0.5    BL    0.01
```

The results are given in the standard output.

```
Genotype file     : sample.geno.txt
Mvar              : 0.500000
Method            : BL
Kappa             : 0.010000
Phi               : 1.000000
Inbreeding coef. : 0.000000
# individuals     : 100
# markers         : 1000

Hyperparameters
Phi    Omega    (Kappa)
1.000000    0.119573    (0.010000)
```

The last row displays the hyperparameter values. φ is set to 1 by default. ω is calculated as 0.119573. The assumed κ value is given in the parenthesis. The φ value is changed by specifying the -b option.

```
$./hyperpara    sample.geno.txt    0.5    BL    0.01    -b 5
……
Phi    Omega    (Kappa)
5.000000    0.597865    (0.010000)
```

Information on input files and arguments is omitted in this example.

Ex. 2) Calculate multiple hyperparameter value sets of BayesC with the follosing settings: *Mvar* = 0.5; *κ* = 0.01, 0.1, and 1.

```
$./hyperpara    sample.geno.txt    0.5    BayesC    0.01    -k 0.1    -k 1
……
Nu   S2   Kappa
5.000000   0.071744   0.010000
5.000000   0.007174   0.100000
5.000000   0.000717   1.000000
```

Note the multiple use of the -k option.

Ex. 3) Calculate multiple hyperparameter value sets of SSVS with the following settings; *Mvar* = 0.5; *κ* = 0.01 and 0.1; *A* = 0.9 and 0.99.

```
$./hyperpara    sample.geno.txt    0.5    SSVS    0.01   -k 0.1    -a 0.9    -a 0.99
……
c   Nu   S2   Kappa   (A)
0.001122   5.000000   0.064569   0.010000   (0.900000)
0.012346   5.000000   0.006457   0.100000   (0.900000)
0.000102   5.000000   0.071026   0.010000   (0.990000)
0.001122   5.000000   0.007103   0.100000   (0.990000)
```

Note the multiple use of the -a option. This command line returns all combinations of the given hyperparameter values and assumptions. The assumed *A* values are also displayed in parentheses.

Ex. 4) Calculate the hyperparameter values of BayesB with the following settings: *Mvar* = 0.5; *κ* = 0.01. The inbred lines are analyzed.

```
$./hyperpara    sample.geno.txt    0.5    BayesB    0.01    -f 1
……
Nu S2 Kappa
5.000000   0.035872   0.010000
```

Ex. 5) Calculate the hyperparameter values of EBL with the following settings: *Mvar* = 0.5; *κ* = 0.01. Consider the marker genotypes as general variables.

```
$./hyperpara    sample.geno.txt    0.5    EBL    0.01    -t
……
Phi Omega Psi Theta (Kappa)
0.100000   0.100000   1.000000   0.119482   (0.010000)
```

6. R function *vigor*

The manual of the R function *vigor* is also provided as R documentation. To view this manual, type ?vigor on the R console. The usage of *vigor* is

```
    vigor (Pheno, Geno, Method = c("BL", "EBL", "wBSR", "BayesB", "BayesC", "SSVS", "MIX"),
          Hyperparameters, Function = "fitting", Nfold = 10, CVFoldTuning = 5,
          Partition=NULL, Covariates = "Intercept", Threshold = 2+log10(ncol(Geno)),
          Maxiterations=1000, RandomIni=FALSE, Printinfo=TRUE)
```

The first four arguments are mandatory; the remaining arguments are optional.

6-1. Mandatory arguments

The four mandatory arguments of the R function *vigor* are *Pheno*, *Geno*, *Method*, and *Hyperparameters*.

Arguments

Pheno

Pheno is a vector of phenotypic values (response variables). Its length is the number of individuals (*N*). Missing values (coded as NA) are allowed.

Geno

Geno is an *N* (number of individuals) × *P* (number of markers) matrix of marker genotypes. Marker genotypes are encoded by single numerals, for example, −1 (AA), 0 (AB), and 1 (BB), or 0 (AA), 1 (AB), or 2 (BB). Doubles (e.g., 0.8) are also allowed. **The number and ordering of individuals should be identical in *Pheno* and *Geno*.**

Method

The available regression methods, along with their abbreviations, are listed below. Because *Method* is given in string format, it requires a double-quotation (e.g., “BL”).

■	“BL”	: Bayesian Lasso
■	“EBL”	: Extended Bayesian Lasso
■	“wBSR”	: weighted Bayesian shrinkage regression
■	“BayesB”	: BayesB
■	“BayesC”	: BayesC
■	“SSVS”	: Stochastic search variable selection
■	“MIX”	: Bayesian mixture model

Hyperparameters

Hyperparameter values are required by the regression method. *Hyperparameters* is a vector containing a single set (combination) of hyperparameter values, or a matrix of multiple hyperparameter sets. For example, when a single hyperparameters set is input to BayesB, we can specify *Hyperparameters* as a vector V:

```
> V <- c (5,   1,   0.01)
> V
[1]   5.00   1.00   0.01
```

This set includes the values v = 5, S^2 = 1, and κ = 0.01. To specify multiple sets, we can declare a matrix *M*;

