A New Hyperprior Distribution for Bayesian Regression Model with Application in Genomics

Renato Rodrigues Silva [1,*]

1 Institute of Mathematics and Statistics, Federal University of Goiás, Goiânia, Goiás, Brazil

***renato.rrsilva@ufg.br**

Abstract

In the regression analysis, there are situations where the model have more predictor variables than observations of dependent variable, resulting in the problem known as "large p small n". In the last fifteen years, this problem has been received a lot of attention, specially in the genome-wide context. Here we purposed the bayes H model, a bayesian regression model using mixture of two scaled inverse chi square as hyperprior distribution of variance for each regression coefficient. This model is implemented in the R package BayesH.

Introduction

In the regression analysis, there are situations where the model have more predictor variables than observations of dependent variable, resulting in the problem known as "large p small n" [1].

To figure out this problem, there are already exists some methods developed as ridge regression [2], least absolute shrinkage and selection operator (LASSO) regression [3], bridge regression [4], smoothly clipped absolute deviation (SCAD) regression [5] and others. This class of regression models is known in the literature as regression model with penalized likelihood [6]. In the bayesian paradigma, there are also some methods purposed as stochastic search variable selection [7], and Bayesian LASSO [8].

Recently, the "large p small n" problem has been receiving more attention for scientist who works with animal or plant genetics, specifically to apply in genome-wide selection studies [9].

Genome-wide selection is a approach in quantitative genetics to predict the breeding value of the individuals from a testing population based on estimates of the molecular marker effects from training population. The training population is comprised by individuals which were genotyped and phenotyped while in the testing population the individuals are only genotyped [10], [11]. Genotyping refers to obtain the genetic makeup of individuals through some technology [12] and phenotyping is a measure of some economic importance traits as yield, height and etc [11, 13].

With advent of the high throughput genotyping plataforms, nowadays is possible to define a statistical model to identify association between molecular markers and an observed phenotype. In these models, the effects of all markers are estimated simultaneously, capturing even small effects [10, 14].

In the context of genome-wide selection, many animal and plant breeders developed some bayesian regression models to make prediction of complex traits when there are

more covariables than observations of response variable. In the cornerstone publication [10], Bayes A and Bayes B models were presented. In the Bayes A, the scaled-t density were used as prior distribution of marker effects, while in the Bayes B the prior distribution were modeled using a mixture of a point of mass at zero and a scaled-t density. More recently, the use of a mixture of a point of mass at zero and a Gaussian slab were purposed. This model is known in the literature as Bayes Cπ [15, 17–19].

However, there are issues in these models which should have been taken into account. The prior distribution is always influential, therefore its choice is crucial. In this paper we proposed the fit of an Bayesian regression model with mixture of two scaled inverse chi square as hyperprior distribution of variance for each regression coefficient (bayes H model). Until our knowledge, it has never reported before.

An advantage of the model is the flexibility. Depending on values chosen for hyperparameters, is possible to obtain equivalent models to (Bayes Ridge Regression and Bayes A) or even to select variable via Gibbs Sampling in a broad sense. To illustrate to application of the Bayes H model, we analyzed some simulated and real datasets.

Materials and Methods

Simulated Data

The aim these simulations were compare effects of prior distribution in the prediction of complex traits in some situations such as presence or absence of strong linkage disequilibrium or oligogenic or poligenic genetic architecture. The parameter settings of four scenarios generated are presented below. The phenotype were calculated using the equation described by (2).

Table 1. Parameter settings for four simulated scenarios

Scenario	Population *	N. individuals	N. markers	N. QTL's	Distribution of QTL's effect
1	heterogeneous stock mice	250	2500	50	$Gamma(3; 0, 75)$
2	heterogeneous stock mice	250	2500	10	$Gamma(3; 0, 75)$
3	random mating	250	2500	50	$Gamma(3; 0, 75)$
4	random mating	250	2500	10	$Gamma(3; 0, 75)$

* Heterogeneous stock mice were sampled from subsampling of mice dataset , which had already analized by [25] and avalaible in BGLR library of R statistical software [19]; random mating were sampled from Bernoulli distribution with allele frequency equal to 0.5.

For each scenario the predictive performance between Bayes ridge regression and Bayes H model were compared. The table (1) displays the values of hyperparameters used in mixture of the scaled inverse chi-squared distribution. Depending on the values assigned to hyperparameters, differents model can be defined. For example, using the hyperprior A (1), the model will be equivalent to Bayes A model [10]. On the other hand, the use of hyperprior B can be considered as a variable selecion model in a broad sense.

