Chapter 11
Bayesian Approach to Genomic Selection

Introduction to Bayesian Method and MCMC Algorithm

Bayesian method is often used in complicated problems where a maximum likelihood method does not have an easy solution. It requires high dimensional multiple numerical integration, which was not feasible prior to the development of high power computers. With computers, we can perform numerical integration very efficiently. The Markov chain Monte Carlo (MCMC) algorithm is a way to perform numerical integration. Bayesian method divides a complicated problem into many small problems (parameters). Each small problem (parameter) is often very simple, e.g., with known distribution conditional on all other problems (parameters) and can be simulated by sampling a random variable from the known distribution. With the Bayesian method, every parameter is considered a random variable with a prior distribution. After incorporating the data, we update (or modify) the prior to obtain a posterior distribution for the variable of interest. A random number is sampled from the posterior distribution. The posterior distribution here means the conditional posterior distribution because it depends on known values of all other variables. After all variables are sampled, this only completes one cycle of the MCMC. The process continues for many cycles. We will then collect the posterior sample to obtain some simple statistic, e.g., mean and standard deviation, for each variable (parameter).

We now use a simple regression method to demonstrate the MCMC algorithm, although the simple regression method is so “simple” that there is no need to perform MCMC. The linear model in matrix notation is

$$y = X_0 b_0 + X_1 b_1 + e \tag{1}$$

where X_0 is a unity vector of length *n* (every element of this matrix is one), b_0 is the intercept, b_1 is the regression coefficient, and $e \sim N(0, I_n \sigma^2)$ is the residual errors where I_n is an identity matrix with dimension $n \times n$ and σ^2 is the residual variance. In this simple problem, the data are represented by $D = \{y, X_0, X_1\}$, the parameters are $\theta = \{b_0, b_1, \sigma^2\}$.

Likelihood

We first define the likelihood as the density of observed variables given parameters. In our case, it is the distribution of the response variable *y* given the values of the parameters. The likelihood is the normal density,

$$\begin{aligned} p(y \mid \theta) &= N(y \mid X_0 b_0 + X_1 b_1, \sigma^2) \\ &= \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} (y - X_0 b_0 - X_1 b_1)^T (y - X_0 b_0 - X_1 b_1) \right] \end{aligned} \tag{2}$$

Note that the notation for a normal distribution $N(x \mid \mu, \sigma^2)$ is a general one so that you do not have to write the long expression of the density function. We often ignore the constant in the density because it is irrelevant to the parameters. The constant here in the likelihood is $(2\pi)^2$.

Prior distribution

We need to define prior distributions for the three parameters. In the simple regression analysis, we often choose the following priors for the parameters,

$$\begin{aligned} &p(b_0) = N(b_0 \mid 0, 1\text{E}10) \\ &p(b_1) = N(b_0 \mid 0, \sigma_b^2) \\ &p(\sigma^2) = \text{Scaled} - \text{Inverse} - \chi^2(\tau, \omega), \text{ where } \tau = \omega = 1\text{E-}10 \\ &p(\sigma_b^2) = 1/\sigma_b^2 = \text{Jefferys' Prior} = \text{general}\left(\text{-log}(\sigma_b^2)\right) \end{aligned} \tag{3}$$

Sometimes we can choose $p(b_0) \propto 1$ and $p(\sigma^2) \propto 1$ as the prior distributions for the two parameters. These priors are called uniform priors or uninformative priors or flat priors. These priors mean that there is no prior information for these parameters. We often write $p(b_0) = 1$ and $p(\sigma^2) = 1$, and ignore them in the formulation of the Bayesian method. Let

$$p(b_1) = N(b_1 \mid 0, \sigma_b^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{b_1^2}{2\sigma_b^2}\right) \tag{4}$$

be the prior distribution for the regression coefficient. This prior is called the shrinkage prior because the prior mean is zero and will be used later in genomic prediction when we deal with a large number of regression coefficients. If $p(b_0) = 1$ and $p(\sigma^2) = 1$ are used as the priors, the joint prior for all three parameters is

$$p(\theta) = p(b_0) p(b_1) p(\sigma^2) = p(b_1) = N(b_1 \mid 0, \sigma_b^2) \tag{5}$$

Posterior distribution

Combining the likelihood and the prior distribution, we get the joint distribution of the response variable and the parameters,

$$p(y, \theta) = p(y \mid \theta) p(\theta) = N(y \mid X_0 b_0 + X_1 b_1, \sigma^2) N(b_1 \mid 0, \sigma_b^2) \tag{6}$$

This joint distribution is proportional to the posterior distribution,

$$p(\theta \mid y) = \frac{p(y, \theta)}{p(y)} \propto p(y \mid \theta) p(\theta) = N(y \mid X_0 b_0 + X_1 b_1, \sigma^2) N(b_1 \mid 0, \sigma_b^2) \tag{7}$$

where the marginal distribution for the response variance

$$p(y) = \int p(y, \theta) d\theta \tag{8}$$

is irrelevant to the parameters and thus is ignored.

Conditional posterior distribution

Conditional posterior distribution is the distribution of a single variable given the data and the values of all other variables. The conditional posterior for the intercept is

$$p(b_0 \mid y, b_1, \sigma^2) = \frac{p(b_0, b_1, \sigma^2 \mid y)}{p(b_1, \sigma^2)} \propto p(b_0, b_1, \sigma^2 \mid y) \tag{9}$$

After some simplification, we can see that this distribution is normal,

$$p(b_0 \mid y, b_1, \sigma^2) = N(b_0 \mid \tilde{b}_0, V_{\tilde{b}_0}) \tag{10}$$

where

$$\tilde{b}_0 = (X_0^T X_0)^{-1} X_0^T (y - X_1 b_1) = \frac{1}{n}\sum_{j=1}^{n}(y_j - X_{1j} b_1) \tag{11}$$

and

$$V_{\tilde{b}_0} = (X_0^T X_0)^{-1}\sigma^2 = \frac{1}{n}\sigma^2 \tag{12}$$

From this distribution, we can simulate an intercept, which will be treated as a known variable that allows you to derive the conditional posterior distribution for the next variable.