```
> M <- matrix (c (5,   1,   0.01,   5,   1,   0.1), nc = 3, byrow = TRUE )
> M
       [ ,1]   [ ,2]   [ ,3]
[1, ]     5      1   0.01
[2, ]     5      1   0.10
```

In the first set (row 1), v = 5, S^2 = 1, and κ = 0.01; in the second set (row 2), v = 5, S^2 = 1, and κ = 0.1. When *Hyperparameters* is a matrix (i.e., includes multiple hyperparameter sets), but the *Function* is “fitting”, only the first set is used. **Because the number of hyperparameters differs among methods (see Chapters 2. Regression models and 3. Hyperparameters), the lengths of the vectors or the numbers of matrix columns differ among methods. Hyperparameter values should be ordered in the vectors or matrix rows as indicated in Table 2-2.**

6-2. Optional arguments

Arguments

Function

Specifies the functions of *vigor* illustrated in Fig. 1-1.

- “fitting” : model fitting
- “tuning” : model fitting after hyperparameter tuning
- “cv” : cross-validation

The default function is “fitting”.

Nfold

Fold number of CV. This argument is used with the “cv” function.

- n (n>1) : n-fold cross-validation with randomly partitioned individuals
- −1 : leave-one-out CV
- −9 : Execute CV for a specified *Partition* argument (see below).

The default setting of *Nfold* is 10.

CVFoldTuning

The cross-validation for tuning the hyperparameters requires an integer fold number. The *CVFoldTuning* argument is used with the “tuning” and “cv” functions. The default is 5.

Partition

The *Partition* matrix defines the partitioning of individuals in CV and is used in the “cv” function with *Nfold* set to −9. Similar to the partition file of the CLP *vigor*, ***Partition* specifies the individuals to be predicted (i.e., the individuals that are not used for training) at each fold**.

Ex. 1). The following matrix is the *Partition* matrix of 19 individuals in a five-fold CV.

16	5	17	13	9
12	18	3	14	6
8	7	11	15	19
1	10	2	4	−9

The matrix columns specify the test individuals at each fold. The elements correspond to the row numbers of *Geno* and the vector indices of *Pheno*. In this example, individuals 16, 12, 8, and 1 are excluded from training and predicted at the first fold. Individuals 5, 18, 7, and 10 are excluded and predicted at the second fold. Spaces in the matrix are filled with “−9” (note the missing fourth individual at the fifth fold).

Ex. 2). *Partition* can be employed in random sampling, which may sample some individuals more than once. For example, consider 19 individuals and split five times, with four individuals tested at each split. The corresponding *Partiti*on matrix is

18	3	11	16	13
17	8	13	13	18
7	15	14	19	7
1	13	12	7	2

Herein, individuals 18, 13, and 7 are repeatedly selected for testing.

Covariates

Specifies covariates other than marker genotypes (i.e., the z_{ij} terms in the linear regression equation in **Chapter 2. Regression methods**). The covariates can be a string “Intercept” or an N (number of individuals) × F (number of covariates) matrix.

- “Intercept” : Only the intercept is included in the model (the default)
- $N \times F$ matrix : The covariates are the matrix elements. Both integers and doubles are allowed. **When a matrix is given, *vigor* does not add the intercept, but instead regards the first column of the matrix as the intercept.**

Ex. 1). The following is the covariate file of three individuals. The first column is the intercept. The second and third columns are the covariates. The total number of covariates is three.

1	0.2	11
1	0.4	9
1	1.2	18

Ex. 2). Again consider three individuals. Suppose that the first, second, and third individuals are respectively cultivated in fields “A”, “B”, and “C”. Three field effects are represented in the following covariate matrix.

1	0	0
1	1	0
1	0	1

Herein, the second and third covariates (columns) indicate the relative effects of fields “B” and “C” on field “A”. The intercept (first column) can be regarded as the mean of field “A”. The number of covariates is three.

Ex. 3). Consider a matrix whose covariates are the probabilities that individuals belong to sub-populations (a so-called Q matrix). In this case, the Q matrix can be used as the covariates without the intercept. For example, a Q matrix of four sub-populations can be expressed as the following covariate file.

0.6	0.1	0.1	0.2
0.05	0.45	0.5	0.0
0.2	0.6	0.1	0.1

Here, the first individual belongs to sub-populations one, two, three, and four with probabilities of 0.6, 0.1, 0.1, and 0.2, respectively. The number of covariates is four.

Threshold

This variable is the convergence threshold. Larger values indicate stricter thresholds. The convergence criterion is given in **Appendix A-3**. The default is 2 + log10(*P*) where *P* is the number of markers.

Maxiterations

Maximum number of iterations. Default is 1000.

RandomIni

If TRUE, the initial values are randomized. The default is FALSE. The default initial values are given in **Appendix A-3**.

Printinfo

If TRUE (the default condition), the run information is printed to the console.

6-3. Error messages

Messages	Descriptions
Check the length of Hyperparameters	The length of the *Hyperparameters* vector dose not match the number of hyperparameters required by the regression method.
Check the number of columns of Hyperparameters	The number of columns in the *Hyperparameters* matrix dose not match the number of hyperparameters required by the regression method.
Covariate specification error	*Covariates* should be “Intercept” or a numerical matrix.
Function specification error	Accepted *Function* inputs are “fitting”, “tuning”, and “cv”.
Hyperparameters should be positive	Hyperparameter values of BL and EBL should be positive.
is.matrix(Covariates) is not TRUE	*Covariates* should be “Intercept” or a numerical matrix.
is.matrix (Geno) is not TRUE	*Geno* should be a matrix.
is.matrix(Hyperparameters)|is.vector(Hyperparameters) is not TRUE	*Hyperparameters* should be a vector or matrix.
is.vector (Pheno) is not TRUE	*Pheno* should be a vector.
Kappa should be 0<Kappa<=1	Ensure that $0 < \kappa \leq 1$.
Maxiterations>0 is not TRUE	*Maxiterations* should be a positive integer.
Method specification error	Accepted *Method* inputs are BL, EBL, wBSR, BayesB, BayesC, SSVS, and MIX.
NA in Covariates is not allowed	The *Covariates* matrix cannot contain missing values (NA).
NA in Geno is not allowed	The *Geno* matrix cannot contain missing value (NA).
Nfold specification error	Accepted values of *Nfold* are −1, −9, and n ($n > 1$).
nrow(Covariates)==length(Pheno) is not TRUE	The number of rows (individuals) in *Covariates* does not match the number of elements (individuals) in *Pheno*. Note that the individuals in both objects should also have the same ordering.

continued

nrow(Geno)==length(Pheno) is not TRUE	The number of rows (individuals) of *Geno* does not match the number of elements (individuals) of *Pheno*. Note that the individuals in both objects should also have the same ordering.
Nu should be >0	v (Nu) should be >0.
S2 should be >= 0	S^2 should be >0.
Partition matrix error	The *Partition* matrix should not contain strings, integers larger than *N* (number of individuals), or 0.
Partition should be specified when Nfold==-9	Although *Nfold* = −9, *Partition* is not specified.

6-4. Output lists

Vigor has two output list formats one for ***Model fitting*** and ***Model fitting after hyperparameter tuning***, the other for ***cross-validation***. The parameters are defined in Table 2-1 and **Appendix A**.

6-4-1. The output list of “fitting” or “tuning”

The outputs of hyperparameter tuning are highlighted in blue font.

$LB

Lower bound of the marginal log likelihood of data (**Y**). The lower bound is defined in **Appendix A-2**.

$ResidualVar

Residual variances ($1/\tau_0^2$) during iterations. **Reported in the standardized scale**.

$Beta

Posterior means of marker effects, i.e., E[**β**|**Y**]. The beta of wBSR is E[**β**|**Y**]E[**γ**|**Y**].

$Sd.beta

Posterior uncertainty (standard deviation) of marker effects, i.e., the square root of V[**β**|**y**]. The posterior uncertainty of wBSR is the square root of E[$\boldsymbol{\beta}^2$|**Y**]V[**γ**|**Y**] + V[**β**|**Y**]E[**γ**|**Y**]^2.

$Tau2

Posterior mean of τ_p^2, output by the BL and EBL models.

$Sigma2

Posterior mean of σ^2 (in BayesC and SSVS), σ_p^2 (in wBSR and BayesB), or σ_A^2 and σ_B^2 (in MIX).

$Alpha

Posterior means of covariate effects (E[**α**|**Y**]).

$Sd.alpha

Posterior uncertainty of covariate effects (square root of V[**α**|**Y**]).

$Lambda2

Posterior means of λ^2, output by the BL model. **Reported in the standardized scale**.

$Delta2

Posterior means of δ^2, output by the EBL model. **Reported in the standardized scale.**

$Eta2

Posterior means of η_p^2, output by the EBL model. **Reported in the standardized scale.**

$Gamma

Posterior means of γ_p, output by the wBSR model.

$Rho

Posterior means of ρ_p, output by the BayesB, BayesC, SSVS, and MIX models.

$MSE

A data frame with (2 + number of hyperparameters) columns. This data frame is output when *Function* is set to"tuning". For example, a data frame in the BL model takes the form

Set	Phi	Omega	MSE
1	1	0.0012	2.51
2	1	0.1196	2.08

The column “Set” contains the row numbers of the *Hyperparameters* matrix. “Phi” and “Omega” are the specified hyperparameter values, and “MSE” is that obtained in CV. In this example, the second set yields the lower MSE, so is used in the model fitting.

6-4-2. The output list of “cv”

Terms highlighted in blue are the outputs of hyperparameter tuning (executed when multiple hyperparameter sets are specified).

$Prediction

A data frame with four columns, labeled Test, Y, Yhat, and BV

$Test

Tested samples (the row numbers in *Pheno*/*Geno*/*Covariates*).

$Y

Phenotypic values of the tested samples (true values).

$Yhat

Predicted phenotypic values, obtained by summing the marker and covariate effects (i.e., $\sum_{f=1}^{F} z_{if}\hat{\alpha}_f + \sum_{p=1}^{P} x_{ip}\hat{\beta}_p$ where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated effects).

$BV

Predicted breeding values, obtained by summing the marker effects (i.e., $\sum_{p=1}^{P} x_{ip}\hat{\beta}_p$)

$MSE

A data frame with (3 + number of hyperparameters) columns, output by the hyperparameter tuning procedure. For example, a data frame in the wBSR model takes the form,

Fold	ChosenSet	Nu	S2	Kappa	MSE
1	1	5	0.0007	1.00	1.84
2	1	5	0.0007	1.00	1.85
3	3	5	0.0717	0.01	1.64
4	3	5	0.0717	0.01	1.89
5	2	5	0.0072	0.10	1.49

“Fold” is the fold number of the CV, and “ChosenSet” denotes the chosen hyperparameter set (i.e., set with the least MSE) at each fold. The set numbers correspond to the row numbers of the *Hyperparameters* matrix. “MSE” is the MSE obtained by inserting the specified hyperparameters “Nu”, “S2”, and “Kappa” into CV.

$Partition

A matrix containing the partition of individuals in the CV. This matrix is output when the individuals are randomly partitioned (i.e., when *Nfold* > 1), and can be input to *vigor* as the *Partition* argument (see Example 4 in **Section 6-5. Examples of usage**).

6-5. Examples of usage

This section analyzes some sample data provided with the package. These examples are also illustrated in the R documentation of *vigor*. First, to load the library and read sample data, type

```
>library (vigor)
>data (sampledata)
```

Ex. 1) Analyze the "Height" trait by executing ***Model fitting*** with BL, and make a simple Manhattan plot. The last row displays the markers with significant effects on the trait (P<0.05).

```
>Result <- vigor (Pheno$Height, Geno, "BL", c(1,1), Covariates=Covariates)
>plot (abs (Result$Beta), pch=20) #Manhattan plot
>which ((abs (Result$Beta)-1.96 * Result$Sd.beta)>0) #Significant markers (P<0.05)
```

In this example, the used hyperparameter values are $\varphi = 1$ and $\omega = 1$.

Ex. 2) Execute ***Model fitting*** with BayesC, excluding the covariates.

```
>Result <- vigor (Pheno$Height, Geno, "BayesC", c(5, 1, 0.01))
>plot (abs (Result$Beta), pch=20)
>which ((abs (Result$Beta)-1.96 * Result$Sd.beta)>0)
>print (Result$Alpha) #Intercept is automatically added to the model
```

The hyperparameter values are $v = 5$, $S^2 = 1$, and $\kappa = 0.01$.

Ex. 3) Execute ***Model fitting after hyperparameter tuning*** with BayesB, given a matrix of two hyperparameter sets.

```
>H <- matrix ( c(5, 1, 0.001, 5, 1, 0.01), nc=3, byrow=TRUE )
>print (H)
>Result <- vigor (Pheno$Height, Geno, "BayesB", H, Function="tuning", Covariates = Covariates)
>plot (abs (Result$Beta), pch=20)
>print (Result$MSE)    #the set with the lowest MSE was used.
```

In the first and second set, we set $\kappa = 0.001$ and $\kappa = 0.01$, respectively.

When *Function* is set to “fitting”, only the first set is used in the regression, even when a *Hyperparameter* matrix is given. To repeat analyses for different sets, the following loop can be iterated:

```
>Result <- as.list (numeric(2))
>for (set in 1:2) {
+ Result [[set]] <- vigor (Pheno$Height, Geno, "BayesB", H [set, ], Covariates = Covariates)
>}
```

Ex. 4) Execute a six-fold ***Cross-validation*** using BL. Because two hyperparameter sets (φ = 1 and ω = 0.01, and φ = 1 and ω = 0.1) are given, hyperparameter tuning is automatically performed at each fold (i.e., CV for hyperparameter tuning).

```
>H <- matrix (c(1, 0.01, 1, 0.1), ncol=2, byrow=TRUE)
>Result <- vigor (Pheno$Height, Geno, "BL", H, Function="cv", Nfold=6, Covariates=Covariates)
>plot (Result$Prediction$Y, Result$Prediction$Yhat)    #plot true and predicted values
>cor (Result$Prediction$Y, Result$Prediction$Yhat)    #accuracy
>print (Result$MSE)    #see which the set used at each fold.
>print (Result$Partition)    #see the partition of CV
```

Execute CV with the same partition using BayesC.