Table 2. Information about hyperparameters used in each component of mixture

Hyperprior	ν_1	s_1	ν_2	s_2
A	5	0.04	5	0.04
B	5	0.5	7	0.002

The values of hyperparameters for prior distribuition for σ^2 were defined as follows: degree of freedom equal to 5 and scale parameter equal to 0.1. Figure (1) shows the

influence of hyperprior distribution for τ^2 in the marginal prior for β_j. Assuming $\sigma^2 = 1$, it is observed the use of hyperprior A (mixture with same scale parameters) the marginal prior resulting for β_j is a t-scaled distribution. On the contrary, when hyperprior B is used the marginal prior obtaining for β_j is a mixture of t-scaled distribution with the same location parameters but different scales parameters, which results a distribution with tails heavier and sharper peaks than t-scaled.

Fig 1. Hyperprior distribution for τ^2 and marginal prior distribution for β_j

Real Data

The real dataset is comprised by 10346 polymorphic markers scored in a mice population of 1814 individuals. The phenotype measured was body mass index (BMI) [28]. The dataset were previously by [25], which further details about about heterogeneous stock mice population can be found. It is important to mention the dataset is avalaible in R package BGLR [19].

Predictions of the complex traits in the mice dataset was done in two step. First of all, a mixed model was fitted to remove the population structure and kinship effect of dataset. In the second step, Bayes H or Bayesian Ridge Regression were fitted to make predictions considering the BLUP's predicted from mixed model as response variable.

The inference of population was based on clustering of the loadings of two top principal components obtained from to genomic relationship matrix. The clustering was done using Gaussian mixture models implemented in the library Mclust of R statistical software [23], [29] e [30]. Several mixture models were fitted and bayesian information criterion was used to select the best model. The candidate models differ each other in relation to covariance matrix of each component of mixture. The general structure of covariance matrix is $\Sigma_k = \lambda D_k A_k D_k^{'}$ where Σ_k is the covariance matrix of kth component of mixture model, D_k is the orthogonal matrix of eigenvectors, A_k is the diagonal matrix whose elements are proportional to the eigenvalues and λ is a scale parameter [30].

Phenotypic Analysis - Mixed Model

Before to predict the molecular breeding value of BMI the phenotypic analysis was done. Phenotypic analysis consisted fitting the mixed model to heterogeneous stock mice population and predict the best linear unbiased predictor (BLUP) for each individual [10, 14].

The mixed model is defined by

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$ is the vector of response variable; $\mathbf{X}$ is the incidence matrix of the fixed effects; $\boldsymbol{\beta}$ is the vector of fixed effects that represents (litter, gender, year and population strucuture); $\mathbf{Z}$ the incidence matrix of random effects; $\mathbf{b}$ the vector of random effects that follows Gaussian distribution with mean 0 and variance σ_b^2 and $\boldsymbol{\epsilon}$ the random error.

Population structure was infered using the results from fitting of Gaussian Mixture Models implemented in the package mclust, a library of R statistical software [23, 29, 30].

Genomic Selection - Statistical Model

The statistical model is defined by

$$y_i = \mu + \sum_{k=1}^{p} \beta_k x_{ik} + \epsilon_i \tag{2}$$

where y_i is the $i-th$ observation of response variable; β_k is the $k-th$ regression coefficient of model; x_{ik} is a explanatory variable for $i-th$ individual and $k-th$ explanatory variable and ϵ_i is the random error for $i-th$ individual that follows $N(0, \sigma^2)$.

Prior Distributions

Considering the fitting of the model, the prior distribution for intercept μ, is defined by

$$\mu \sim N(0, \omega^2)$$

where ω^2 is a hyperparameter. In practice, we used a large number for ω^2 to set up this prior distribution as vague.

The prior distribution for each β_k given a value of τ_k^2 is Gaussian, i.e,

$$\beta_k | \tau_k^2, \sigma^2 \sim N(0, \tau_k^2 \sigma^2) \tag{3}$$

Here, is the novelty of the manuscript. The hyperprior distribution for τ_k^2 is conditioned a latent random variable Z_k. Hence, the hyperprior distribution for τ_k^2 follows the mixture of the two components scaled inverse chi square distribution, i.e

$$\begin{cases} Z_k \sim Bernoulli(1, \pi) \\ \tau_k^2 | Z_k = 1, \sim \text{Scaled-Inv}\chi^2(\nu_1, s_1^2) \\ \tau_k^2 | Z_k = 0, \sim \text{Scaled-Inv}\chi^2(\nu_2, s_2^2) \end{cases} \tag{4}$$

The hyperprior distribution of π is Beta distribution with parameters (α, γ). In pratice, we adopted $\alpha = 1$ and $\gamma = 1$ to obtain a vague hyperprior.