The conditional distribution of the regression coefficient given other variables is

$$p(b_1 \mid y, b_0, \sigma^2) \propto p(b_0, b_1, \sigma^2 \mid y) \tag{13}$$

This distribution is also normal,

$$p(b_1 \mid y, b_0, \sigma^2) = N(b_1 \mid \tilde{b}_1, V_{\tilde{b}_1}) \tag{14}$$

where

$$\tilde{b}_1 = (X_1^T X_1 + \sigma^2 / \sigma_b^2)^{-1} X_1^T (y - X_0 b_0) \tag{15}$$

and

$$V_{\tilde{b}_1} = (X_1^T X_1 + \sigma^2 / \sigma_b^2)^{-1}\sigma^2 \tag{16}$$

From this distribution, a new regression coefficient can be sampled.

Given b_0 and b_1, we can find a conditional distribution for σ^2, which is

$$p(\sigma^2 \mid y, b_0, b_1) \propto p(b_0, b_1, \sigma^2 \mid y) \tag{17}$$

This distribution happens to be a scaled inverse chi-square distribution,

$$\begin{aligned} p(\sigma^2 \mid y, b_0, b_1) &= \text{Scale} - \text{Inv} - \chi^2(\sigma^2 \mid \tilde{n}, s^2) \\ &= \frac{(\tilde{n}s^2/2)^{\tilde{n}/2}}{\Gamma(\tilde{n}/2)} \frac{1}{(\sigma^2)^{(\tilde{n}+2)/2}} \exp\left(-\frac{\tilde{n}s^2}{2\sigma^2}\right) \end{aligned} \tag{18}$$

where $\tilde{n} = n - 2$ and

$$s^2 = \frac{1}{n-2}\sum_{j=1}^{n}(y_j - X_{0j} b_0 - X_{1j} b_1)^2 \tag{19}$$

To simulate σ^2 from this distribution, you first simulate a variable from χ^2_{n-2} distribution and then take

$$\sigma^2 = \frac{\tilde{n}s^2}{\chi^2_{n-2}} = \frac{1}{\chi^2_{n-2}}\sum_{j=1}^{n}(y_j - X_{0j} b_0 - X_{1j} b_1)^2 \tag{20}$$

Summary of the MCMC steps

Step 0. Set $t=0$ and initialize $\theta^{(t)}=\left\{b_0^{(t)},b_1^{(t)},\sigma^{2(t)}\right\}$

Step 1. Sample b_0 from $N(b_0\,|\,\tilde{b}_0,V_{\tilde{b}_0})$

Step 2. Sample b_1 from $N(b_1\,|\,\tilde{b}_1,V_{\tilde{b}_1})$

Step 3. Sample σ^2 from $\text{Scale}-\text{Inv}-\chi^2(\sigma^2\,|\,\tilde{n},s^2)$

Step 4. Increment *t* by 1, i.e., $t=t+1$, and go back to step 1.
Step 5. If $t=T$, then stop, where T is a preset large number, say 100000.

Post MCMC analysis

Delete the values of the first 1000 cycles (burn-in) and then save one observation in every 20 cycles (thinning) to form a posterior sample of the parameters: $\left\{\theta^{(1)},\theta^{(2)},...,\theta^{(N)}\right\}$. The posterior mean of the parameters are the Bayesian estimates of the parameters, which are

$$\hat{\theta}=\frac{1}{N}\sum_{k=1}^{N}\theta^{(k)} \tag{21}$$

and the posterior variance matrix of the parameters is

$$\text{var}(\hat{\theta})=\frac{1}{N}\sum_{k=1}^{N}(\theta^{(k)}-\hat{\theta})(\theta^{(k)}-\hat{\theta})^T \tag{22}$$

PROC REG

Simulated data with 100 observations. The data were generated in a data step using the following statements,

```
data one;
  b0=10;
  b1=5;
  sigma2=15;
  x0=1;
  do i=1 to 100;
     x1=normal(0)*sqrt(10)+10;
     e=normal(0)*sqrt(sigma2);
     y=x0*b0+x1*b1+e;
     output;
  end;
  keep x0 x1 y;
run;
```

The first 10 observations are shown in **Table 1.** We first use PROC REG to estimate the parameters using the following SAS code,

```
ods graphics off;
proc reg data=one;
   model y =x0 x1/noint;
quit;
```

Table 1

First 10 observations of a simulated dataset with sample size 100

x0	x1	y
1	15.79237	84.41699
1	20.80989	115.9174
1	4.111151	29.71018
1	29.07647	157.3909
1	19.91511	104.142
1	32.90716	170.0362
1	-1.07753	2.186243
1	17.62758	96.64255
1	19.11299	104.3976
1	0.356999	12.88888

Table 2

Output of the regression analysis for PROC REG

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	668468	334234	23652.5	<.0001
Error	98	1384.83673	14.13099		
Uncorrected Total	100	669853			

Root MSE	3.75912	R-Square	0.9979
Dependent Mean	62.66835	Adj R-Sq	0.9979
Coeff Var	5.99844		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x0	1	9.87309	0.53306	18.52	<.0001
x1	1	5.02919	0.03600	139.69	<.0001

The estimate regression coefficients and the residual error variance are

$$\begin{bmatrix} b_0 \\ b_1 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 9.87309 \\ 5.02919 \\ 14.13099 \end{bmatrix} \tag{23}$$

PROC MCMC

We now use PROC MCMC to analyze the data. The SAS code for PROC MCMC is

```
filename aa "simuldata.csv";
filename bb "postsample.csv";
data one;
   infile aa dlm=',' firstobs=2;
   input x0 x1 y;
run;
ods graphics on;
proc mcmc data=one outpost=postsample nmc=100000
          thin=100 seed=246810 nbi=1000 ntu=3000
          monitor=(b0 b1 sigma2 sigma2b)
          diag=Geweke(f1=0.25 f2=0.25);
     ods select  PostSumInt Geweke TADpanel;
     parms b0 0;
     parms b1 0;
     parms sigma2 1;
     parms sigma2b 1;
     begincnst;
        tau=1e-10;
        omega=1e-10;
     endcnst;
     prior b0 ~ normal(mean = 0, var = 1e10);
     prior b1 ~ normal(mean = 0, var = sigma2b);
     prior sigma2 ~ sichisq(tau,omega);
     prior sigma2b ~ general(-log(sigma2b));
     mu = b0*x0+b1*x1;
     model y ~ normal(mean = mu, var = sigma2);
run;
ods graphics off;

proc export data=postsample outfile=bb dbms=csv replace;
run;
```

In this analysis, we actually used a hierarchical prior for the regression coefficient. We have a normal prior for $b_1 \sim N(0,\sigma_b^2)$ and then a Jeffery's prior for σ_b^2, which is $p(\sigma_b^2) \propto 1 / \sigma_b^2$. In PROC MCMC, the Jeffery's prior is denoted by

$$\sigma_b^2 \sim \text{general}(-\log(\sigma_b^2))$$

It took seven seconds to complete the MCMC sampling with 100000 iterations. The posterior sample is stored in a file called “postsample.csv”. Part of the posterior sample is shown in **Table 3**.