```
>H <- matrix (c(5, 1, 0.01, 5, 1, 0.1), nc=3, byrow=TRUE)
>Result2 <- vigor(Pheno$Height, Geno, "BayesC", H, Function="cv", Nfold=−9,
+ Partition=Result$Partition, Covariates=Covariates)
>cor(Result2$Prediction$Y, Result2$Prediction$Yhat)    #accuracy
```

7. R function *hyperpara*

The function *hyperpara* calculates the values of the hyperparameters that influence on the inference, based on several assumptions of the genetic architecture. The calculation equations are given in **Appendix B**. The manual of the R function *hyperpara* is also provided as R documentation. To view this manual, type ?hyperpara on the R console. The usage of *hyperpara* is

```
    hyperpara(Geno, Mvar, Method = c("BL", "EBL", "wBSR", "BayesB", "BayesC", "SSVS", "MIX"),
              Kappa, A = 0.9, Xtype="Geno", f = 0, BL.Phi = 1, EBL.Phi = 0.1, EBL.Omega = 0.1,
              Psi = 1, Nu = 5, Printinfo = FALSE)
```

The first four arguments are mandatory; the remaining arguments are optional.

7-1. Mandatory arguments

The mandatory arguments of the R function *hyperpara* are *Geno*, *Mvar*, *Method*, and *Kappa*.

Arguments

Geno

> *Geno* is an *N* (number of individuals) × *P* (number of markers) matrix of marker genotypes. **Marker genotypes should be encoded as 0 (AA), 1 (AB), or 2 (BB).** Doubles between 0 and 2 (e.g., 0.8) are also allowed.

Mvar

> *Mvar* specifies the assumed proportion of the phenotypic variance (i.e., variance of response variables) that can be explained by the markers. In BL and EBL regression, 0 < *Mvar* <1; in the other methods, 0 < *Mvar* ≤ 1.

Method

> Seven regression methods are available. The methods and their abbreviations are listed below.
>
> - BL : Bayesian Lasso
> - EBL : Extended Bayesian Lasso
> - wBSR : weighted Bayesian shrinkage regression
> - BayesB : BayesB
> - BayesC : BayesC
> - SSVS : Stochastic search variable selection
> - MIX : Bayesian mixture model

Kappa

> The assumed proportion of markers with non-zero effects. In MIX and SSVS regression, 0 < *Kappa* < 1; in the other methods, 0 < *Kappa* ≤ 1. *Kappa* can be a vector.

7-2. Optional arguments

Arguments

A

In SSVS and MIX regression, *A* denotes the proportion of *Mvar* that can be explained by markers assigned to the normal prior distribution with the larger variance. *A* should satisfy 0 < *A* < 1 (its default value is 0.9). *A* can be a vector.

Xtype

Specifies the type of predictor variables. This option is recommended when the predictor variables are not marker genotypes (**Appendix B**).

- "Geno" : marker genotypes
- "Var" : variables other than marker genotypes

The default *Xtype* is "Geno".

f

Inbreeding coefficient of the genotyped population. When analyzing inbred species, *f* is set to 1. Its default value is 0.

BL.Phi

φ value of BL. Default is 1. *BL.Phi* can be a vector.

EBL.Phi

φ value of EBL. Default is 0.1. *EBL.Phi* can be a vector.

EBL.Omega

ω value of EBL. Default is 0.1. *EBL.Omega* can be a vector.

Psi

ψ value of EBL. Default is 1. *Psi* can be a vector.

Nu

ν (Nu) value in the wBSR, BayesB, BayesC, SSVS, and MIX models. Default is 5. *Nu* can be a vector.

Printinfo

If TRUE, the run information is printed to the console, and a histogram of the minor allele frequencies is presented. Default is FALSE.

7-3. Error messages

Messages	Descriptions
A should be 0<A<1	*A* should satisfy $0 < A < 1$
BL.Phi should be >0	φ of BL should be >0
EBL.Omega should be >0	ω of EBL should be >0
EBL.Phi should be >0	φ of EBL should be >0
f should be a scalar (0<=f<=1)	*f* should be a scalar in the range $0 \leq f \leq 1$.
Genotypes should be coded as 0 (homo), 1 (hetero), and 2 (homo)	Genotypes should be encoded as 0, 1, and 2. Doubles between 0 and 2 are also allowed.
Kappa should be 0<Kappa<=1	κ should satisfy $0 < \kappa \leq 1$.
Kappa should be 0<Kappa<1 when SSVS or MIX	In the SSVS and MIX models, κ should satisfy $0 < \kappa < 1$.
Mvar should be 0<Mvar<=1	*Mvar* should satisfy $0 < Mvar \leq 1$
Mvar should be 0<Mvar<1 when BL or EBL	In the BL and EBL models, *Mvar* should satisfy $0 < Mvar < 1$.
NA in Geno is not allowed	*Geno* cannot contain missing value (NA).
Nu should be >2	ν (Nu) should be > 2.
Psi should be >=0	ψ of EBL should be ≥ 0.
S2 should be >= 0	S^2 should be ≥ 0.
Xtype specification error	Accepted *Xtype* string are “Geno” and “Var”.

7-4. Output

The function *hyperpara* outputs a vector when a single hyperparameter set (combination) is created and a matrix when multiple hyperparameter sets are created.

7-5. Examples of usage

This section analyzes some sample data provided with the package. These examples are also illustrated in the R documentation of *vigor*. First, to load the library and read the sample data, type

```
>library (vigor)
>data (sampledata)
```

Ex. 1) Calculate the hyperparameter values of BL with the following settings: *Mvar* = 0.5; *κ* = 0.01.

```
> hyperpara (Geno, 0.5, "BL", 0.01, Printinfo=TRUE)
```

The default φ value is 1. To change the φ value, input the *BL.Phi* argument.

```
> hyperpara (Geno, 0.5, "BL", 0.01, BL.Phi=5)
```

Ex. 2) Calculate the multiple hyperparameter value sets of BayesC with the following settings: *Mvar* = 0.5; *κ* = 0.01, 0.1, and 1.

```
> hyperpara (Geno, 0.5, "BayesC", c(0.01, 0.1, 1))
```

Ex. 3) Use of the output vector as an argument of vigor.

```
> Result <- vigor (Pheno$Height, Geno, "wBSR", hyperpara (Geno, 0.5, "wBSR", 0.01))
```

Ex. 4) Calculate the multiple hyperparameter value sets of SSVS with the following settings: *Mvar* = 0.5; *κ* = 0.01 and 0.1; *A* = 0.9 and 0.99.

```
> hyperpara (Geno, 0.5, "SSVS", c(0.01,0.1), c(0.9,0.99))
```

Ex. 5) Calculate the hyperparameter values of BayesB with the following settings: *Mvar* = 0.5; *κ* = 0.01. The inbred lines are analyzed.

```
> hyperpara (Geno, 0.5, "BayesB", 0.01, f=1)
```

Ex. 6) Calculate the hyperparameter values of EBL with the following settings: *Mvar* = 0.5; *κ* = 0.01. The marker genotypes are treated as general variables.