Finally, the prior distribution for σ is

$$\sigma^2 \sim \text{Scaled-Inv}\chi^2(\nu_\sigma, s_\sigma^2) \tag{5}$$

where the hyperparameters ν_σ, s_σ^2 represents the degree of freedom and scale parameters of the scaled inverse chi-square distribution.

Likelihood and Posterior Distribution

The likelihood is defined by

$$Pr(\mathbf{y}, \mu, \boldsymbol{\beta}, \tau^2, \sigma^2) = \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{k=1}^{p} \beta_k x_{ik} \right)^2 \right\} \tag{6}$$

Hence, the joint posterior distribution is given by

$$Pr(\mu, \boldsymbol{\beta}, \tau_k^2, \sigma^2 \mathbf{z}, \pi | \mathbf{y}) \propto Pr(\mathbf{y} | \mu, \boldsymbol{\beta}, \sigma^2) Pr(\mu) \prod_{k=1}^{p} \left[Pr(\beta_k | \tau_k^2) Pr(\tau_k^2 | Z_k) Pr(Z_k | \pi) \right] Pr(\pi) \times Pr(\sigma^2) \tag{7}$$

$$
\begin{aligned}
Pr(\mu, \boldsymbol{\beta}, \tau^2, \sigma^2 \mathbf{z}, \pi | \mathbf{y}) \quad \propto \quad & \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{k=1}^{p} \beta_k x_{ik} \right)^2 \right\} \\
& \exp\left\{ -\frac{1}{2\omega^2} \mu^2 \right\} \prod_{k=1}^{p} \left[\left(\pi \frac{1}{\tau_k^{2(\frac{\nu_1}{2}+1)}} \exp\left\{ -\frac{\nu_1 s_1^2}{2\tau^2} \right\} \right)^{I(Z_k=1)} \right. \\
& \left. \left((1-\pi) \frac{1}{\tau_k^{2(\frac{\nu_2}{2}+1)}} \exp\left\{ -\frac{\nu_2 s_2^2}{2\tau_k^2} \right\} \right)^{I(Z_k=0)} \frac{1}{(\tau_k^2)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2\tau_k^2} \beta_k^2 \right\} \right] \\
& \frac{1}{B(\alpha, \gamma)} \pi^{\alpha-1} (1-\pi)^{\gamma-1} \frac{1}{(\sigma^2)^{(\frac{\nu_\sigma}{2}+1)}} \exp\left\{ -\frac{\nu_\sigma s_\sigma^2}{2\sigma^2} \right\}
\end{aligned}
$$

Gibbs sampling algorithm

Gibbs sampling was used to obtain a sequence of observed values of the parameters [20], [21]. The full conditional posterior distribution for μ is given by

$$
\mu^{(g)} | \boldsymbol{\beta}^{(g-1)}, \tau_k^{2(g-1)}, \sigma^{2(g-1)}, \pi^{(g-1)}, \mathbf{z}^{(g-1)}, \mathbf{y} \sim N\left(\tilde{\mu}^{(g)}, \tilde{\sigma}^{2(g)} \right) \tag{8}
$$

where

$$
\tilde{\mu}^{(g)} = \frac{\frac{\sum_{i=1} y_i^{*(g)}}{\sigma^{2(g-1)}}}{\left(\frac{n}{\sigma^{2(g-1)}} + \frac{1}{\omega^{2(g-1)}} \right)};
$$

$$
\tilde{\sigma}^{2(g)} = \frac{1}{\left(\frac{n}{\sigma^{2(g-1)}} + \frac{1}{\omega^{2(g-1)}} \right)};
$$

$$
y_i^{*(g)} = y_i - \sum_{k=1}^{p} \beta_k^{(g-1)} x_{ik}
$$

and g is the counter of Gibbs sampling algorithm.