Table 3
Part of the posterior sample from which posterior estimates of the parameters are inferred

Iteration	b0	b1	sigma2	sigma2b	LogPrior	LogLike	LogPost
1001	9.3744	5.0811	16.0858	205.5	-47.8987	-274.8	-322.7
1101	10.4913	5.0102	15.2132	147.2	-47.3647	-274.2	-321.6
1201	10.0425	5.0074	15.6063	228.9	-48.0222	-273.8	-321.8
1301	10.0704	4.998	13.7709	676.5	-49.4863	-273.7	-323.2
1401	9.7275	5.0382	10.7779	40.1894	-45.3035	-275.1	-320.4
1501	9.8308	5.0504	13.5104	88.0961	-46.5357	-273.6	-320.2
1601	10.0694	5.0124	14.4674	1132.6	-50.3012	-273.5	-323.8
1701	10.7341	4.9494	17.3131	38.8449	-45.7259	-276.5	-322.2
1801	9.5579	5.0646	13.0114	851.9	-49.7718	-273.9	-323.7
1901	9.8679	5.0661	14.3162	23.7207	-45.0217	-274.4	-319.4
2001	9.9898	5.04	10.9804	45.2675	-45.4654	-275.1	-320.5
2101	10.092	5.0387	15.341	271.7	-48.2539	-273.9	-322.2
2201	9.7559	4.9921	13.2423	204.2	-47.6927	-274.9	-322.6
2301	9.7759	5.051	13.9782	30.752	-45.261	-273.6	-318.8
2401	10.6428	4.9823	11.5602	125.6	-46.8653	-275.5	-322.4
2501	10.0167	4.9869	18.9411	30.2963	-45.5381	-276.3	-321.8
2601	9.3623	5.0872	15.3742	178.9	-47.655	-274.8	-322.4
2701	9.8299	5.0021	12.9616	4.404	-44.6962	-274.1	-318.8
2801	9.5447	5.0431	15.7474	7.265	-44.5514	-273.9	-318.4

The output of PROC MCMC are shown in the following tables and figures.

Table 4
Posterior estimate of parameters

Posterior Summaries and Intervals					
Parameter	N	Mean	Standard Deviation	95% HPD Interval	
b0	1000	9.8575	0.5178	8.7936	10.8648
b1	1000	5.0296	0.0356	4.9695	5.1138
sigma2	1000	14.4891	2.0847	10.6288	18.6853
sigma2b	1000	198.6	445.0	2.2300	919.1

Compared with PROC REG, the estimates are very close. The reason for the close similarity is that we the priors distributions we choose are all uninformative.

PROC REG

$$\begin{bmatrix} b_0 \\ b_1 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 9.87309 \\ 5.02919 \\ 14.13099 \end{bmatrix} \tag{24}$$

PROC MCMC

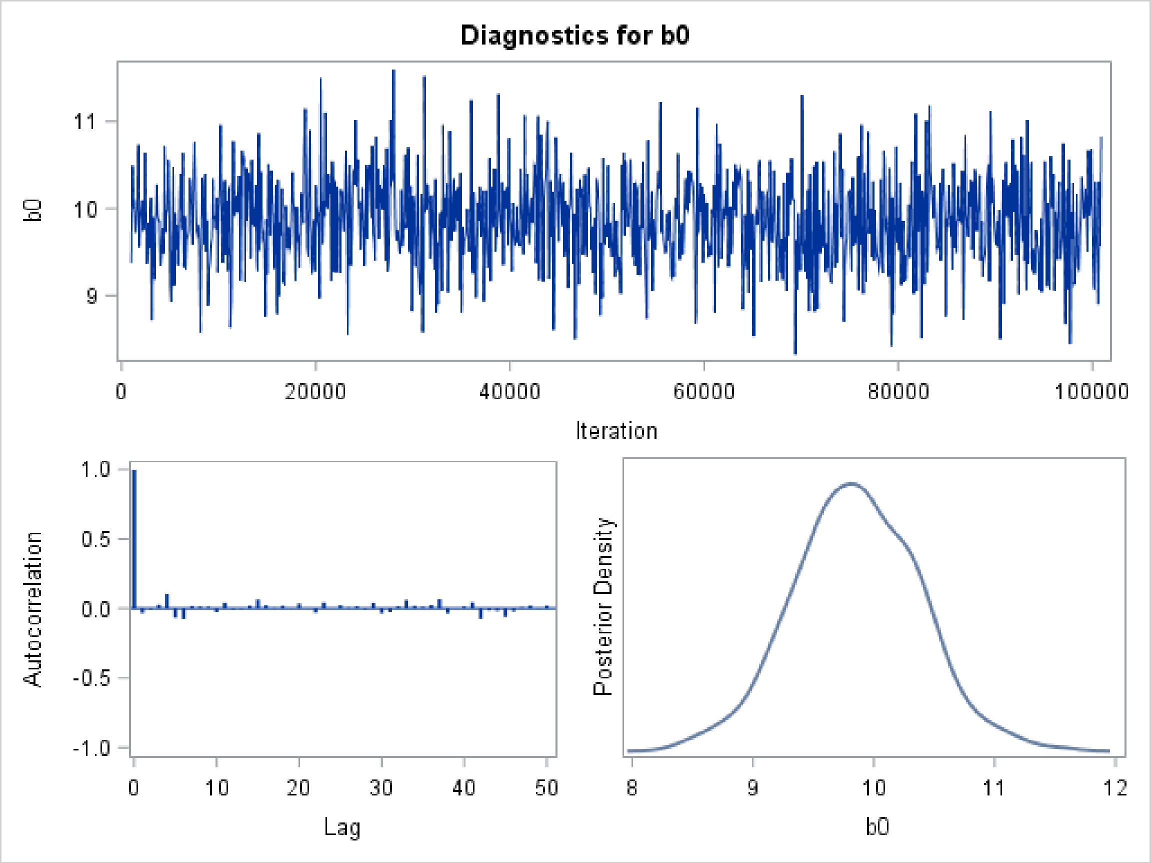
$$\begin{bmatrix} b_0 \\ b_1 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 9.8575 \\ 5.0296 \\ 14.4891 \end{bmatrix} \tag{25}$$