```
> hyperpara (Geno, 0.5, "EBL", 0.01, Xtype="Var")
```

8. Citation

For VIGoR:

Onogi, A. and H. Iwata, 2015 VIGoR: variational Bayesian inference for genome-wide regression (in prep.)

We hope that users also cite the papers that proposed the regression methods or variational Bayesian algorithms (see also Table 2-3).

Carbonetto, P. and M. Stephens, 2012 Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. 7: 73-108.

George, E. I. and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88: 881-889.

Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Hayashi, T. and H. Iwata, 2010 EM algorithm for Bayesian estimation of genomic breeding values.. BMC Genet. 11:3.

Hayashi, T. and H. Iwata, 2013 A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics 14:34.

Karkkainen, H. P. and M. J. Sillanpaa, 2012a Back to basics for Bayesian model building in genomic selection. Genetics 191:969-987.

Karkkainen, H. P. and M. J. Sillanpaa, 2012b Robustness of Bayesian multilocus association models to cryptic relatedness. Ann. Hum. Genet. 76:510-523.

Li, Z. and M. J. Sillanpaa, 2012 Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190:231-249.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen et al., 2009 The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119-1126.

Mutshinda, C. M. and M. J. Sillanpaa, 2010 Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186:1067-1075.

Onogi, A. 2015 Documents for VIGoR. https://github.com/Onogi/VIGoR.

Park, T. and G. Casella, 2008 The Bayesian lasso. J. Am. Stat. Assoc. 103: 681-686.

Appendix A - Algorithms

A-1 Model structure

The linear regression model assumed in VIGoR is, for individual *i*,

$$y_i = \sum_{j=1}^{F} z_{ij}\alpha_j + \sum_{p=1}^{P} \gamma_p x_{ip}\beta_p + \varepsilon_i ,$$

where y_i is the observed value (i.e., phenotypic value), *F* is the number of covariates other than markers, z_{ij} is the covariate corresponding to the effect α_j, *P* is the number of markers, γ_p is the indicator variable that takes 0 or 1, x_{ip} is the genotype of marker *p*, β_p is the effect of marker *p*, and ε_i is the residual. Likelihood is defined as,

$$N\left(\sum_{j=1}^{F} z_{ij}\alpha_j + \sum_{p=1}^{P} \gamma_p x_{ip}\beta_p , \frac{1}{\tau_0^2}\right),$$

where *N* indicates a normal distribution, and τ_0^2 is the precision of the residuals. For all of the regression methods, the prior distribution of τ_0^2 is $\frac{1}{\tau_0^2}$, and that of α_j is assumed to be proportional to a constant value. Except for wBSR, the indicator variables (γ_p) are fixed to 1. The prior distributions of β_p and γ_p are presented in Table 2-1.

In the development of VIGoR, we modified the VB algorithms for BL and EBL from those in Li and Sillanpaa (2012): in their parameterization, the marker effect is independent from the residual variance, whereas in VIGoR, it is conditional on the residual variance. In our experience, this manipulation slightly accelerated convergence. We also modified the hierarchical structure of MIX: in Luan *et al*. (2009), the two marker effect variances were drawn from the same scaled inverse-chi-squared distribution, whereas, in VIGoR, they were drawn from different priors of which the scale parameters had different magnitudes determined by *c*. This was intended to facilitate the clustering of markers according to their effect sizes. Carbonetto and Stephens (2012) introduced a variational Bayesian algorithm for linear regression with a spike and slab prior which is equivalent to BayesC. The difference between the algorithm for BayesC in VIGoR and that in Carbonetto and Stephens (2012) is that Carbonetto and Stephens used the importance sampling technique to infer the posterior distribution of σ^2, τ_0^2, and κ, whereas we inferred the factorized posteriors of σ^2 and τ_0^2 and used a fixed value for κ.

A-2. Variational Bayesian inference

VB approximates the marginal log likelihood of data (***y***) by maximizing the lower bound of the marginal log likelihood. The lower bound can be written by using Jensen’s inequality as

$$\begin{aligned}\log p(\boldsymbol{y}) &= \log \int q(\boldsymbol{\theta}) \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &\geq \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}\end{aligned}, \qquad \text{(A1)}$$

where *q* denotes any probability distribution of ***θ***. In VB, factorized posterior distributions are used as *q*. That is,

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{P} q_i(\theta_i \mid \boldsymbol{y}),$$

where *P* is the number of parameters. Hereinafter, we denote the factorized posterior distributions as pseudo posterior distributions. The lower bound can be maximized with regard to $q_i(\theta_i \mid \boldsymbol{y})$ by setting

$$q_i(\theta_i \mid \boldsymbol{y}) \propto \exp\left(E_{q_j, j\neq i}\left[\log p(\boldsymbol{y},\boldsymbol{\theta})\right]\right)$$

because

$$\begin{aligned}\int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} &= \int q_i(\theta_i \mid \boldsymbol{y}) \prod_{j\neq i} q_j(\theta_j \mid \boldsymbol{y}) \left[\log p(\boldsymbol{y},\boldsymbol{\theta}) - \log q_i(\theta_i \mid \boldsymbol{y})\right] d\boldsymbol{\theta} \\ &\quad - \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \sum_{j\neq i} \log q_j(\theta_j \mid \boldsymbol{y}) d\boldsymbol{\theta} \\ &= \int q_i(\theta_i \mid \boldsymbol{y}) \left\{ \int \prod_{j\neq i} q_j(\theta_j \mid \boldsymbol{y}) \log p(\boldsymbol{y},\boldsymbol{\theta}) d\theta_{j\neq i} - \log q_i(\theta_i \mid \boldsymbol{y}) \right\} d\theta_i + const. \\ &= -KL\left\{ q_i(\theta_i \mid \boldsymbol{y}) \parallel \exp\left(E_{q_j, j\neq i}\left[\log p(\boldsymbol{y},\boldsymbol{\theta})\right]\right) \right\} + const.\end{aligned},$$

where *KL* indicates the Kullback–Leibler divergence. Here, we use $\int q_j(\theta_j \mid \boldsymbol{y}) d\theta_j = 1$, which holds for any *j*. The maximization of the lower bound is equivalent to the minimization of the *KL* divergence between the posterior and the pseudo posterior distributions of ***θ***. That is,

$$\begin{aligned}\log p(\boldsymbol{y}) &= \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} - \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta} \mid \boldsymbol{y}) \log \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \boldsymbol{y})} d\boldsymbol{\theta} + KL\left[q(\boldsymbol{\theta} \mid \boldsymbol{y}) \parallel p(\boldsymbol{\theta} \mid \boldsymbol{y})\right]\end{aligned}$$

In VB, as well as in Markov chain Monte Carlo, the posterior uncertainty of parameters can be inferred in addition to the posterior means. More detailed descriptions of VB can be found elsewhere, e.g. Bishop (2006) or Murphy (2012).

A-3. Update procedures

In VB, parameters are iteratively updated from initial values until convergence. The initial values of regression coefficients (*α_p* and *β_p*) are 0, and that of τ_0^2 is $\dfrac{100}{V\left[\boldsymbol{y}\right]}$, where *V[**y**]* indicates the phenotypic variance. The iterative update is stopped when $\dfrac{\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\right\|^2}{\left\|\boldsymbol{\theta}^*\right\|^2} < 10^{-m}$, where $\|\ \|$ is the Euclidean norm, ***θ*** is the vector that contains all parameter values at the previous iteration, ***θ**** is the vector consisting of newly updated parameter values at the iteration, and *m* defines the criterion for convergence, which is set to 2 + log10 (*P*) where *P* is the number of markers.

A-3-1. Bayesian lasso (BL)

The joint log posterior distribution is

$$\frac{N}{2}\log\tau_0^2 + \frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F} z_{ij}\alpha_j - \sum_{p=1}^{P} x_{ip}\beta_p\right)^2$$

$$-\log\tau_0^2 + \frac{P}{2}\log\tau_0^2 + \frac{1}{2}\sum_{p=1}^{P}\log\tau_p^2 - \frac{\tau_0^2}{2}\sum_{p=1}^{P}\tau_p^2\beta_p^2 - 2\sum_{p=1}^{P}\log\tau_p^2 - \frac{1}{2}\sum_{p=1}^{P}\frac{\lambda^2}{\tau_p^2} + (\phi-1)\log\lambda^2 - \varpi\lambda^2 + Const.$$

where *Const.* indicates the constant term. In VB, the joint posterior distribution of parameters is factorized into pseudo posterior distributions for each parameter as described above. The pseudo posterior distribution of a parameter is obtained by taking expectations of the joint posterior log likelihood with regard to the pseudo posterior distributions of the remaining parameters. In BL, the pseudo posterior distribution of α_j is a normal distribution with

$$E\left[\alpha_j\right] = \Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N} z_{ij}\left(y_i - \sum_{k\neq j}^{F} E\left[\alpha_k\right] z_{ik} - \sum_{p=1}^{P} E\left[\beta_p\right] x_{ip}\right)$$