The full conditional posterior distribution for τ_k^2 given $Z_k = 1$ is defined by

$$
\tau_k^{2(g)} | \mu^{(g)}, \boldsymbol{\beta}^{(g-1)}, \pi^{(g-1)}, \sigma^{2(g-1)}, Z_k = 1, \mathbf{y} \sim \text{Scaled-Inv}\chi^2 \left(\nu_1 + 1, \frac{\frac{\beta_k^{2(g-1)}}{\sigma^{2(g-1)}} + \nu_1 s_1^2}{\nu_1 + 1} \right) \tag{9}
$$

Likewise, for given $Z_k = 0$ we have

$$
\tau_k^{2(g)} | \mu^{(g)}, \boldsymbol{\beta}^{(g-1)}, \pi^{(g-1)}, \sigma^{2(g-1)}, Z_k = 0, \mathbf{y} \sim \text{Scaled-Inv}\chi^2 \left(\nu_2 + 1, \frac{\frac{\beta_k^{2(g-1)}}{\sigma^{2(g-1)}} + \nu_2 s_2^2}{\nu_2 + 1} \right) \tag{10}
$$

The values of z_k are obtained computing the probability Z_k given $\tau_k^{(2g)}$ and $\pi^{(g)}$, i.e,

$$
Pr(Z_k = 1 | \tau_k^{2(g)}, \pi^{(g-1)}) = \frac{\pi f_1(\tau_k^2)}{\pi f_1(\tau_k^2) + (1-\pi) f_2(\tau_k^2)} \tag{11}
$$

where $f_1(\tau^2)$ $f_2(\tau^2)$ are probability density functions of scaled inverse chi square distribution with parameters (ν_1, s_1^2) and (ν_2, s_2^2), respectively.

Moreover,

$$Z_k \sim Bernoulli(1, Pr(Z_k = 1|\tau_k^{2(g)}, \pi^{(g-1)})). \tag{12}$$

The full conditional posterior distribution for π is defined by

$$\pi^{(g)}|\mu^{(g)}, \boldsymbol{\beta}^{(g-1)}, \tau_k^{2(g)}, \sigma^{2(g-1)}, \mathbf{z}^{(g)}, \mathbf{y} \sim \text{Beta}\left(\alpha + \sum_{z_k:z_k=1} z_k^{(g)}, \gamma + p - \sum_{z_k:z_k=1} z_k^{(g)}\right). \tag{13}$$

The procedure to sampling β_k given $Z_k = 1$ from the full conditional posterior distribution was adapted from the strategy purposed by [22], i.e

$$\beta_k^{(g)}|\mu^{(g)}, \tau_k^{2(g)}, \sigma^{2(g-1)}, \pi^{g)}, \boldsymbol{\beta}_{-k}^{(g-1)}, Z_k = 1, \mathbf{y} \sim N(\dot{\beta}_k^{(g)}, \dot{\sigma}^{2(g)}) \tag{14}$$

where

$$\dot{\beta}_k^{(g)} = \frac{\frac{\sum_{i=1} x_{ij} y_{ik}^{**(g)}}{\sigma^{2(g-1)}}}{\frac{\sum_{i=1} x_{ij}^2}{\sigma^{2(g-1)}} + \frac{1}{\sigma^{2(g-1)}\tau_k^{2(g)}}};$$

$$\dot{\sigma}_k^{2(g)} = \frac{1}{\frac{\sum_{i=1} x_{ij}^2}{\sigma^{2(g-1)}} + \frac{1}{\sigma^{2(g-1)}\tau_k^{2(g)}}};$$

and

$$y_{ik}^{**(g)} = y_i - \sum_{j \neq k} \beta_j^{(g-1)} x_{ij}.$$

Finally, the full conditional posterior distribution for σ^2 is given by

$$\sigma^{2(g)}|\mu^{(g)}, \boldsymbol{\beta}^{(g)}, \tau_k^{2(g)}, \mathbf{y} \sim \text{Scaled-Inv}\chi^2\left(\nu_\sigma + n, \frac{\text{SSE} + \text{SSB} + \nu_\sigma s_\sigma^2}{\nu_\sigma + n + p}\right) \tag{15}$$

where $\text{SSE} = \sum_{i=1}^n \left(y_i - \mu - \sum_{k=1}^p \beta_k^{(g)} x_{ik}\right)^2$ and $\text{SSB} = \sum_{k=1}^p \frac{\beta_k^2}{\tau_k^2}$.

For Bayesian Ridge Regression, there are a unique τ^2 hyperparameter. The prior distribution for τ^2 and σ^2 follows scaled inverse chi square with hyperparameters (ν, s^2) and (ν_σ, s_σ^2).

