# GENOMIC SELECTION USING A FAST EM ALGORITHM
# 1. UNDERSTANDING THE METHODOLOGY

**R.K. Shepherd[1] and J.A. Woolliams[2]**

[1]FABIE, CQUniversity, Rockhampton, QLD 4702
[2]The Roslin Institute & R(D)SVS, University of Edinburgh, Roslin, Midlothian, EH25 9PS UK

## SUMMARY

The paper uses the technical computer software Mathematica® to explain the features inherent in the procedure emBayesB which is a fast EM algorithm for implementing genomic selection by mapping QTL in genome-wide dense SNP marker data. The prior mixture for a SNP effect and the bimodal shape of the posterior distribution of a SNP effect are displayed graphically, along with visualisations and calculations of how emBayesB estimates genomic breeding values. The companion paper (Shepherd *et al.* 2009) uses emBayesB to analyse simulated data.

## INTRODUCTION

Genomic selection is a new tool for genetic improvement in animal breeding which uses genome-wide dense SNP markers to ensure all QTL are in linkage disequilibrium (LD) with at least one marker. The first step in genomic selection is the estimation of SNP effects using phenotype and genotype data in a reference population (training data), followed by calculation of genomic breeding values (GEBV) using only marker genotypes (and previously estimated SNP effects) in the population for selection (validation data). Mixed model methods and Bayesian MCMC (Markov Chain Monte Carlo) methods have been recommended for genomic selection. Bayesian MCMC methods generally have the highest accuracy of predicting GEBV but are slow computationally (Lund *et al.* 2009). An Expectation Maximisation (EM) algorithm can use valuable information in a prior distribution as in a Bayesian approach and is usually much faster. This paper describes an EM algorithm called emBayesB for genomic selection and explains visually the features inherent in emBayesB using the technical computer software Mathematica®.

## EM APPROACH FOR 1 SNP

It is instructive to first visualise the estimation of the effect of one SNP. Then the algorithm is extended to the estimation for *m* SNP where *m* is usually much larger than *n*, the number of individuals.

**Data model for 1 SNP.** The linear model $\mathbf{y} = \mathbf{b}g + \mathbf{e}$ is used to relate record $y_i$ of individual *i* to the SNP effect $g$ where element $b_i$ of vector $\mathbf{b}$ is the number (0, 1 or 2) of reference SNP alleles for individual *i*. We standardise $\mathbf{b}$ so that $\mathbf{1}'\mathbf{b} = 0$ and $\mathbf{b}'\mathbf{b} = n$. The errors are assumed normal and independent so that $\mathbf{y} \mid g \sim N\left(\mathbf{b}g, \mathbf{I}\sigma_e^2\right)$. Using maximum likelihood (ML) we find that the likelihood distribution of the estimate of the SNP effect $g$ given the data is normal ie. $g \mid \mathbf{y} \sim N\left(g_L, \sigma^2\right)$ where the *best* estimate of the SNP effect is the mean $g_L$ $(= \frac{1}{n}\mathbf{b}'\mathbf{y})$ which is the weighted data average and $\sigma^2 = \frac{1}{n}\sigma_e^2$. The two likelihoods displayed in Figure 1A (for $g_L = 0.6$ and $\sigma_e^2 = 1$) illustrate the finding that, as *n* increases, the likelihood becomes narrower ie. we are more confident about the ML (or *best*) estimate of $g$ with more information.