and

$$V\left[\alpha_j\right] = \Lambda_j,$$

where $\Lambda_j^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} z_{ij}^2$.

The pseudo posterior distribution of β_p is also a normal distribution with

$$E\left[\beta_p\right] = \mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{k\neq p}^{P} E\left[\beta_k\right] x_{ik}\right)$$

and

$$V\left[\beta_p\right] = \mathrm{H}_p,$$

where $\mathrm{H}_p^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}^2 + E\left[\tau_p^2\right]E\left[\tau_0^2\right]$.

The pseudo posterior distribution of τ_p^2 is an inverse-Gaussian distribution with

$$E\left[\tau_p^2\right] = \mu_p$$

and

$$E\left[\frac{1}{\tau_p^2}\right] = \frac{1}{\mu_p} + \frac{1}{\xi_p},$$

where $\mu_p = \sqrt{\dfrac{E\left[\lambda^2\right]}{E\left[\beta_p^2\right]E\left[\tau_0^2\right]}}$ and $\xi_p = E\left[\lambda^2\right]$. Here we use $E\left[X^r\right] = \mu^r \sum_{s=0}^{r-1} \dfrac{(r-1+s)!}{s!(r-1-s)!}\left(\dfrac{2\xi}{\mu}\right)^{-s}$ and $E\left[X^{-r}\right] = \dfrac{E\left[X^{r+1}\right]}{\mu^{2r+1}}$ when *X* follows an inverse Gaussian distribution with parameters *μ* and *ζ* (Chhikara and Folks 1989).

The pseudo posterior distribution of τ_0^2 is a gamma distribution with

$$E\left[\tau_0^2\right] = \frac{a_1}{b_1},$$

where $a_1 = \frac{1}{2}(N+P)$ and

$$b_1 = \frac{1}{2}\left\{\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right]z_{ij} - \sum_{p=1}^{P} E\left[\beta_p\right]x_{ip}\right)^2 + \sum_{j=1}^{F} V\left[\alpha_j\right]\sum_{i=1}^{N} z_{ij}^2 + \sum_{p=1}^{P} V\left[\beta_p\right]\sum_{i=1}^{N} x_{ip}^2 + \sum_{p=1}^{P} E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right\}$$

The pseudo posterior distribution of λ^2 is also a gamma distribution with

$$E\left[\lambda^2\right] = \frac{a_2}{b_2},$$

where $a_2 = P + \phi$ and $b_2 = \frac{1}{2}\sum_{p=1}^{P} E\left[\frac{1}{\tau_p^2}\right] + \varpi$.

The moments of these pseudo posterior distributions are iteratively updated from the initial values until convergence. The iterative update monotonically increases the lower bound of the marginal log likelihood of data, p(***y***). From Eq. A1, the lower bound of p(***y***) can be written as

$$\int q(\boldsymbol{\theta})\log p(\mathbf{Y},\boldsymbol{\theta})d\boldsymbol{\theta} - \int q(\boldsymbol{\theta})\log q(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

The first term denotes the expectation of the joint log posterior distribution with regard to the pseudo posterior distributions. The second term denotes the expectation of the log pseudo posterior distributions with regard to the pseudo posterior distributions. The lower bound of BL is

$$\sum_{p=1}^{P}\left[-\frac{1}{2}E\left[\delta^2\right]E\left[\frac{1}{\tau_p^2}\right]-\frac{1}{2}\log E\left[\lambda^2\right]+\frac{1}{2}\log V\left[\beta_p\right]-\frac{1}{2}E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right]$$

$$-\varpi E\left[\lambda^2\right]-a_1\log b_1+\log\Gamma\left(a_1\right)-a_2\left(\log b_2-1\right)+\log\Gamma\left(a_2\right)+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right],$$

$$-\frac{N-P-F}{2}\log 2\pi+\phi\log\varpi-\log\Gamma\left(\phi\right)+\frac{2P+F}{2}-P\log 2$$

where Γ( ) indicates the gamma function.

A-3-2. Extended Bayesian lasso (EBL)

The joint log posterior distribution is

$$\frac{N}{2}\log\tau_0^2+\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2$$

$$-\log\tau_0^2+\frac{P}{2}\log\tau_0^2+\frac{1}{2}\sum_{p=1}^{P}\log\tau_p^2-\frac{\tau_0^2}{2}\sum_{p=1}^{P}\tau_p^2\beta_p^2-2\sum_{p=1}^{P}\log\tau_p^2-\frac{1}{2}\sum_{p=1}^{P}\frac{\delta^2\eta_p^2}{\tau_p^2}$$

$$+\left(\phi-1\right)\log\delta^2-\varpi\delta^2+\left(\psi-1\right)\sum_{p=1}^{P}\log\eta_p^2-\theta\sum_{p=1}^{P}\eta_p^2+Const.$$

As in BL, the prior distributions of the marker effects are conditioned by the residual variance.

The pseudo posterior distribution of α_j is a normal distribution with

$$E\left[\alpha_j\right]=\Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)$$

and

$$V\left[\alpha_j\right]=\Lambda_j,$$

where $\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$.

The pseudo posterior distribution of β_p is also a normal distribution with

$$E\left[\beta_p\right]=\mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\beta_k\right]x_{ik}\right)$$

and

$$V\left[\beta_p\right]=\mathrm{H}_p,$$

where $\mathrm{H}_p^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}^2 + E\left[\tau_p^2\right]E\left[\tau_0^2\right]$.

The pseudo posterior distribution of τ_p^2 is an inverse-Gaussian distribution with

$$E\left[\tau_p^2\right] = \sqrt{\frac{E\left[\delta^2\right]E\left[\eta_p^2\right]}{E\left[\beta_p^2\right]E\left[\tau_0^2\right]}}$$

and

$$E\left[\frac{1}{\tau_p^2}\right] = \sqrt{\frac{E\left[\beta_p^2\right]E\left[\tau_0^2\right]}{E\left[\delta^2\right]E\left[\eta_p^2\right]}} + \frac{1}{E\left[\delta^2\right]E\left[\eta_p^2\right]}.$$

The pseudo posterior distribution of τ_0^2 is a gamma distribution with

$$E\left[\tau_0^2\right] = \frac{a_1}{b_1},$$

where $a_1 = \frac{1}{2}(N+P)$ and

$$b_1 = \frac{1}{2}\left\{\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right]z_{ij} - \sum_{p=1}^{P} E\left[\beta_p\right]x_{ip}\right)^2 + \sum_{j=1}^{F} V\left[\alpha_j\right]\sum_{i=1}^{N} z_{ij}^2 + \sum_{p=1}^{P} V\left[\beta_p\right]\sum_{i=1}^{N} x_{ip}^2 + \sum_{p=1}^{P} E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right\}$$

.

The pseudo posterior distributions of δ^2 and η_p^2 are also gamma distributions; the expectations are

$$E\left[\delta^2\right] = \frac{a_2}{b_2},$$

where $a_2 = P + \phi$ and $b_2 = \frac{1}{2}\sum_{p=1}^{P} E\left[\eta_p^2\right]E\left[\frac{1}{\tau_p^2}\right] + \varpi$, and

$$E\left[\eta_p^2\right] = \frac{a_3}{b_{3,p}},$$

where $a_3 = 1 + \psi$ and $b_{3,p} = \frac{1}{2}E\left[\delta^2\right]E\left[\frac{1}{\tau_p^2}\right] + \theta$.

The lower bound of log p(***Y***) is

$$\begin{aligned}&\sum_{p=1}^{P}\left[-\frac{1}{2}E\left[\delta^2\right]E\left[\eta_p^2\right]E\left[\frac{1}{\tau_p^2}\right]-\frac{1}{2}\log\left(E\left[\delta^2\right]E\left[\eta_p^2\right]\right)+\frac{1}{2}\log V\left[\beta_p\right]-\frac{1}{2}E\left[\tau_p^2\right]E\left[\beta_p^2\right]\right]\\&+\sum_{p=1}^{P}\left[-\theta E\left[\eta_p^2\right]-a_3\left(\log b_{3,p}-1\right)+\log\Gamma\left(a_3\right)\right]\\&-\varpi E\left[\delta^2\right]-a_1\log b_1+\log\Gamma\left(a_1\right)-a_2\left(\log b_2-1\right)+\log\Gamma\left(a_2\right)+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]\\&-\frac{N-P-F}{2}\log 2\pi+\phi\log\varpi-\log\Gamma\left(\phi\right)+P\left[\psi\log\theta-\log\Gamma\left(\psi\right)\right]+\frac{2P+F}{2}-P\log 2\end{aligned} \quad .$$

A-3-3. Weighted Bayesian shrinkage regression (wBSR)

The joint log posterior density is

$$\begin{aligned}&\frac{N}{2}\log\tau_0^2-\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}\gamma_p x_{ip}\beta_p\right)^2\\&-\log\tau_0^2-\frac{1}{2}\sum_{p=1}^{P}\log\sigma_p^2-\frac{1}{2}\sum_{p=1}^{P}\frac{\beta_p^2}{\sigma_p^2}+\left(-\frac{\nu}{2}-1\right)\sum_{p=1}^{P}\log\sigma_p^2-\sum_{p=1}^{P}\frac{\nu S^2}{2\sigma_p^2}\ .\\&+\left[\sum_{p=1}^{P}\gamma_p\right]\log\kappa+\left[P-\sum_{p=1}^{P}\gamma_p\right]\log\left(1-\kappa\right)+Const.\end{aligned}$$

The pseudo posterior distribution of α_j is a normal distribution with

$$E\left[\alpha_j\right]=\Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\gamma_p\right]E\left[\beta_p\right]x_{ip}\right)$$

and

$$V\left[\alpha_j\right]=\Lambda_j\,,$$

where $\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$.

The pseudo posterior distribution of β_p is also a normal distribution with

$$E\left[\beta_p\right]=\mathrm{H}_p E\left[\gamma_p\right]E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\gamma_k\right]E\left[\beta_k\right]x_{ik}\right)$$

and

$$V\left[\beta_p\right]=\mathrm{H}_p\,,$$

where $\mathrm{H}_p^{-1}=E\left[\gamma_p^2\right]E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}^2+E\left[\frac{1}{\sigma_p^2}\right]$.

The pseudo posterior distribution of σ_p^2 is a scaled inverse-chi-square distribution with

$$E\left[\sigma_p^2\right]=\frac{\nu_p S_p^2}{\nu_p-2},$$

where $\nu_p=\nu+1$ and $S_p^2=\dfrac{E\left[\beta_p^2\right]+\nu S^2}{\nu+1}$, and

$$E\left[\frac{1}{\sigma_p^2}\right]=\frac{1}{S_p^2}.$$

The pseudo posterior distribution of τ_0^2 is a gamma distribution with

$$E\left[\tau_0^2\right]=\frac{a_1}{b_1},$$

where $a_1=\dfrac{N}{2}$ and