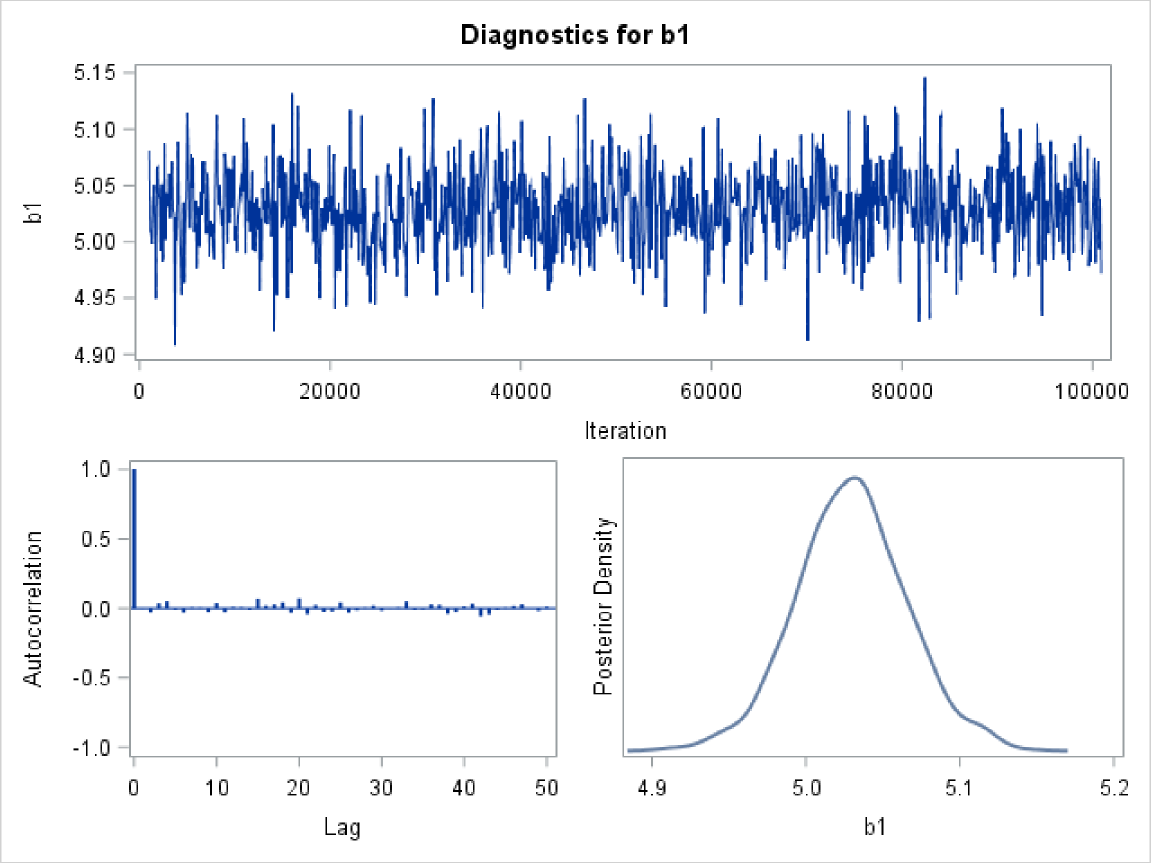
Convergence diagnostic result shows that all parameters have converged to their corresponding stationary distribution.