The full conditional posterior distribution for τ^2 is

$$\tau^{2(g)}|\mu^{(g)}, \boldsymbol{\beta}^{(g-1)}, \pi^{(g-1)}, \sigma^{2(g-1)}, \mathbf{y} \sim \text{Scaled-Inv}\chi^2\left(\nu + p, \frac{\sum_{k=1}^p \frac{\beta_k^{(}2(g-1)}{\sigma^{2(g-1)}} + \nu s^2}{\nu + p}\right)$$

and for σ^2, we have

$$\sigma^{2(g)}|\mu^{(g)}, \boldsymbol{\beta}^{(g)}, \tau^{2(g)}, \mathbf{y} \sim \text{Scaled-Inv}\chi^2\left(\nu_\sigma + n + p, \frac{\text{SSE} + \text{SSB} + \nu_\sigma s_\sigma^2}{\nu_\sigma + n + p}\right)$$

Consequently, the full conditional posterior for β_k parameters is given by

$$\beta_k^{(g)}|\mu^{(g)}, \tau^{2(g-1)}, \sigma^{2(g-1)}, \pi^{(g-1)}, \boldsymbol{\beta}_{-k}^{(g-1)}, Z_k = 1, \mathbf{y} \sim N(\dot{\beta}_k^{(g)}, \dot{\sigma}^{2(g)})$$

where

$$\dot{\beta}_k^{(g)} = \frac{\frac{\sum_{i=1} x_{ij} y_{ik}^{**(g)}}{\sigma^{2(g-1)}}}{\frac{\sum_{i=1} x_{ij}^2}{\sigma^{2(g-1)}} + \frac{1}{\sigma^{2(g-1)}\tau^{2(g)}}};$$

$$\dot{\sigma^2}_k^{(g)} = \frac{1}{\frac{\sum_{i=1} x_{ij}^2}{\sigma^{2(g-1)}} + \frac{1}{\sigma^{2(g-1)}\tau^{2(g)}}}.$$

In summary, the Gibbs sampling algorithm can be defined by

For 1 to G do it:

1- Generate $\mu^{(g)}$ from (8).

2- Generate $\tau^{2(g)}$ from (9).

3- Generate $\mathbf{z}^{(g)}$ from (11) and (12).

4- Generate $\pi^{(g)}$ from (13).

5- Generate each $\beta_k^{(g)}$ from (14).

6- Generate $\sigma^{2(g)}$ from (15).

where G is the number of iterations.

This algorithm was implemented in a R package [23] called BayesH avalaible at *https://cran.r-project.org/web/packages/BayesH/index.html*.

Mathematical Details about Prior Distribution

In this section we are going to show some details about conditional prior distribution for $\beta_k|\tau_k^2$ given a prior distribuition for $\tau_k^2|Z_k$. This demonstration is based on [17]. The distribution of hyperparameter τ_k^2 for given $Z_k = 1$ is described by

$$f_{\tau_k^2|z_k=1}(\tau_k^2) = \frac{\left(\frac{s_1^2\nu_1}{2}\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \exp\left\{-\frac{1}{2}\left[\frac{s_1^2\nu_1}{\tau_k^2}\right]\right\} \frac{1}{\tau_k^{2\left(1+\frac{\nu_1}{2}\right)}},$$

and the prior distribution for β_k given σ^2 and τ_k^2 is defined by

$$f_{\beta_k|\sigma^2,z_k=1} = \frac{1}{\sqrt{2\pi\left(\sigma^2\tau_k^2\right)}} \exp\left\{-\frac{1}{2}\left[\frac{\beta_k^2}{\tau_k^2\sigma^2}\right]\right\}.$$

Consequently, the prior distribution for β_k conditioned to σ^2 and $Z_k = 1$ is given by

$$\begin{aligned} f_{\beta_k|\sigma^2,z_k=1} &= \int_0^\infty \frac{\left(\frac{s_1^2\nu_1}{2}\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \exp\left\{-\frac{1}{2}\left[\frac{s_1^2\nu_1}{\tau_k^2} + \frac{\beta_k^2}{\tau_k^2\sigma^2}\right]\right\} \frac{1}{\tau_k^{2\left(1+\frac{\nu_1}{2}\right)}} d\tau_k^2 \\ &\propto \int_0^\infty \exp\left\{-\frac{1}{2}\left[\frac{s_1^2\nu_1}{\tau_k^2} + \frac{\beta_k^2}{\tau_k^2\sigma^2}\right]\right\} \frac{1}{\tau_k^{2\left(1+\frac{\nu_1}{2}\right)}} d\tau_k^2 \end{aligned} \quad (16)$$