$$\begin{aligned}b_1=&\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{p=1}^{P}E\left[\gamma_p\right]E\left[\beta_p\right]x_{ip}\right)^2\\&+\frac{1}{2}\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\frac{1}{2}\sum_{p=1}^{P}\left(E\left[\gamma_p^2\right]E\left[\beta_p^2\right]-E\left[\gamma_p\right]^2E\left[\beta_p\right]^2\right)\sum_{i=1}^{N}x_{ip}^2\end{aligned}.$$

The posterior distribution of γ_p is a Bernoulli distribution with

$$E\left[\gamma_p\right]=\frac{\kappa\exp(R)}{\kappa\exp(R)+(1-\kappa)\exp\left(R^*\right)},$$

where

$$R=-\frac{E\left[\tau_0^2\right]}{2}\left\{\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E[\gamma_k]E[\beta_k]x_{ik}-E\left[\beta_p\right]x_{ip}\right)^2+W_p+V\left[\beta_p\right]\sum_{i=1}^{N}x_{ip}^2\right\},$$

and

$$R^*=-\frac{E\left[\tau_0^2\right]}{2}\left\{\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E[\gamma_k]E[\beta_k]x_{ik}\right)^2+W_p\right\},$$

where

$$W_p=\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\sum_{k\neq p}^{P}\left(E\left[\gamma_k^2\right]E\left[\beta_k^2\right]-E[\gamma_k]^2E[\beta_k]^2\right)\sum_{i=1}^{N}x_{ik}^2.$$

The lower bound of log $p(\boldsymbol{Y})$ is

$$\sum_{p=1}^{P}\left[-\frac{v_p}{2}\log\frac{v_pS_p^2}{2}+\log\Gamma\left(\frac{v_p}{2}\right)+\frac{1}{2}\log V\left[\beta_p\right]+E\left[\gamma_p\right]\log\frac{\kappa}{E\left[\gamma_p\right]}+(1-\kappa)\log\frac{(1-\kappa)}{\left(1-E\left[\gamma_p\right]\right)}\right]$$
$$-a_1\log b_1+\log\Gamma\left(a_1\right)+P\left[\frac{v}{2}\log\frac{vS^2}{2}-\log\Gamma\left(\frac{v}{2}\right)\right]+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right] \qquad .$$
$$-\frac{N-F}{2}\log 2\pi+\frac{P+F}{2}$$

A-3-4. BayesB

The joint log posterior distribution is

$$\frac{N}{2}\log\tau_0^2-\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2$$
$$-\log\tau_0^2+\sum_{p=1}^{P}\rho_p\left[-\frac{1}{2}\log\sigma_p^2-\frac{\beta_p^2}{2\sigma_p^2}\right]+\left(-\frac{\nu}{2}-1\right)\sum_{p=1}^{P}\log\sigma_p^2-\frac{\nu S^2}{2}\sum_{p=1}^{P}\frac{1}{\sigma_p^2} .$$
$$+\left[\sum_{p=1}^{P}\rho_p\right]\log\kappa+\left[P-\sum_{p=1}^{P}\rho_p\right]\log\left(1-\kappa\right)+Const.$$

The pseudo posterior distribution of α_j is a normal distribution with

$$E\left[\alpha_j\right]=\Lambda_jE\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}\left(y_i-\sum_{k\neq j}^{F}E\left[\alpha_k\right]z_{ik}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)$$

and

$$V\left[\alpha_j\right]=\Lambda_j ,$$

where $\Lambda_j^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}z_{ij}^2$.

For β_p and ρ_p, we consider the joint posterior distribution to be,

$$q\left(\beta_p,\rho_p\right)\propto-\frac{E\left[\tau_0^2\right]}{2}\left[\beta_p^2\sum_{i=1}^{N}x_{ip}^2-2\beta_p\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}z_{ij}E\left[\alpha_j\right]-\sum_{k\neq p}^{P}x_{ik}E\left[\beta_k\right]\right)\right]$$
$$+\rho_p\left[\frac{1}{2}\Phi\left(\frac{\tilde{v}_p}{2}\right)-\frac{1}{2}\log\frac{\tilde{v}_p\tilde{S}_p^2}{2}-\frac{\beta_p^2}{2\tilde{S}_p^2}\right]+\rho_p\log\kappa+\left(1-\rho_p\right)\log\left(1-\kappa\right) \quad , \qquad \text{(A2)}$$

where $\tilde{v}_p=v+E\left[\rho_p\right]$, $\tilde{S}_p^2=\dfrac{vS^2+E\left[\beta_p^2\right]}{\tilde{v}_p^2}$, and Φ indicates the digamma function, that is,

$$\Phi\left(x\right)=\frac{d\Gamma\left(x\right)}{dx}\Gamma\left(x\right)^{-1}=\Gamma'\left(x\right)\Gamma\left(x\right)^{-1} .$$

The digamma function stems from

$$\int q(x)\log x dx = -\Phi\left(\frac{\nu}{2}\right)+\log\frac{\nu S^2}{2}$$, (A3)

where *q*(*x*) indicates the density of $\chi^{-2}\left(\nu, S^2\right)$ and is written as

$$q(x)=\Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}} x^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right).$$

Eq. A3 can be derived as follows: first, differentiate a part of *q(x)* with respect to v:

$$\frac{d}{d\nu}\left[x^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)\right]=x^{-\frac{\nu}{2}-1}\left(-\frac{S^2}{2x}\right)\exp\left(-\frac{\nu S^2}{2x}\right)-\frac{1}{2}x^{-\frac{\nu}{2}-1}\log x\exp\left(-\frac{\nu S^2}{2x}\right).$$

By multiplying the both sides by $D=\Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}$, and differentiating the right side with respect to *x*, we obtain

$$\begin{aligned} D\frac{d}{d\nu}\left[x^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)\right] &= -\frac{S^2}{2}\int\frac{1}{x}Dx^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)dx-\frac{1}{2}\int\log xDx^{-\frac{\nu}{2}-1}\exp\left(-\frac{\nu S^2}{2x}\right)dx \\ &= -\frac{S^2}{2}\int q(x)\frac{1}{x}dx-\frac{1}{2}\int q(x)\log x dx \end{aligned}.$$

Using $\int D^{-1}q(x)dx = D^{-1}$, we obtain

$$D\frac{dD^{-1}}{d\nu}=-\frac{S^2}{2}\int q(x)\frac{1}{x}dx-\frac{1}{2}\int q(x)\log x dx$$

This can be rewritten as

$$\begin{aligned} \int q(x)\log x dx &= -2D\frac{dD^{-1}}{d\nu}-S^2\int q(x)\frac{1}{x}dx \\ &= -2D\frac{dD^{-1}}{d\nu}-1 \end{aligned}$$

Herein, because

$$\begin{aligned} \frac{dD^{-1}}{d\nu} &= \frac{d}{d\nu}\left[\Gamma\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}\right] \\ &= \frac{1}{2}\Gamma'\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}+\Gamma\left(\frac{\nu}{2}\right)\frac{d}{d\nu}\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}} \\ &= \frac{1}{2}\Gamma'\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}-\frac{1}{2}\Gamma\left(\frac{\nu}{2}\right)\left(\log\frac{\nu S^2}{2}+1\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}} \end{aligned},$$

we obtain

$$\begin{aligned}\int q(x)\log x dx &= -2D\frac{dD^{-1}}{d\nu}-1\\ &= -2\Gamma\left(\frac{\nu}{2}\right)^{-1}\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}\left[\frac{1}{2}\Gamma'\left(\frac{\nu}{2}\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}-\frac{1}{2}\Gamma\left(\frac{\nu}{2}\right)\left(\log\frac{\nu S^2}{2}+1\right)\left(\frac{\nu S^2}{2}\right)^{-\frac{\nu}{2}}\right]-1\\ &= -\Gamma\left(\frac{\nu}{2}\right)^{-1}\Gamma'\left(\frac{\nu}{2}\right)+\log\frac{\nu S^2}{2}+1-1\\ &= -\Gamma\left(\frac{\nu}{2}\right)^{-1}\Gamma'\left(\frac{\nu}{2}\right)+\log\frac{\nu S^2}{2}\\ &= -\Phi\left(\frac{\nu}{2}\right)+\log\frac{\nu S^2}{2}\end{aligned}$$

The pseudo posterior distribution of ρ_p can be obtained by integrating out β_p in Eq. A2 as

$$\begin{aligned}q(\rho_p=1) &= \int q(\beta_p,\rho_p=1)d\beta_p\\ &\propto \frac{H_p}{2}\left[E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}z_{ij}E\left[\alpha_j\right]-\sum_{k\neq p}^{P}x_{ik}E\left[\beta_k\right]\right)\right]^2\\ &\quad+\frac{1}{2}\log H_p+\frac{1}{2}\Phi\left(\frac{\tilde{v}_p}{2}\right)-\frac{1}{2}\log\frac{\tilde{v}_p\tilde{S}_p^2}{2}+\log\kappa\\ &\propto F_p+\log\kappa\end{aligned},$$

where $H_p^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}^2+\frac{1}{\tilde{S}_p^2}$. Similarly, we obtain $q(\rho_p=0)\propto\log(1-\kappa)$. Consequently,

$$E\left[\rho_p\right]=\frac{\kappa\exp(F_p)}{\kappa\exp(F_p)+(1-\kappa)}.$$

The pseudo posterior distribution of β_p can be obtained by integrating out ρ_p in Eq. A2 as

$$\begin{aligned}q(\beta_p) &= q(\beta_p,\gamma_p=1)+q(\beta_p,\gamma_p=0)\\ &= q(\beta_p\mid\gamma_p=1)q(\gamma_p=1)+q(\beta_p\mid\gamma_p=0)q(\gamma_p=0)\end{aligned}.$$

Thus, $E\left[\beta_p\right]=E\left[\beta_p\mid\rho_p=1\right]E\left[\rho_p\right]$ and $E\left[\beta_p^2\right]=E\left[\beta_p^2\mid\rho_p=1\right]E\left[\rho_p\right]$. From Eq. A2, $q(\beta_p\mid\gamma_p=1)$ turns out to be a normal distribution with

$$E\left[\beta_p\mid\rho_p=1\right]=\mathrm{H}_pE\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{k\neq p}^{P}E\left[\beta_k\right]x_{ik}\right)$$

and

$$V\left[\beta_p \mid \rho_p = 1\right] = \mathrm{H}_p \, .$$

The second moment can be obtained as

$$E\left[\beta_p^2 \mid \rho_p = 1\right] = V\left[\beta_p \mid \rho_p = 1\right] + E\left[\beta_p \mid \rho_p = 1\right]^2 .$$

The pseudo posterior distribution of σ^2 is a scaled inverse-chi-squared distribution with

$$E\left[\frac{1}{\sigma^2}\right] = \frac{1}{\tilde{S}_p^2} \, .$$

The pseudo posterior of τ_0^2 is a gamma distribution with

$$E\left[\tau_0^2\right] = \frac{a_1}{b_1} \, ,$$

where $a_1 = \dfrac{N}{2}$ and

$$\begin{aligned} b_1 = {} & \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{p=1}^{P} E\left[\beta_p\right] x_{ip}\right)^2 + \frac{1}{2}\sum_{j=1}^{F} V\left[\alpha_j\right]\sum_{i=1}^{N} z_{ij}^2 \\ & + \frac{1}{2}\sum_{p=1}^{P} E\left[\rho_p\right]\left(E\left[\beta_p^2 \mid \rho_p = 1\right] - E\left[\rho_p\right] E\left[\beta_p \mid \rho_p = 1\right]^2\right)\sum_{i=1}^{N} x_{ip}^2 \, . \end{aligned}$$

The lower bound of log p(***Y***) is