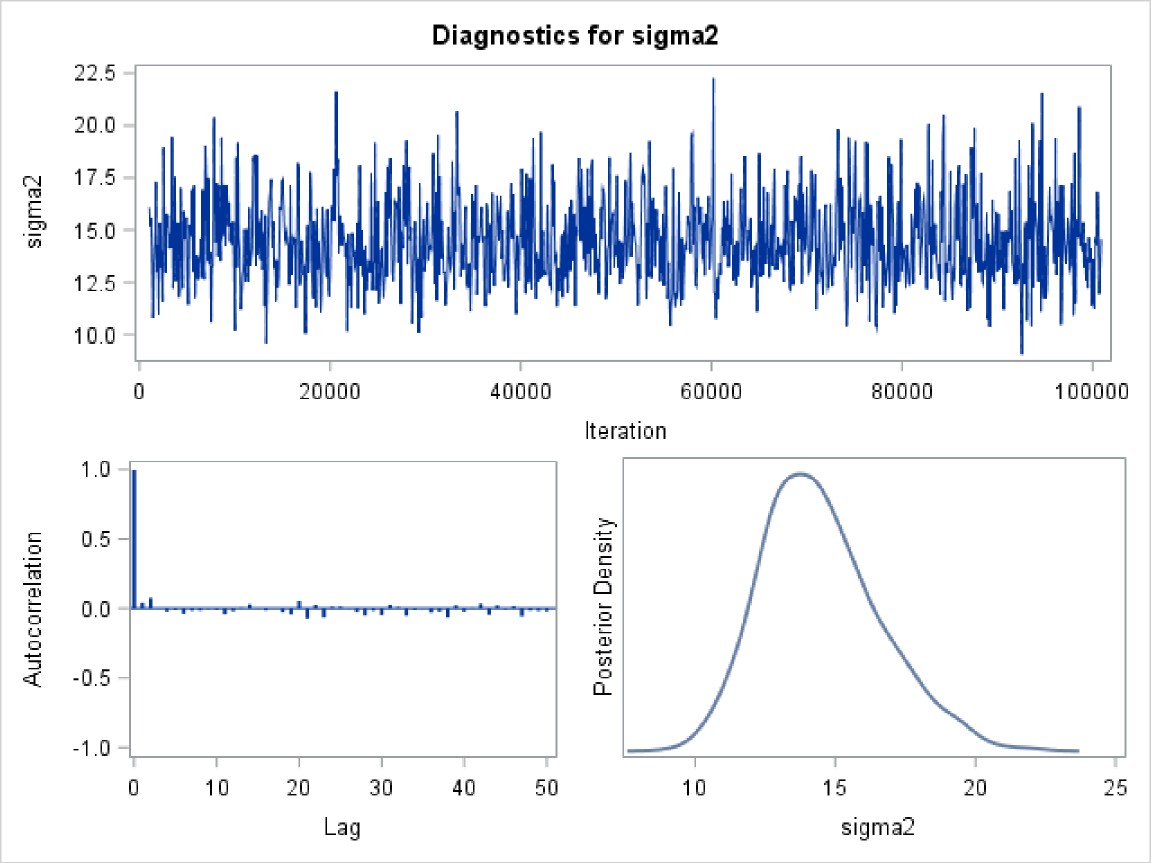
Table 5
Convergence diagnostics

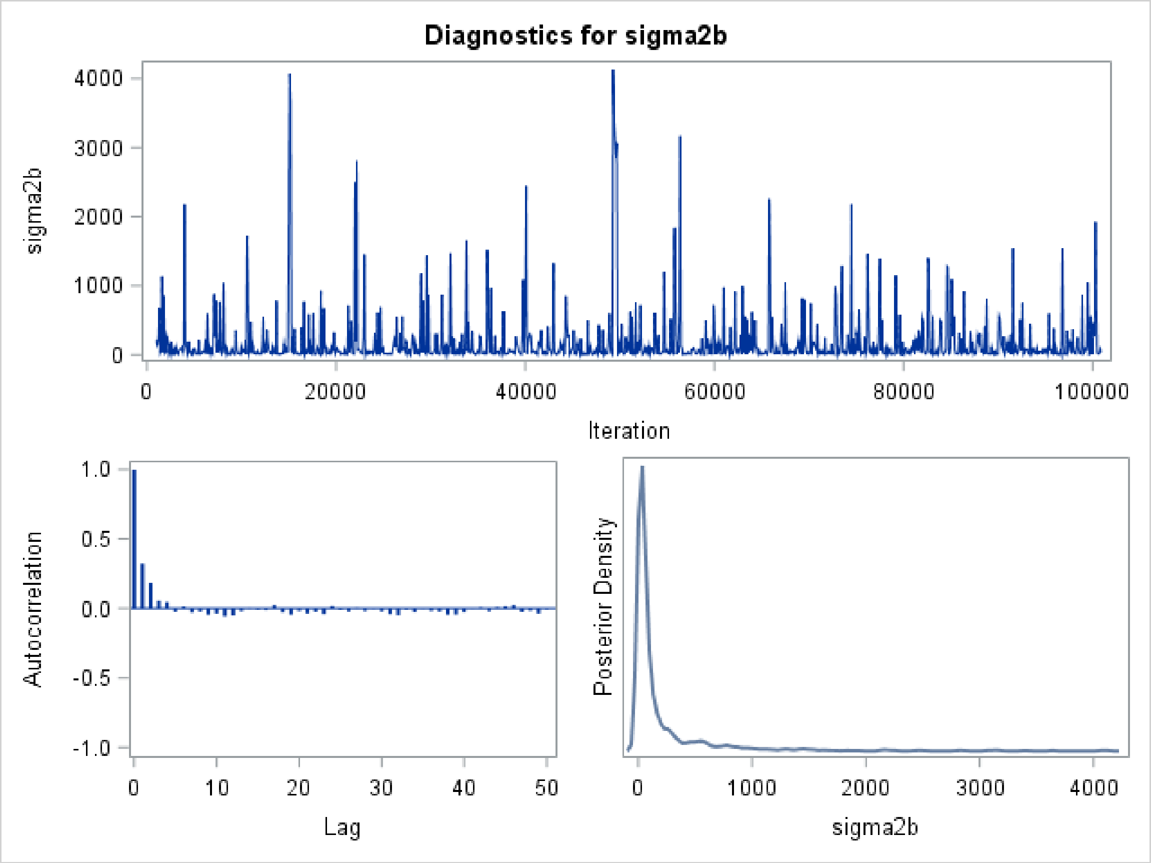
Geweke Diagnostics		
Parameter	z	Pr > |z|
b0	1.3026	0.1927
b1	-1.7654	0.0775
sigma2	0.1658	0.8683
sigma2b	0.3432	0.7314

The following Figures show the Trace-Autocorrelation-Density (TAD) plot panels for the three parameters (b_0 , b_1 and σ^2) and the prior variance of b_1 (σ_b^2).


Diagnostics for b0
b0
11
10
9
0
20000
40000
60000
80000
100000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
8
9
10
11
12
b0


Diagnostics for b1
b1
5.15
5.10
5.05
5.00
4.95
4.90
0
20000
40000
60000
80000
100000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
4.9
5.0
5.1
5.2
b1


Diagnostics for sigma2
sigma2
22.5
20.0
17.5
15.0
12.5
10.0
0
20000
40000
60000
80000
100000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
10
15
20
25
sigma2


Diagnostics for sigma2b
sigma2b
4000
3000
2000
1000
0
0
20000
40000
60000
80000
100000
Iteration
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
Lag
Posterior Density
0
1000
2000
3000
4000
sigma2b

References

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* **193**(2): 327-345.

George EI, McCulloch RE. 1993. Variable selection via Gibbs sampling. *J Am Stat Assoc* **88**(423): 881-889.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819-1829.

Wang H, Zhang Y, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S. 2005. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**(1): 465-480.

Xu S. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics* **163**(2): 789-801.

Xu S. 2007. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**(2): 513-521.

Yi NJ, George V, Allison DB. 2003. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**(3): 1129-1138.

Bayesian Method and MCMC Algorithm

Part of the text is duplicate of “Introduction to Bayesian Method and MCMC Algorithm” described previously.

