

## Exercise on Genomic Selection Bayes A Model

### Part 1) Fitting a genomic selection model

The data set file (csv format) is called reference.csv. This file has 50 rows (plus header) and 11 columns. Each row refers to an individual (an animal or plant); the first column is a phenotypic measurement ( $y$ ) and the remaining 10 columns are genotypes for diallelic markers coded as 0, 1 and 2 for genotypes aa, Aa and AA.

Please fit a Bayes A model for these data using the 10 markers simultaneously, with their additive effects only, i.e., the model should be as:

$$y_i = \mu + \sum_{j=1}^{10} x_{ij}g_j + \varepsilon_i,$$

where  $y_i$  is the phenotypic observation on individual  $i$  ( $i = 1, 2, \dots, 50$ ),  $\mu$  is the overall (phenotypic) mean,  $x_{ij}$  is the genotype of genetic marker  $j$  on individual  $i$ ,  $g_j$  is the genetic effect associated with marker  $j$ , and  $\varepsilon_i$  is a residual effect, assumed  $\varepsilon_i \sim N(0, \sigma_e^2)$ .

As discussed in class, in the Bayes A formulation it is assumed that the gene (marker) effects have normal prior distributions centered on zero and with gene specific variances  $\sigma_j^2$ , i.e.,  $g_j | \sigma_j^2 \sim N(0, \sigma_j^2)$ , and that the residual and gene variances ( $\sigma_e^2$  and  $\sigma_j^2$ ,  $i=1, \dots, 10$ ) have prior scaled inverse chi-square distributions.

For fitting such a model, an R code is provided. What you need to do is just to change the address of the file reference.csv (wherever it is on your computer). On the Appendix you'll find the R code with some additional comments regarding other commands and information you may want to change on the code when running your analysis.

After fitting your model, please answer the following questions:

1. Please print the trace plot with the first 1,000 iterations (samples) of the MCMC for the residual variance.
2. What are the posterior means of the marker effects? Is there any marker with a substantially larger effect than the others?
3. Please print the posterior density of the marker effect with the highest (absolute) value.
4. Please provide the posterior means and standard deviations for  $\mu$  and  $\sigma_e^2$ .

## APPENDIX

**Note:** On the code below, the parameter is defined as 'mu', the gene effects are denote by a vector 'g', and the variances  $\sigma_j^2$  and  $\sigma_e^2$  are defined as 'gvar' and 'vare', respectively.

### # R coding for the Bayes A approach of genomic selection

```
nmarkers <- 10 #number of markers
nrecords <- 50 #number of records
numit <- 1000 #number of iterations
```

Here you can change the number of MCMC iterations if you wish

```
x<-
read.table('/Users/Rosa/Dropbox/2013/teaching/ESALQ/Lab
/genomic selection/reference.csv',sep=',',header=TRUE)
```

Here you need to put the correct address for the file on your computer

```
y<-as.matrix(x[,1])
x<-as.matrix(x[,2:11])
```

```
# Setting up storage vectors and matrices for MCMC output
gStore <- array(0,c(numit,nmarkers))
gvarStore <- array(0,c(numit,nmarkers))
vareStore <- array(0,c(numit))
muStore <- array(0,c(numit))
ittstore <- array(0,c(numit))
```

```
# Gibbs sampling
```

```
# Starting values and some additional arrays
g <- array(0.01,c(nmarkers))
mu <- 50
gvar <- array(0.1,c(nmarkers))
ones <- array(1,c(nrecords))
e <- array(0,c(nrecords))
```

```
for (l in 1:numit) {
```

```
# Sampling vare
e <- y - x%*%g - mu
vare <- (t(e)%*%e)/rchisq(1,nrecords-2)
```

```
# Sampling the mean
mu <- rnorm(1,(t(ones)%*%y -t(ones)%*%x%*%g)/nrecords,sqrt(vare/nrecords))
```

```

# Sampling gvar
for (j in 1:nmarkers) {
# gvar[j] <- (0.002+g[j]*g[j])/rchisq(1,4.012+1) # Meuwissen et al. (2001) prior
# gvar[j] <- (0.002+g[j]*g[j])/rchisq(1,1) # Xu (2003) prior
gvar[j] <- (0.002+g[j]*g[j])/rchisq(1,0.998) # Te Braak et al. (2006) prior
}

# Sampling the marker effects
z <- array(0,c(nrecords))
for (j in 1:nmarkers) {
gtemp <- g
gtemp[j] <- 0
for (i in 1:nrecords) {
z[i] <- x[i,j]
}
mean <- ( t(z)%*%y-t(z)%*%x%*%gtemp-t(z)%*%ones*mu ) /
(t(z)%*%z+vare/gvar[j])
g[j] <- rnorm(1,mean,sqrt(vare/(t(z)%*%z+vare/gvar[j])))
}

# store the sampled values
for (j in 1:nmarkers) {
gStore[l,j] <- g[j]
gvarStore[l,j] <- gvar[j]
}
vareStore[l] <- vare
muStore[l] <- mu
ittstore[l] <- l
}

# For visual inspection of convergence of MCMC for each parameter (Trace plot)
# plot(ittstore[1:1000],gStore[1:1000,1],type='l')

# To look at the posterior density of each parameter
# plot(density(gStore[101:1000,1]))

# To get the posterior means of each parameter (e.g., gene effects)
# mean(gStore[101:1000,1])

```

Here you will need to make the necessary changes to answer questions 1-4.