$$\begin{aligned} & \sum_{p=1}^{P}\left[\frac{E\left[\rho_p\right]}{2}\log V\left[\beta_p \mid \rho_p = 1\right] + E\left[\rho_p\right]\log\frac{\kappa}{E\left[\rho_p\right]} + (1-\kappa)\log\frac{(1-\kappa)}{\left(1 - E\left[\rho_p\right]\right)}\right] \\ & - a_1 \log b_1 + \log\Gamma\left(a_1\right) + P\left[\frac{v}{2}\log\frac{vS^2}{2} - \log\Gamma\left(\frac{v}{2}\right)\right] - \sum_{p=1}^{P}\left[\frac{\tilde{v}_P}{2}\log\frac{\tilde{v}_P \tilde{S}_P^2}{2} - \log\Gamma\left(\frac{\tilde{v}_p}{2}\right)\right] + \frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right] . \\ & - \frac{N-F}{2}\log 2\pi + \frac{\sum_{p=1}^{P} E\left[\rho_p\right] + F}{2} \end{aligned}$$

A-3-5. BayesC

The joint log posterior distribution is

$$\begin{aligned}&\frac{N}{2}\log\tau_0^2-\frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}z_{ij}\alpha_j-\sum_{p=1}^{P}x_{ip}\beta_p\right)^2\\&-\log\tau_0^2+\sum_{p=1}^{P}\rho_p\left[-\frac{1}{2}\log\sigma^2-\frac{\beta_p^2}{2\sigma^2}\right]+\left(-\frac{\nu}{2}-1\right)\log\sigma^2-\frac{\nu S^2}{2\sigma^2}.\\&+\left[\sum_{p=1}^{P}\rho_p\right]\log\kappa+\left[P-\sum_{p=1}^{P}\rho_p\right]\log\left(1-\kappa\right)+Const.\end{aligned}$$

The pseudo posterior distributions of *α_j* and τ_0^2 are the same as those in BayesB. For β_p and ρ_p, as we do in BayesB, we consider the joint posterior distribution to be,

$$\begin{aligned}q\left(\beta_p,\rho_p\right)\propto&-\frac{E\left[\tau_0^2\right]}{2}\left[\beta_p^2\sum_{i=1}^{N}x_{ip}^2-2\beta_p\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}z_{ij}E\left[\alpha_j\right]-\sum_{k\neq p}^{P}x_{ik}E\left[\beta_k\right]\right)\right]\\&+\rho_p\left[\frac{1}{2}\Phi\left(\frac{\tilde{v}}{2}\right)-\frac{1}{2}\log\frac{\tilde{v}\tilde{S}^2}{2}-\frac{\beta_p^2}{2\tilde{S}^2}\right]+\rho_p\log\kappa+\left(1-\rho_p\right)\log\left(1-\kappa\right)\end{aligned},\quad\text{(A4)}$$

where $\tilde{v}=v+\sum_{j=1}^{P}E\left[\rho_j\right]$ and $\tilde{S}^2=\dfrac{vS^2+\sum_{j=1}^{P}E\left[\beta_j^2\right]}{\tilde{v}}$.

The pseudo posterior distribution of ρ_p can be obtained by integrating out β_p in Eq. A4 as

$$\begin{aligned}q\left(\rho_p=1\right)&=\int q\left(\beta_p,\rho_p=1\right)d\beta_p\\&\propto\frac{H_p}{2}\left[E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}\left(y_i-\sum_{j=1}^{F}z_{ij}E\left[\alpha_j\right]-\sum_{k\neq p}^{P}x_{ik}E\left[\beta_k\right]\right)\right]^2\\&\quad+\frac{1}{2}\log H_p+\frac{1}{2}\Phi\left(\frac{\tilde{v}}{2}\right)-\frac{1}{2}\log\frac{\tilde{v}\tilde{S}^2}{2}+\log\kappa\\&\propto F_p+\log\kappa\end{aligned},$$

where $H_p^{-1}=E\left[\tau_0^2\right]\sum_{i=1}^{N}x_{ip}^2+\frac{1}{\tilde{S}}$. Similarly, we obtain $q\left(\rho_p=0\right)\propto\log\left(1-\kappa\right)$. Consequently,

$$E\left[\rho_p\right]=\frac{\kappa\exp\left(F_p\right)}{\kappa\exp\left(F_p\right)+\left(1-\kappa\right)}.$$

The pseudo posterior distribution of β_p can be obtained by integrating out ρ_p in Eq. A4 as done in BayesB. We obtain

$$E\left[\beta_p\right]=E\left[\beta_p\mid\rho_p=1\right]E\left[\rho_p\right],$$

$$E\left[\beta_p^2\right]=E\left[\beta_p^2\mid\rho_p=1\right]E\left[\rho_p\right],$$

$$E\left[\beta_p \mid \rho_p = 1\right] = \mathrm{H}_p E\left[\tau_0^2\right] \sum_{i=1}^{N} x_{ip} \left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{k \neq p}^{P} E\left[\beta_k\right] x_{ik} \right)$$

and

$$V\left[\beta_p \mid \rho_p = 1\right] = \mathrm{H}_p .$$

The second moment can be obtained as

$$E\left[\beta_p^2 \mid \rho_p = 1\right] = V\left[\beta_p \mid \rho_p = 1\right] + E\left[\beta_p \mid \rho_p = 1\right]^2 .$$

The pseudo posterior distribution of σ^2 is a scaled inverse-chi-squared distribution with

$$E\left[\frac{1}{\sigma^2}\right] = \frac{1}{\tilde{S}^2} .$$

The lower bound of log p(***Y***) is

$$\begin{aligned}
&\sum_{p=1}^{P} \left[\frac{E\left[\rho_p\right]}{2} \log V\left[\beta_p \mid \rho_p = 1\right] + E\left[\rho_p\right] \log \frac{\kappa}{E\left[\rho_p\right]} + \left(1-\kappa\right) \log \frac{\left(1-\kappa\right)}{\left(1 - E\left[\rho_p\right]\right)} \right] \\
&- a_1 \log b_1 + \log \Gamma\left(a_1\right) + \frac{\nu}{2} \log \frac{\nu S^2}{2} - \log \Gamma\left(\frac{\nu}{2}\right) - \frac{\tilde{\nu}}{2} \log \frac{\tilde{\nu} \tilde{S}^2}{2} + \log \Gamma\left(\frac{\tilde{\nu}}{2}\right) + \frac{1}{2} \sum_{j=1}^{F} \log V\left[\alpha_j\right] . \\
&- \frac{N-F}{2} \log 2\pi + \frac{\sum_{p=1}^{P} E\left[\rho_p\right] + F}{2}
\end{aligned}$$

A-3-6. Stochastic search variable selection (SSVS)

The joint posterior distribution is

$$\begin{aligned}
&\frac{N}{2} \log \tau_0^2 - \frac{\tau_0^2}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{F} z_{ij} \alpha_j - \sum_{p=1}^{P} x_{ip} \beta_p \right)^2 \\
&- \log \tau_0^2 - \frac{1}{2} \left[\sum_{p=1}^{P} \rho_p \right] \log \sigma^2 - \frac{1}{2} \sum_{p=1}^{P} \frac{\rho_p \beta_p^2}{\sigma^2} - \left[P - \sum_{p=1}^{P} \rho_p \right] \frac{1}{2} \log c\sigma^2 - \frac{1}{2} \sum_{p=1}^{P} \frac{\left(1-\rho_p\right) \beta_p^2}{c\sigma^2} \\
&+ \left(-\frac{\nu}{2} - 1 \right) \log \sigma^2 - \sum_{p=1}^{P} \frac{\nu S^2}{2\sigma^2} + \left(-\frac{\nu}{2} - 1 \right) \log c\sigma^2 - \sum_{p=1}^{P} \frac{\nu S^2}{2c\sigma^2} \\
&+ \left[\sum_{p=1}^{P} \rho_p \right] \log \kappa + \left[P - \sum_{p=1}^{P} \rho_p \right] \log\left(1-\kappa\right) + Const.
\end{aligned} .$$

The pseudo posterior distribution of α_j is a normal distribution with

$$E\left[\alpha_j\right] = \Lambda_j E\left[\tau_0^2\right] \sum_{i=1}^{N} z_{ij} \left(y_i - \sum_{k \neq j}^{F} E\left[\alpha_k\right] z_{ik} - \sum_{p=1}^{P} E\left[\beta_p\right] x_{ip} \right)$$

and

$$V\left[\alpha_j\right] = \Lambda_j\,,$$

where $\Lambda_j^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} z_{ij}^2$.

The posterior distribution of β_j is a normal distribution with

$$E\left[\beta_p\right] = \mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{k\neq p}^{P} E\left[\beta_k\right] x_{ik}\right)$$

and

$$V\left[\beta_p\right] = \mathrm{H}_p\,,$$

where

$$\mathrm{H}_p^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}^2 + E\left[\frac{1}{\sigma^2}\right]\left[E\left[\rho_p\right]\left(1-\frac{1}{c}\right)+\frac{1}{c}\right].$$

The pseudo posterior distribution of σ^2 is a scaled inverse-chi-squared distribution with the degree of freedom,

$$\tilde{\nu} = \nu + P\,,$$

the scale parameter,

$$\tilde{S}^2 = \frac{\sum_{p=1}^{P} E\left[\rho_p\right]E\left[\beta_p^2\right] + \frac{1}{c}\sum_{p=1}^{P}\left(1-E\left[\rho_p\right]\right)E\left[\beta_p^2\right] + \nu S^2}{\nu + P}\,,$$

and

$$E\left[\frac{1}{\sigma^2}\right] = \frac{1}{\tilde{S}^2}\,.$$

The pseudo posterior distribution of ρ_p is a Bernoulli distribution with

$$E\left[\rho_p\right] = \frac{\kappa\exp\left(-\frac{1}{2}E\left[\frac{1}{\sigma^2}\right]E\left[\beta_p^2\right]\right)}{\kappa\exp\left(-\frac{1}{2}E\left[\frac{1}{\sigma^2}\right]E\left[\beta_p^2\right]\right) + \frac{(1-\kappa)}{\sqrt{c}}\exp\left(-\frac{1}{2c}E\left[\frac{1}{\sigma^2}\right]E\left[\beta_p^2\right]\right)}\,.$$

The pseudo posterior distribution of τ_0^2 is a gamma distribution with

$$E\left[\tau_0^2\right] = \frac{a_1}{b_1}$$

where

$$a_1 = \frac{N}{2}$$

and

$$b_1 = \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{p=1}^{P} E\left[\beta_p\right] x_{ip} \right)^2 + \sum_{j=1}^{F} V\left[\alpha_j\right]\sum_{i=1}^{N} z_{ij}^2 + \sum_{p=1}^{P} V\left[\beta_p\right]\sum_{i=1}^{N} x_{ip}^2 \,.$$

The lower bound of log p(***Y***) is

$$\sum_{p=1}^{P}\left[\frac{1}{2}\log V\left[\beta_p\right] + E\left[\rho_p\right]\log\frac{\kappa}{E\left[\rho_p\right]} + \left(1-\kappa\right)\log\frac{\left(1-\kappa\right)}{\left(1-E\left[\rho_p\right]\right)}\right]$$

$$-a_1\log b_1 + \log\Gamma\left(a_1\right) + \frac{\nu}{2}\log\frac{\nu S^2}{2} - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{\tilde{\nu}}{2}\log\frac{\tilde{\nu}\tilde{S}^2}{2} + \log\Gamma\left(\frac{\tilde{\nu}}{2}\right) + \frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right].$$

$$-\frac{N-F}{2}\log 2\pi + \frac{P+F}{2} - \frac{1}{2}\left(P - \sum_{p=1}^{P} E\left[\rho_p\right]\right)\log c$$

A-3-7. Bayesian mixture regression (MIX)

The joint log posterior distribution is

$$\frac{N}{2}\log\tau_0^2 - \frac{\tau_0^2}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{F} z_{ij}\alpha_j - \sum_{p=1}^{P} x_{ip}\beta_p \right)^2$$

$$-\log\tau_0^2 - \frac{1}{2}\left[\sum_{p=1}^{P}\rho_p\right]\log\sigma_A^2 - \frac{1}{2}\sum_{p=1}^{P}\frac{\rho_p\beta_p^2}{\sigma_A^2} - \left[P - \sum_{p=1}^{P}\rho_p\right]\frac{1}{2}\log\sigma_B^2 - \frac{1}{2}\sum_{p=1}^{P}\frac{\left(1-\rho_p\right)\beta_p^2}{\sigma_B^2}$$