1. Bayesian regression

Bayesian method is often used in complicated problems where a maximum likelihood method does not have an easy solution. It requires high dimensional multiple numerical integration, which was not feasible prior to the development of high power computers. With computers, we can perform numerical integration very efficiently. The Markov chain Monte Carlo (MCMC) algorithm is a way to perform numerical integration. Bayesian method divides a complicated problem into many small problems (parameters). Each small problem (parameter) is often very simple, e.g., with known distribution conditional on all other problems (parameters) and can be simulated by sampling a random variable from the known distribution. With the Bayesian method, every parameter is considered a random variable with a prior distribution. After incorporating the data, we update (or modify) the prior to obtain a posterior distribution for the variable of interest. A random number is sampled from the posterior distribution. The posterior distribution here means the conditional posterior distribution because it depends on known values of all other variables. After all variables are sampled, this only completes one cycle of the MCMC. The process continues for many cycles. We will then collect the posterior sample to obtain some simple statistic, e.g., mean and standard deviation, for each variable (parameter).

We now use a simple regression method to demonstrate the MCMC algorithm, although the simple regression method is so “simple” that there is no need to perform MCMC. The linear model in matrix notation is

$$y = J_n a + Xb + e \qquad (26)$$

where J_n is a unity vector of length *n* (every element of this matrix is one), a is the intercept, *b* is the regression coefficient, and $e \sim N(0, I_n\sigma^2)$ is the residual errors where I_n is an identity matrix with dimension $n \times n$ and σ^2 is the residual variance. In this simple problem, the data are represented by $D = \{y, X\}$, the parameters are $\theta = \{a, b, \sigma^2\}$.

Likelihood: We first define the likelihood as the density of observed variables given parameters. In our case, it is the distribution of the response variable *y* given the values of the parameters. The likelihood is the normal density,

$$\begin{aligned} p(y \mid \theta) &= N(y \mid J_n a + Xb, \sigma^2) \\ &= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(y - J_n a - Xb)^T (y - J_n a - Xb) \right] \end{aligned} \qquad (27)$$

Note that the notation for a normal distribution $N(x \mid \mu, \sigma^2)$ is a general one so that you do not have to write the long expression of the density function. We often ignore the constant in the density because it is irrelevant to the parameters. The constant here in the likelihood is $(2\pi)^2$.

Prior distribution: We need to define prior distributions for the three parameters. Let $p(a)\propto 1$ and $p(\sigma^2)\propto 1$ be the prior distributions for the two parameters. These prior are called uniform priors or uninformative priors or flat priors. These priors mean that there is no prior information for these parameters. We often write $p(a)=1$ and $p(\sigma^2)=1$, and ignore them in the formulation of the Bayesian method. Let

$$p(b)=N(b\,|\,0,\sigma_b^2)=\frac{1}{\sqrt{2\pi}\sigma_b}\exp\left(-\frac{b^2}{2\sigma_b^2}\right) \tag{28}$$

be the prior distribution for the regression coefficient. This prior is called the shrinkage prior because the prior mean is zero and will be used later in genomic prediction when we deal with a large number of regression coefficients. So, the joint prior for all three parameters is

$$p(\theta)=p(a)p(b)p(\sigma^2)=p(b)=N(b\,|\,0,\sigma_b^2) \tag{29}$$

Posterior distribution: Combining the likelihood and the prior distribution, we get the joint distribution of the response variable and the parameters,

$$p(y,\theta)=p(y\,|\,\theta)p(\theta)=N(y\,|\,J_n a+Xb,\sigma^2)N(b\,|\,0,\sigma_b^2) \tag{30}$$

This joint distribution is proportional to the posterior distribution,

$$p(\theta\,|\,y)=\frac{p(y,\theta)}{p(y)}\propto p(y\,|\,\theta)p(\theta)=N(y\,|\,J_n a+Xb,\sigma^2)N(b\,|\,0,\sigma_b^2) \tag{31}$$

where the marginal distribution for the response variance

$$p(y)=\int p(y,\theta)d\theta \tag{32}$$

is irrelevant to the parameters and thus is ignored.

Conditional posterior distribution: Conditional posterior distribution is the distribution of a single variable given the data and the values of all other variables. The conditional posterior for the intercept is

$$p(a\,|\,y,b,\sigma^2)=\frac{p(a,b,\sigma^2\,|\,y)}{p(b,\sigma^2)}\propto p(a,b,\sigma^2\,|\,y) \tag{33}$$

After some simplification, we can see that this distribution is normal,

$$p(a\,|\,y,b,\sigma^2)=N(a\,|\,\tilde{a},V_{\tilde{a}}) \tag{34}$$

where

$$\tilde{a}=(J_n^T J_n)^{-1}J_n^T(y-Xb)=\frac{1}{n}\sum_{j=1}^{n}(y_j-X_j b) \tag{35}$$

and

$$V_{\tilde{a}}=(J_n^T J_n)^{-1}\sigma^2=\frac{1}{n}\sigma^2 \tag{36}$$

From this distribution, we can simulate an intercept, which will be treated as a known variable that allows you to derive the conditional posterior distribution for the next variable.

The conditional distribution of the regression coefficient given other variables is

$$p(b \mid y, a, \sigma^2) \propto p(a, b, \sigma^2 \mid y) \tag{37}$$

This distribution is also normal,

$$p(b \mid y, a, \sigma^2) = N(b \mid \tilde{b}, V_{\tilde{b}}) \tag{38}$$

where

$$\tilde{b} = (X^T X \ + \sigma^2 / \sigma_b^2)^{-1} X^T (y - J_n a) \tag{39}$$

and

$$V_{\tilde{b}} = (X^T X \ + \sigma^2 / \sigma_b^2)^{-1} \sigma^2 \tag{40}$$

From this distribution, a new regression coefficient can be sampled.