To solve the integrate written in (16) we have to make the change of variable

$$u = \frac{1}{2\tau_k^2}\left(\frac{\nu_1 s_1^2\sigma^2 + \beta_k^2}{\sigma^2}\right)$$

and find $d\tau_k^2$ in terms of u and du. It is implies

$$\tau_k^2 = \frac{1}{2u}\left(\frac{\nu_1 s_1^2\sigma^2 + \beta_k^2}{\sigma^2}\right) \tag{17}$$

and

$$d\tau_k^2 = -\frac{1}{2}u^{-2}\left(\frac{\nu_1 s_1^2\sigma^2 + \beta_k^2}{\sigma^2}\right)du \tag{18}$$

Substituing both (17) and (18) in (16) we have

$$\begin{aligned}
f_{\beta_k|\sigma^2,z_k=1} &\propto \int_0^\infty \left(\frac{\nu_1 s_1^2\sigma^2 + \beta_k^2}{\sigma^2}\right)^{\left(-\frac{\nu_1+1}{2}\right)} u^{\frac{\nu_1+1}{2}-1}\exp\{-u\}\,du \\
&\propto \left(\frac{\nu_1 s_1^2\sigma^2 + \beta_k^2}{\sigma^2}\right)^{\left(-\frac{\nu_1+1}{2}\right)} \int_0^\infty u^{\frac{\nu_1+1}{2}-1}\exp\{-u\}\,du \\
&\propto \left(\frac{\nu_1 s_1^2\sigma^2 + \beta_k^2}{\sigma^2}\right)^{\left(-\frac{\nu_1+1}{2}\right)} \Gamma\left(\frac{\nu_1+1}{2}\right) \\
&\propto \left(\nu_1 s_1^2\sigma^2 + \beta_k^2\right)^{\left(-\frac{\nu_1+1}{2}\right)} \Gamma\left(\frac{\nu_1+1}{2}\right) \\
&\propto \left(1 + \frac{\beta_k^2}{\nu_1 s_1^2\sigma^2}\right)^{\left(-\frac{\nu_1+1}{2}\right)} \\
&\propto \left(1 + \frac{\beta_k^2}{\nu_1 \tilde{s}_{1\sigma^2}}\right)^{\left(-\frac{\nu_1+1}{2}\right)}
\end{aligned} \tag{19}$$

showing that (19) is a kernel of scaled t distribution [24], [17] with degree of freedom ν_1 and scale parameter $\tilde{s}_{1\sigma^2} = s_1^2\sigma^2$.

Likewise, for $Z_k = 0$, we have

$$\beta_k|\sigma^2, z_k = 0 \sim \text{Scaled t}\left(0, \nu_2, \tilde{s}_{2\sigma^2}\right),$$

where $\tilde{s}_{2\sigma^2} = s_2^2\sigma^2$.

Consequently,

$$\begin{cases} \beta_k|\sigma^2, Z_k = 1, \sim \text{Scaled-t}(\nu_1, \tilde{s}_{1\sigma^2}) & \text{with probability} \quad \pi \\ \beta_k|\sigma^2, Z_k = 0, \sim \text{Scaled-t}(\nu_2, \tilde{s}_{2\sigma^2}) & \text{with probability} \quad (1-\pi) \end{cases}$$

showing that for $(\nu_1 = \nu_2)$ and $(s_1^2 = s_2^2)$, the prior distribution for each β_k of bayes H model is equivalent the prior distribuition of bayes A model. Furthermore, for s_1^2 or s_2^2 tending to zero, the prior distribution for each β_k is equivalent to bayes B model. There are other possibilities, for example, tending s_1^2 or s_2^2 to infinity, a mixture distribution of slab Gaussian and t-scaled distribuition is obtained as prior for each β_k.

Results and Discussion

In this study, we purposed a new hyperprior for bayesian regression model to predict complex trait. This model were applied in real and simulated datasets.

Results from cross validation studies shows the prediction accuracy of Bayes H model is slight higher than Bayesian Ridge Regression in scenarios where dataset were generated from heterogeneous stock mice population and quite higher for dataset simulated from random mating. Hence, the type population, consequently, the strength of linkage disequilibrium is more influential in the prediction than number of QTLs (2). However, comparing two datasets generated from random mating, the number of QTLs caused a increase of prediction accuracy in Bayes H model. It was not observed difference in the prediction accuracy of Bayes H when were used different hyperpriors.