$$+\left(-\frac{\nu}{2}-1\right)\log\sigma_A^2 - \sum_{p=1}^{P}\frac{\nu S^2}{2\sigma_A^2} + \left(-\frac{\nu}{2}-1\right)\log\sigma_B^2 - \sum_{p=1}^{P}\frac{\nu c S^2}{2\sigma_B^2}$$

$$+\left[\sum_{p=1}^{P}\rho_p\right]\log\kappa + \left[P - \sum_{p=1}^{P}\rho_p\right]\log\left(1-\kappa\right) + Const.$$

The pseudo posterior distribution of *α_j* is a normal distribution with

$$E\left[\alpha_j\right] = \Lambda_j E\left[\tau_0^2\right]\sum_{i=1}^{N} z_{ij}\left(y_i - \sum_{k\neq j}^{F} E\left[\alpha_k\right] z_{ik} - \sum_{p=1}^{P} E\left[\beta_p\right] x_{ip} \right)$$

and

$$V\left[\alpha_j\right] = \Lambda_j \,,$$

where $\Lambda_j^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} z_{ij}^2$.

The posterior distribution of β_j is a normal distribution with

$$E\left[\beta_p\right] = \mathrm{H}_p E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}\left(y_i - \sum_{j=1}^{F} E\left[\alpha_j\right] z_{ij} - \sum_{k\neq p}^{P} E\left[\beta_k\right] x_{ik} \right)$$

and

$$V\left[\beta_p\right] = \mathrm{H}_p ,$$

where

$$\mathrm{H}_p^{-1} = E\left[\tau_0^2\right]\sum_{i=1}^{N} x_{ip}^2 + E\left[\frac{1}{\sigma_A^2}\right]E\left[\rho_p\right] + E\left[\frac{1}{\sigma_B^2}\right]\left(1 - E\left[\rho_p\right]\right).$$

The pseudo posterior distribution of σ_A^2 and σ_B^2 are scaled inverse-chi-squared distributions

where, for σ_A^2,

$$\tilde{\nu}_A = \nu + \sum_{p=1}^{P} E\left[\rho_p\right],$$

$$\tilde{S}_A^2 = \frac{\sum_{p=1}^{P} E\left[\rho_p\right] E\left[\beta_p^2\right] + \nu S^2}{\nu + \sum_{p=1}^{P} E\left[\rho_p\right]} ,$$

and

$$E\left[\frac{1}{\sigma_A^2}\right] = \frac{1}{\tilde{S}_A^2} ,$$

and, for σ_B^2,

$$\tilde{\nu}_B = \nu + P - \sum_{p=1}^{P} E\left[\rho_p\right],$$

$$\tilde{S}_B^2 = \frac{\sum_{p=1}^{P}\left(1 - E\left[\rho_p\right]\right) E\left[\beta_p^2\right] + \nu c S^2}{\nu + P - \sum_{p=1}^{P} E\left[\rho_p\right]} ,$$

and

$$E\left[\frac{1}{\sigma_B^2}\right] = \frac{1}{\tilde{S}_B^2} .$$

The pseudo posterior distribution of ρ_p is a Bernoulli distribution with

$$E\left[\rho_p\right] = \frac{\kappa D}{\kappa D + \left(1 - \kappa\right) D^*} ,$$

where

$$D=\sqrt{\frac{2}{\sum_{p=1}^{P}E\left[\rho_p\right]E\left[\beta_p^2\right]+\nu S^2}}\exp\left\{\Phi\left[\frac{1}{2}\left(\nu+\sum_{p=1}^{P}E\left[\rho_p\right]\right)\right]-\frac{1}{2}E\left[\beta_p^2\right]E\left[\frac{1}{\sigma_A^2}\right]\right\}$$

and

$$D^*=\sqrt{\frac{2}{\sum_{p=1}^{P}\left(1-E\left[\rho_p\right]\right)E\left[\beta_p^2\right]+\nu c S^2}}\exp\left\{\Phi\left[\frac{1}{2}\left(\nu+P-\sum_{p=1}^{P}E\left[\rho_p\right]\right)\right]-\frac{1}{2}E\left[\beta_p^2\right]E\left[\frac{1}{\sigma_B^2}\right]\right\}.$$

The pseudo posterior of τ_0^2 is a gamma distribution with

$$E\left[\tau_0^2\right]=\frac{a_1}{b_1}$$

where

$$a_1=\frac{N}{2}$$

and

$$b_1=\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{F}E\left[\alpha_j\right]z_{ij}-\sum_{p=1}^{P}E\left[\beta_p\right]x_{ip}\right)^2+\sum_{j=1}^{F}V\left[\alpha_j\right]\sum_{i=1}^{N}z_{ij}^2+\sum_{p=1}^{P}V\left[\beta_p\right]\sum_{i=1}^{N}x_{ip}^2 .$$

The lower bound of log p(***Y***) is

$$\begin{aligned}
&\sum_{p=1}^{P}\left[\frac{1}{2}\log V\left[\beta_p\right]+E\left[\rho_p\right]\log\frac{\kappa}{E\left[\rho_p\right]}+(1-\kappa)\log\frac{(1-\kappa)}{\left(1-E\left[\rho_p\right]\right)}\right]\\
&-a_1\log b_1+\log\Gamma\left(a_1\right)-\frac{\tilde{\nu}_A}{2}\log\frac{\tilde{\nu}_A\tilde{S}_A^2}{2}+\log\Gamma\left(\frac{\tilde{\nu}_A}{2}\right)-\frac{\tilde{\nu}_B}{2}\log\frac{\tilde{\nu}_B\tilde{S}_B^2}{2}+\log\Gamma\left(\frac{\tilde{\nu}_B}{2}\right).\\
&+\frac{1}{2}\sum_{j=1}^{F}\log V\left[\alpha_j\right]+\nu\log\frac{\nu S^2}{2}-2\log\Gamma\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\log c\\
&-\frac{N-F}{2}\log 2\pi+\frac{P+F}{2}
\end{aligned}$$

Appendix B - Hyperparameter values

Suppose that phenotypic variance (i.e., variance of response variables) is standardized such that the mean and variance is 0 and 1, respectively, as done by VIGoR. When the effect variance of marker *j* is σ_j^2, the proportion of markers with non-zero effects is κ, and linkage equilibrium is assumed, *Mvar* can be represented as

$$Mvar = \kappa \sum_{j=1}^{P} \sigma_j^2 2\left(1+f\right) p_j \left(1-p_j\right), \qquad \text{(B1)}$$

where *f* is the inbreeding coefficient, and p_j is the allele frequency for the marker *j* (Habier *et al*. 2011). For BL, because σ_j^2 can be written as $\frac{1}{\tau_j^2}\left(1-Mvar\right)$ (Table 2-1), we obtain

$$Mvar = \kappa \sum_{j=1}^{P} \frac{1}{\tau_j^2}\left(1-Mvar\right) 2\left(1+f\right) p_j \left(1-p_j\right). \qquad \text{(B2)}$$

The expectations of marker effect variance and λ^2 are $E\left[\frac{1}{\tau_j^2}\right] = \frac{2}{\lambda^2}$ and $\frac{\varphi}{\omega}$, respectively. By plugging these expectations into Eq. B2 and solving the equation with regard to ω, we obtain

$$\varpi = \frac{\varphi}{4\kappa\left(1+f\right)\sum_{j}^{P} p_j\left(1-p_j\right)\left(\frac{1}{Mvar}-1\right)}.$$

Thus, we can determine ω if we give values for φ and κ. For EBL, with a similar approach, we obtain

$$\theta = \frac{\psi\varphi}{4\kappa\varpi\left(1+f\right)\sum_{j}^{P} p_j\left(1-p_j\right)\left(\frac{1}{Mvar}-1\right)}.$$

For wBSR and BayesC, because the expectation of marker effect variance is $\frac{\nu S^2}{\left(\nu-2\right)}$, by plugging this into Eq. B1, we obtain

$$S^2 = \frac{\left(\nu-2\right)Mvar}{\nu\kappa\left(1+f\right)\sum_{j=1}^{P} 2p_j\left(1-p_j\right)}.$$

For SSVS, *Mvar* can be written as

$$Mvar = \sigma^2 \sum_{j\in G_1}\left(1+f\right)2p_j\left(1-p_j\right) + c\sigma^2 \sum_{j\in G_2} 2\left(1+f\right)p_j\left(1-p_j\right),$$

where G_1 and G_2 represents the groups of markers that are assigned to normal distributions with larger and smaller variances respectively. Because the prior expectations of the sizes of G_1 and G_2 are κ*P* and (1−κ)*P*, respectively, we obtain

$$MvarA = \sigma^2 \kappa \left(1+f\right) \sum_{j=1}^{P} 2 p_j \left(1-p_j\right)$$

$$Mvar\left(1-A\right) = c\sigma^2 \left(1-\kappa\right)\left(1+f\right) \sum_{j=1}^{P} 2 p_j \left(1-p_j\right),$$

where *A* represents the proportion of *Mvar* that the markers assigned into the group with larger variance can explain. By solving these equations with regards to *c*, we obtain

$$c = \frac{1-A}{A} \frac{\kappa}{1-\kappa}. \qquad \text{(B3)}$$

By using this *c* value, S^2 can be written as

$$S^2 = \frac{\left(v-2\right) Mvar}{v\left[\kappa + c\left(1-\kappa\right)\right]\left(1+f\right) \sum_{j=1}^{P} 2 p_j \left(1-p_j\right)}. \qquad \text{(B4)}$$

We can derive Eqs. B3 and B4 from the model structure of MIX.

When general variables (i.e., covariates besides marker genotypes) are used as predictor variables, variance due to marker genotypes, i.e., $\sum_{j=1}^{P} 2\left(1+f\right) p_j \left(1-p_j\right)$ is simply replaced by the sum of variance of each given predictor, that is, $\sum_{j=1}^{P} V[x_j]$.

References

Bishop, C. M., 2006 Pattern recognition and machine learning. Springer, New York.

Carbonetto, P. and M. Stephens, 2012 Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. 7: 73-108.

Chhikara, R. S., and J. L. Folks, 1989 The inverse Gaussian distribution. Marcel Dekker, New York.

Dempster A. P., N. M. Laird, D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B Met. 39: 1-38.

George, E. I. and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88: 881-889.

Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Hayashi, T. and H. Iwata, 2010 EM algorithm for Bayesian estimation of genomic breeding values.. BMC Genet. 11: 3.

Hayashi, T. and H. Iwata, 2013 A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics 14: 34.

Karkkainen, H. P. and M. J. Sillanpaa, 2012a Back to basics for Bayesian model building in genomic selection. Genetics 191: 969-987.

Karkkainen, H. P. and M. J. Sillanpaa, 2012b Robustness of Bayesian multilocus association models to cryptic relatedness. Ann. Hum. Genet. 76: 510-523.

Li, Z. and M. J. Sillanpaa, 2012 Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190: 231-249.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen *et al*., 2009 The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183: 1119-1126.

Murphy, K. P., 2012 Machine learning: a probabilistic perspective. MIT press, London.

Mutshinda, C. M. and M. J. Sillanpaa, 2010 Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186: 1067-1075.

Park, T. and G. Casella, 2008 The Bayesian lasso. J. Am. Stat. Assoc. 103: 681-686.