Given *a* and *b*, we can find a conditional distribution for σ^2, which is

$$p(\sigma^2 \mid y, a, b) \propto p(a, b, \sigma^2 \mid y) \tag{41}$$

This distribution happens to be a scaled inverse chi-square distribution,

$$\begin{aligned} p(\sigma^2 \mid y, a, b) &= \text{Scale} - \text{Inv} - \chi^2(\sigma^2 \mid \tilde{n}, s^2) \\ &= \frac{(\tilde{n}s^2/2)^{\tilde{n}/2}}{\Gamma(\tilde{n}/2)} \frac{1}{(\sigma^2)^{(\tilde{n}+2)/2}} \exp\left(-\frac{\tilde{n}s^2}{2\sigma^2}\right) \end{aligned} \tag{42}$$

where $\tilde{n} = n - 2$ and

$$s^2 = \frac{1}{n-2} \sum_{j=1}^{n} (y_j - a - X_j b)^2 \tag{43}$$

To simulate σ^2 from this distribution, you first simulate a variable from χ^2_{n-2} distribution and then take

$$\sigma^2 = \frac{\tilde{n}s^2}{\chi^2_{n-2}} = \frac{1}{\chi^2_{n-2}} \sum_{j=1}^{n} (y_j - a - X_j b)^2 \tag{44}$$

Summary of the MCMC steps:

Step 0. Set $t = 0$ and initialize $\theta^{(t)} = \left\{a^{(t)}, b^{(t)}, \sigma^{2(t)}\right\}$

Step 1. Sample *a* from $N(a \mid \tilde{a}, V_{\tilde{a}})$

Step 2. Sample *b* from $N(b \mid \tilde{b}, V_{\tilde{b}})$

Step 3. Sample σ^2 from $\text{Scale} - \text{Inv} - \chi^2(\sigma^2 \mid \tilde{n}, s^2)$

Step 4. Increment *t* by 1, i.e., $t = t + 1$, and go back to step 1.

Step 5. If $t = T$, then stop, where T is a preset large number, say 100000.

Post MCMC analysis: Delete the values of the first 1000 cycles (burn-in) and then save one observation in every 20 cycles (thinning) to form a posterior sample of the parameters: $\left\{\theta^{(1)},\theta^{(2)},...,\theta^{(N)}\right\}$. The posterior mean of the parameters are the Bayesian estimates of the parameters, which are

$$\hat{\theta}=\frac{1}{N}\sum_{k=1}^{N}\theta^{(k)} \tag{45}$$

and the posterior variance matrix of the parameters is

$$\text{var}(\hat{\theta})=\frac{1}{N}\sum_{k=1}^{N}(\theta^{(k)}-\hat{\theta})(\theta^{(k)}-\hat{\theta})^{T} \tag{46}$$

2. BayesA (Bayesian shrinkage estimation)

BayesA and BayesB were proposed by Meuwissen et al. (2001) for genomic selection. BayesA was also used for QTL mapping but it was called Bayesian shrinkage estimation (Xu 2003; Wang et al. 2005). The linear model is

$$y=J_n a+\sum_{k=1}^{m}X_k b_k+e \tag{47}$$

where *a* is the population mean (or intercept), X_k is the genotype indicator for marker *k*, b_k is the marker effect and *m* is the number of markers. Let us define

$$X_{-k}b_{-k}=\sum_{k'=1,k'\neq k}^{m}X_{k'}b_{k'} \tag{48}$$

The linear model can be rewritten as

$$y-X_{-k}b_{-k}=J_n a+X_k b_k+e \tag{49}$$

Defining $y_k=y-X_{-k}b_{-k}$, we get

$$y_k=J_n a+X_k b_k+e \tag{50}$$

Comparing this with the simple regression model introduced before,

$$y=J_n a+Xb+e \tag{51}$$

we see that the two models are exactly the same except that we redefined $y=y_k$, $X=X_k$ and $b=b_k$. So, everything we did there applies here except that we have to deal with *m* regression coefficients.

We now introduce the concept of hierarchical prior. One key issue here is how to choose the prior variance for each regression coefficient. This is the place where BayesA and BayesB can help. Let $p(b_k)=N(b_k\,|\,0,\sigma_k^2)$ be the prior variance for the effect of marker *k*. We also need to define a prior for σ_k^2 so that σ_k^2 can be sampled from the data. This type of Bayesian analysis is called hierarchical prior Bayesian analysis or simply hierarchical Bayesian analysis. Let us choose the following hierarchical prior for σ_k^2,

$$p(\sigma_k^2)=\text{Scale-Inv}-\chi^2(\sigma_k^2\,|\,\tau_0,\omega_0) \tag{52}$$

where τ_0 is a prior degree of freedom and ω_0 is a prior scale parameter. Let us now consider marker *k* only. The parameter vector now includes σ_k^2 as well, $\theta=\left\{a,b_k,\sigma_k^2,\sigma^2\right\}$. The conditional posterior distribution for σ_k^2 is

$$p(\sigma_k^2 \mid y, a, b_k, \sigma^2) \propto p(a, b_k, \sigma_k^2, \sigma^2 \mid y) \tag{53}$$

which is

$$p(\sigma_k^2 \mid y, a, b_k, \sigma^2) = \text{Scale-Inv} - \chi^2(\sigma_k^2 \mid \tau_k, \omega_k) \tag{54}$$

where $\tau_k = \tau_0 + 1$ is the posterior degree of freedom and $\omega_k = (\tau_0\omega_0 + b_k^2)/(\tau_0 + 1)$ is the posterior scale parameter. To sample σ_k^2, you first sample a variable from $\chi^2_{\tau_0+1}$ distribution (this variable is also denoted by $\chi^2_{\tau_0+1}$) and then take

$$\sigma_k^2 = \frac{1}{\chi^2_{\tau_0+1}}(\tau_0\omega_0 + b_k^2) \tag{55}$$

I often choose $(\tau_0, \omega_0) = (0, 0)$ and such a scaled inverse chi-square is called the Jeffrey prior. The sampling process is very simple, taking

$$\sigma_k^2 = \frac{b_k^2}{\chi_1^2} \tag{56}$$

where χ_1^2 is a random variable with a chi-square one distribution. BayesA will not generate sparse model. Although the posterior mean of a regression coefficient can be extremely small but cannot be exactly zero. BayesB will allow you to generate a spare model.