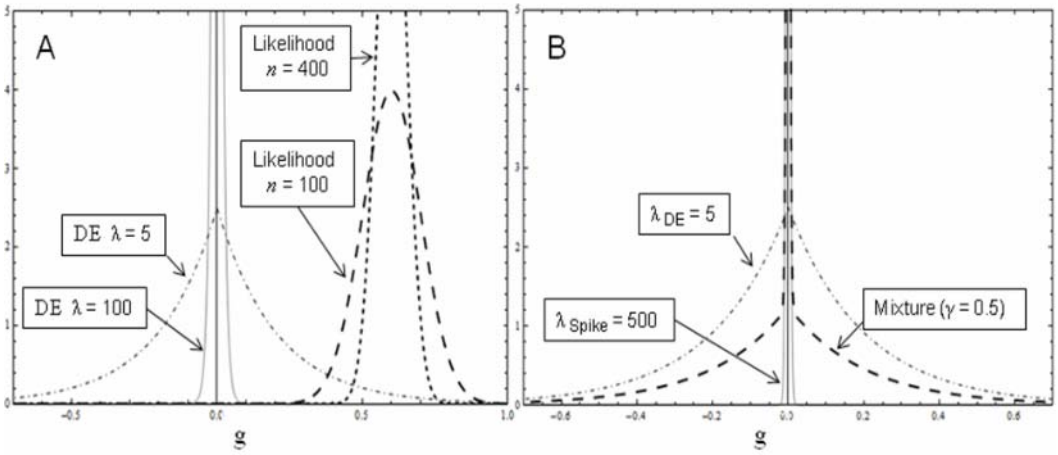


**Figure 1. A. Normal likelihoods and DE distributions. B. Prior mixture for $g$ when $\gamma = 0.5$**

**Prior distribution for 1 SNP.** It is assumed *a priori* that SNP effect $g$ has a probability $1-\gamma$ of being 0 and a probability $\gamma$ **(**due to LD with QTL) of being distributed as a Double Exponential (DE) with parameter $\lambda$ ie. $p(g) = 0.5\lambda \exp(-\lambda \mid g \mid) = PDF_{DE}$. Figure 1A shows the DE shape for two values of $\lambda$. The prior for $g$ can be written as a mixture $\pi(g) = \gamma PDF_{DE} + (1-\gamma)\,\delta(g)$ where $\delta(g)$ is the Dirac Delta function which has all its probability mass at 0. Figure 1B shows the prior mixture for $g$ using a Spike (or DE with $\lambda_{Spike} = 500$) for the Dirac Delta function (as a Dirac Delta function cannot be easily graphed) and a DE with $\lambda_{DE} = 5$ for the 50% ($\gamma = 0.5$**)** chance of the SNP being in LD with QTL. As the Spike's $\lambda$ gets larger and $\gamma$ gets smaller, the prior mixture is often described as a 'spike and slab' prior (see prior mixture in Figure 2A).

**Posterior distribution for 1 SNP.** The posterior for $g$ is illustrated in Figure 2. When the likelihood estimate ($g_L$) is distant from 0 the posterior distribution resembles the likelihood distribution, but is slightly displaced (or regressed) toward 0 as shown in Figure 2A. When $g_L$ is much greater than 0, the mode of the regressed likelihood is $g_L - \lambda_{DE}\sigma^2 \left(= DE_{mode}\right)$ which is the posterior mode for a DE only prior, ie. the LASSO estimate (Yi and Xu 2008) of a SNP effect, as the Spike has no influence if $g_L$ is distant from 0. As $g_L$ gets closer to 0 the posterior becomes bimodal, with the height at 0 increasing the closer $g_L$ is to 0 (Figure 2). This reflects the fact that the true $g$ is more probably 0, the closer $g_L$ is to 0. Using Mathematica® it can be shown that the area under the DE part of the posterior is 0.99, 0.60 and 0.14 for $g_L$ values of 0.2, 0.15 and 0.1 respectively, assuming the parameter values given in Figure 2. These DE areas are basically the posterior probabilities of $g$ being non-zero given the assumed or current parameter estimates. The

posterior probability $\gamma_{post}$ (see Figure 2) of a SNP effect being non-zero (ie. in LD with at least one QTL) form the E-step of the EM algorithm for genomic selection called emBayesB.

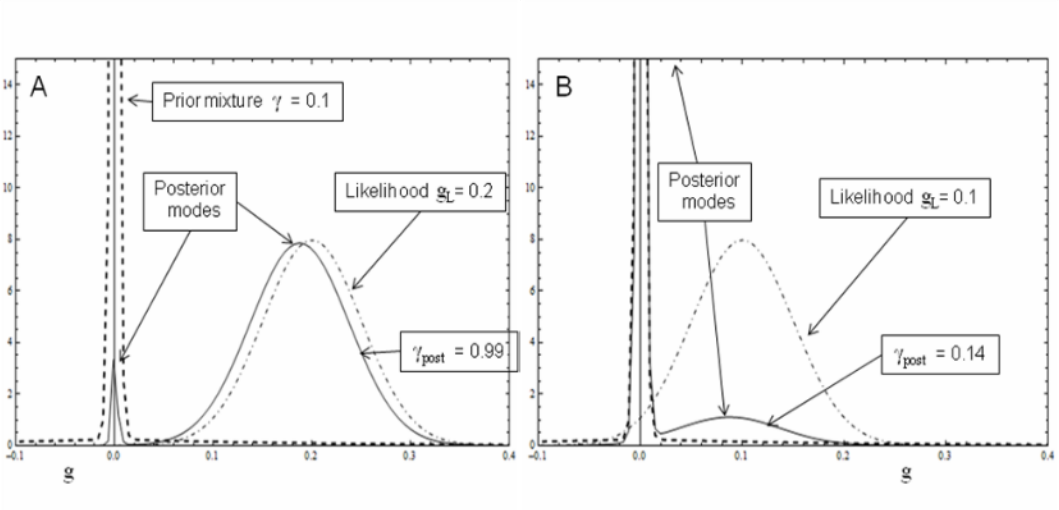


**Figure 2. Bimodal posterior distribution of $g$ as $g_L$ approaches 0 for a 'spike and slab' prior where $\gamma$, $\lambda_{DE}$, $\lambda_{Spike}$, $\sigma_e^2$ and $n$ are 0.1, 5, 500, 1 and 400 respectively.**