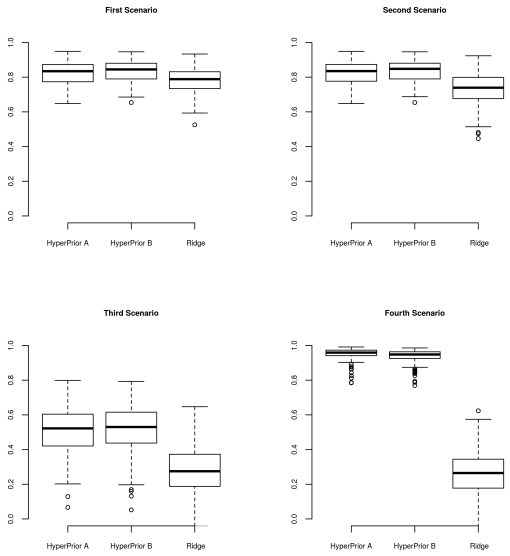

Fig 2. Evaluation of accuracy performance of Bayes H model using 5 fold cross validation. Box plot of Pearson's correlation distribution between observed and predicted values for each simulated scenario.

Figure (3) shows population structure inferred from top two eigenvectors obtained from correlation matrix of mice dataset. Bayesian information criterion indicates the best model is Gaussian mixture with 8 components. Thus, we can infer the presence of eight subpopulations which are the same number of founders of heterogeneous stock mice population 3. The scatterplot of top two eigenvector estimated from to genomic relationship matrix shows the clusters 3. Prediction accuracy of BMI was compared between differents hyperpriors of Bayesian Regression models in the heterogeneous stock mice population. In order to make the comparison, 5 fold cross validation was used. Box plot of Pearson's correlation distribution between observed and predicted values reveals moderate to high accuracy for all models 4. Moreover, the Bayesian ridge regression presented slight higher correlation in regarded to variable selecion model (hyperprior B) and quite higher correlation than model with hyperprior A.

A possible explanation of the fact that Bayes H model outperformed Bayes Ridge Regression only in a simulated dataset is the genetic architecture of the trait BMI. In the simulated data, the QTLs were the unique source of variation considered. Furthermore, the number of QTLs used in the simulations were at most moderate (50). Therefore, models take into consideration that markers have different variances depending on their effects normally predict better than Bayesian Ridge Regression [10, 15]. Using fat percentage dataset in a dairy cattle population, which a single gene explains 50% of the genetic variation, Verbyla et. al [26] reported that predictions from fitting of Bayes Cp have more accuracy than predictions obtained by

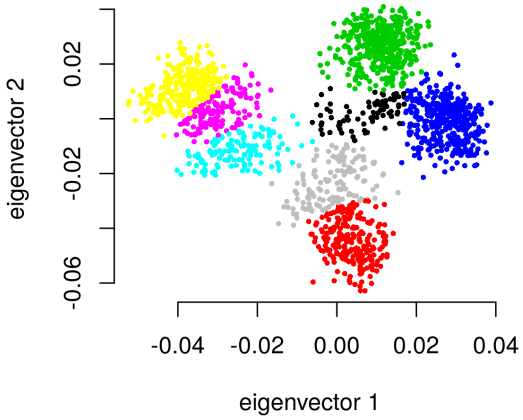

Fig 3. Population structure of heterogeneous stock mice.

RR BLUP. Moreover, Rezende et. al. [27] analyzed data from 17 traits measured in Pinus taeda population comprised of 951 individuals genotyped with 4853 SNPs. They concluded that for trait controlled by few genes, Fusiform rust for example, the models as Bayes A, Bayes Cp had higher predict ability in comparison to RR BLUP. On the other hand, BMI trait is considered a complex trait controlled by a large number of genes [25]. Thus, it is expected the RR BLUP would have a good predictive performance because this model considers homogeneous shrinkage of marker effects. The hypothesis is supported when we considered that genetic architecture of trait can be described by infitesimal model. However, we should have caution with these arguments, Gianola showed heuristically that Bayesian Ridge Regression or RR BLUP does not shrinkage the marker effects the same manner, the best linear unbiased predictor is sample size and allele frequency dependent [17, 18]. Here we would like to speculate another hypothesis about the reason of good predictive performance of the Bayesian Ridge Regression in the real mice dataset. In the real dataset there are many source of genetic variation besides QTLS, such as: background genetic, linkage disequilibrium, epistasis effects and etc. Consequently, the linear model declared in all Bayesian model is not true. Hence, the idea to select the markers that contribute the phenotypic variation does not work well. And this case, the prediction provided by RR BLUP or Bayesian Ridge Regression would be a better approximation.