3. BayesB (stochastic search variable selection)

In statistics, this method is called stochastic search variable selection (SSVS). The method was originally developed by George and McCulloch (1993) and applied to QTL mapping by Yi et al. (2003). Meuwissen et al. (2001) adopted this idea to genomic selection for the first time and gave a new named called BayesB. The majority of steps in BayesB are the same as those in BayesA except that: (1) there is a probability π at which σ_k^2 is set to zero and the π value is interpreted as the proportion of non-segregating markers and (2) if σ_k^2 is not set to zero (with probability $1-\pi$), it is sampled using a Metropolis-Hastings algorithm. Let $\sigma_k^{2(\text{new})}$ be a randomly sampled value of σ_k^2 from its prior distribution, $\text{Scale-Inv} - \chi^2(\sigma_k^2 \mid \tau_0, \omega_0)$. Replace σ_k^2 by $\sigma_k^{2(\text{new})}$ with a probability

$$\alpha = \frac{N(y_k - J_n a \mid 0, X_k X_k^T \sigma_k^{2(\text{new})} + I\sigma^2)}{N(y_k - J_n a \mid 0, X_k X_k^T \sigma_k^2 + I\sigma^2)} \tag{57}$$

Recall the notation of normal distribution $N(x \mid \mu_x, \sigma_x^2)$. With the M-H sampling algorithm, the sample sequence of σ_k^2 may have many repeats because a newly sampled value may not always be accepted (only accepted with a probability α).

The original SSVS algorithm is much easier than its modified version BayesB. An additional Bernoulli variable η_k is introduced to the model to indicate whether b_k should be included in the model (if $\eta_k = 1$) or excluded from the model (if $\eta_k = 0$). The situation of b_k being excluded from the model is equivalent to assigning $b_k = 0$. This explains why it is called variable selection. The model can be sparse because a marker effect can take exactly a value of 0. Depending on the strength of the selection and the actual data, a model can have only a few non-zero marker effects. This is equivalent to modelling b_k with a mixture of two normal distributions,

$$p(b_k) = \eta_k N(b_k \mid 0, \varDelta) + (1 - \eta_k) N(b_k \mid 0, \varDelta^{-1}) \tag{58}$$

where $\varDelta = 10^5$ or any large number (a constant). The prior distribution of η_k is $p(\eta_k) = \text{Bernoulli}(\eta_k \mid \rho)$, where $0 < \rho < 1$ is the parameter of the Bernoulli prior. The conditional posterior distribution of η_k remains Bernoulli,

$$p(\eta_k \mid ...) = \text{Bernoulli}(\eta_k \mid \tilde{\rho}_k) \tag{59}$$

where

$$\tilde{\rho}_k = \frac{\rho N(b_k \mid 0, \Delta)}{\rho N(b_k \mid 0, \Delta) + (1-\rho) N(b_k \mid 0, \Delta^{-1})} \tag{60}$$

is the posterior probability of $\eta_k = 1$.

Note that $\rho = 1 - \pi$ is the proportion of markers with non-zero effects. I used the ρ notation in 2007 (Xu 2007) and assigned a hierarchical prior to ρ, which is $p(\rho) = \text{Beta}(\rho \,|\, 1,1)$. The conditional posterior distribution is

$$p(\rho \mid ...) = \text{Beta}(\rho \mid m_0 + 1, m_1 + 1) \tag{61}$$

where $m_1 = \sum_{k=1}^{m} \eta_k$ and $m_0 = m - m_1$.

4. R program (BGLR)

The Bayesian General Linearized Regression (BGLR) package can perform Bayesian analysis using BayesA, BayesB and many other genomic prediction methods (de los Campos et al. 2013). The sample data are stored in two files: “IMF2-Genotypes.csv” for the genotype data and “IMF2-Phenotypes.csv” for the phenotype data of 278 hybrids in rice.

```
dir<-"C:\\Users\\Xu\\Lecture Notes\\Bayes"
setwd(dir)

gen<-read.csv(file="IMF2-Genotypes.csv",header=T)
phe<-read.csv(file="IMF2-Phenotypes.csv",header=T)
z<-t(as.matrix(gen[,-c(1:4)]))
y<-phe$kgw
foldid<-phe$foldid
n<-length(y)

library(BGLR)

setwd(tempdir())

eta<-list(list(X=z,model="BayesB"))
fm<- BGLR(y=y,ETA=eta,nIter=1500,burnIn=500,thin=10,verbose=F)
bHat<- fm$ETA[[1]]$b
bHat.sd<- fm$ETA[[1]]$SD.b
aHat<-fm$mu
aHat.sd<-fm$SD.mu
yHat<-fm$yHat

nfolds<-max(phe$foldid)
w<-table(foldid)/n
yHatCV<-numeric(n)
yObs<-numeric(n)
for(fold in 1:nfolds){
   yNa<-y
   whichNa<-which(foldid==fold)
   yNa[whichNa]<-NA
   eta<-list(list(X=z,model="BayesB"))
   fm<-BGLR(y=yNa,ETA=eta,nIter=1500,burnIn=500,thin=10,verbose=F)
   yHatCV[whichNa]<-fm$yHat[whichNa]
   yObs[whichNa]<-y[whichNa]
}
pred<-data.frame(yObs,yHatCV,yHat)
cor(pred)^2

setwd(dir)
write.csv(x=pred,file="Result.csv",row.names=F)
```

```
> cor(pred)^2
            yObs    yHatCV     yHat
yObs   1.0000000 0.6631693 0.858992
yHatCV 0.6631693 1.0000000 0.933786
yHat   0.8589920 0.9337860 1.000000
```

Comments:

The Bayesian Alphabet series get longer and longer. The series are already up to BayesR. I am not a big fan of the Bayesian Alphabet series. I prefer a simple model and BayesA is considered one of the simple models. When the model gets complicated, it tries to react to minor fluctuation of the data. Although a complicated model may fit the data well (perfectly sometime), it will reduce model predictability. Ridge regression (also called gBLUP) is the simplest model and it often outperforms all other models. The gBLUP method will be introduced in next chapter.

References

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* **193**(2): 327-345.

George EI, McCulloch RE. 1993. Variable selection via Gibbs sampling. *J Am Stat Assoc* **88**(423): 881-889.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819-1829.

Wang H, Zhang Y, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S. 2005. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**(1): 465-480.

Xu S. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics* **163**(2): 789-801.

Xu S. 2007. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**(2): 513-521.

Yi NJ, George V, Allison DB. 2003. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**(3): 1129-1138.