Table 3. Gaussian mixture model selection using Bayesian Information Criterion

Modelo	BIC
1	16888.60
2	18724.63
3	19643.02
4	19906.28
5	20030.32
6	20199.56
7	20253.73
8	20288.74

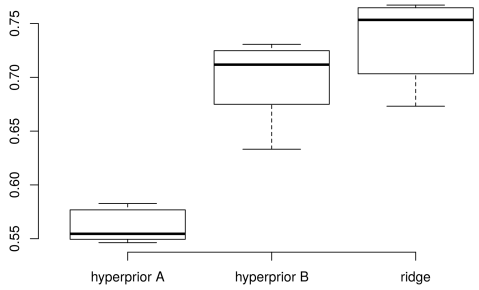

Fig 4. Evaluation of accuracy performance of Bayes H model using 5 fold cross validation. Box plot of Pearson's correlation distribution between observed and predicted values for mice dataset .

References

1. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics. 2013;193(1):327–345. doi:10.1534/genetics.112.143313.

2. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12(1):55–67. doi:10.1080/00401706.1970.10488634.

3. Tibishirani R. Regression Shrinkage and Selection via the Lasso. Journal Royal Statistical Society B. 1996;58(1):267–288. doi:10.2307/2346178.

4. Frank IE, Friedman JH. A Statistical View of Some Chemometrics Regression Tools. Technometrics. 1996;35(2):109–135. doi:10.2307/1269656.

5. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association. 2001;96(456):1348–1360. doi:10.1198/016214501753382273.

6. Fan J, Lv J. A Selective Overview of Variable Selection in High Dimensional Feature Space. Statistica Sinica. 2010;20(01):101–148.

7. George EI, McCulloch RE. Variable Selection via Gibbs Sampling. Journal of the American Statistical Association. 1993;88(423):881–889. doi:10.1080/01621459.1993.10476353.

8. Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008;103(482):681–685. doi:http://dx.doi.org/10.1198/016214508000000337.

9. Heffner EL, Sorrells ME, Jannink JL. Genomic selection for crop improvement. Crop Science. 2009;49(1):1–12. doi:doi:10.2135/cropsci2008.08.0512.

10. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 2001;157(4):1819–1829.

11. Bernardo R. Breeding for Quantitative Traits in Plants 2nd Edition. Woodbury: Stemma Press; 1994.

12. Lateef DD. DNA Marker Technologies in Plants and Applications for Crop Improvements. Journal of Biosciences and Medicines. 2015;3(5):7–18. doi:10.4236/jbm.2015.35002.

13. Hallauer AR, Carena MJ, Miranda Filho JB. Quantitative Genetics in Maize Breeding (Handbook of Plant Breeding, Vol. 6). New York: Springer; 2010.

14. Piepho HP. Ridge regression and extensions for genomewide selection in maize. Crop Science. 2009;49(4):1165–1179. doi:10.2135/cropsci2008.10.0595.

15. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;186(12):1–12. doi:10.1186/1471-2105-12-186.

16. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. 1918.52:34.

17. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. Additive Genetic Variability and the Bayesian Alphabet. Genetics. 2009;183(1):347–363. doi:10.1534/genetics.109.103952.

18. Gianola D. Priors in whole-genome regression: The Bayesian alphabet returns. Genetics. 2013;90(3):525–540. doi:10.1534/genetics.113.151753.

19. Perez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;3(2):483–495. doi:10.1534/genetics.114.164442.

20. Gilks WR, Richardson S, Spiegelhalter D. Markov Chain Monte Carlo in Practice. London: Chapman & Hall; 1996.

21. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, 3rd ed. London: Chapman & Hall; 2004.

22. de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, et al. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics. 2009;182(1):375–385. doi:10.1534/genetics.109.101501.

23. R: A Language and Environment for Statistical Computing; 2014. Available from: `http://www.R-project.org/`.

24. Box GEP, Tiao GC. Bayesian Inference in Statistical Analysis. New York: John Wiley and Sons, Inc; 1992.

25. Legarra A, Robert-Granié C, Manfredi E, Elsen JM. Performance of Genomic Selection in Mice. Genetics. 2008;180(01):611–618. doi:10.1534/genetics.108.088575.

26. Verbyla KL et. al. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genetic Research. 2009;91(5):307–11. doi: 10.1017/S0016672309990243.

27. Rezende MRF, et. al. Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.) Genetics. 2012;190:1503–1510. doi: 10.1534/genetics.111.137026.

28. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JNP, Mott R, et al. Genetic and Environmental Effects on Complex Traits in Mice. Genetics. 2006;174(02):959–984. doi:10.1534/genetics.106.060004.

29. Fraley C, Raftery AE, Murphy TB, Scrucca L. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation; 2012.

30. Fraley C, Raftery AE. Model-based Clustering, Discriminant Analysis and Density Estimation. Journal of the American Statistical Association. 2002;97(458):611–631.