**EM ALGORITHM FOR MANY SNP**

The data model for *m* SNP is $\mathbf{y} = \mathbf{Bg} + \mathbf{e}$ which linearly relates record $y_i$ of individual *i* to the $j^{th}$ SNP effect $g_j$ where element $b_{ij}$ of matrix $\mathbf{B}$ is the number (0, 1 or 2) of reference alleles of SNP *j* for individual *i* (usually standardised). The errors are assumed normal and independent such that $\mathbf{y} \mid \mathbf{g} \sim N\left(\mathbf{Bg}, \mathbf{I}\sigma_e^2\right)$. If we knew precisely which SNP were in LD with QTL (maybe only 100 SNP), then the problem would be much easier. This missing information is crucial in formulating an EM algorithm. We define an unknown variable $z_j$ which indicates if the $j^{th}$ SNP is in LD with QTL ($z_j = 1$) or not ($z_j = 0$). If $z_j = 1$, the SNP effect $g_j$ is assumed to be a Double Exponential random variable with parameter $\lambda$; while if $z_j = 0$, the SNP effect is assumed to be distributed as a Dirac Delta (DD) function which has all its probability mass at 0. We assume *a priori* that a fraction $\gamma$ of the SNPs are in LD with QTL.

Using EM theory we are able to develop an iterative sequence of E and M-steps which converge to maximum *a posteriori* (MAP) parameter estimates. At iteration *k* the E-step involves calculating $\gamma_j^k \left(= E\left[z_j \mid \mathbf{y} \,\&\, \textit{all current estimates}\right]\right)$, the posterior probability that SNP *j* is in LD with QTL given the data and all current parameter estimates (eg. like calculating $\gamma_{post}$ in Figure 2). This is done analytically and fast. Then the M-step uses derived formulae (not shown here) to calculate updated estimates of $g_j, \gamma, \lambda, \sigma_e^2$ given the data and the current values of $\gamma_j^k$. This step is also done very quickly using Gauss-Seidel iteration for the many estimates of $g_j$. Iterating

between the E and M-steps the algorithm converges quickly to produce MAP estimates of $g_j$, ML estimates of $\gamma, \lambda, \sigma_e^2$ and posterior probabilities $\gamma_j^k$ (which are useful for mapping QTL).

## DISCUSSION

The derived formula $\hat{g}_j = \gamma_j^k DE_{mode} + \left(1-\gamma_j^k\right) DD_{mode} = \gamma_j^k DE_{mode}$ gives the MAP estimate of $g_j$. This formula shows that the best estimate of a SNP effect is a weighted average of the two posterior modes. However the mode ($DD_{mode}$) for a Dirac Delta only prior is always 0. So we have that the best estimate of the SNP effect with emBayesB is a proportion of the DE mode. But genetic gain is greatest if the posterior mean is used to estimate each QTL effect (Goddard and Hayes 2007). Using Mathematica®, we can show for Figure 2 that the posterior mean, for $g_L$ values of 0.2, 0.15 and 0.1, is 0.1848, 0.0829 and 0.0124 respectively, while the weighted average of the two posterior modes is 0.1847, 0.0827 and 0.0120 respectively. Hence the weighted average of the two posterior modes is an accurate estimate of the posterior mean of a SNP effect. Bayesian MCMC methods use the estimated posterior mean of each SNP effect in the prediction equation $\mathbf{GEBV = B\hat{g}}$, whereas emBayesB uses the weighted average of the two posterior modes. Hence it is no surprise to find that the accuracy of 0.85 between GEBV and true breeding value for emBayesB is similar to the accuracies of 0.84 to 0.87 for Bayesian MCMC methods when analysing the validation data of the QTLMAS XII dataset (Shepherd *et al.* 2009).

emBayesB works by shrinking the ML estimates of the SNP effects. If the ML estimate is distant from 0 the shrinkage is mainly due to the DE prior and the shrunken estimate is called the LASSO (Yi and Xu 2008). The closer the ML estimate is to 0, the greater is the shrinkage. This is due to the Dirac Delta prior kicking in and reflects the fact that only a small proportion (ie. $\gamma$) are believed non-zero *a priori*. The algorithm combines this prior belief with the data to iteratively derive a probability for each SNP of being non-zero. Then it further regresses the $DE_{mode}$ (or LASSO estimate) for this SNP by its probability of being non-zero. It is this double shrinkage which makes emBayesB so accurate and able to handle all the noise in the data from having so many SNP most of which aren't in LD with QTL. Basically it removes the effects of lots of SNP from the genomic breeding value as these SNP are most probably not in LD with QTL.

Not only is emBayesB accurate when predicting breeding values but it is also very fast. emBayesB uses Gauss Seidel iteration to quickly calculate an analytical *posterior-like* mean for for each SNP effect and iterates until the SNP estimates converge. Bayesian MCMC methods sample sequentially from distributions which eventually converge to the true posterior distribution for each SNP. As there are thousands of SNP, the SNP distributions take a long time to converge.

## ACKNOWLEDGMENTS

This paper reports collaborative research instigated while RKS was on sabbatical at the Roslin Institute with support from CQUniversity and the Roslin Institute.

## REFERENCES

Goddard, M.E. and Hayes, B.J. (2007) *J. Anim. Breed. Genet.* **124**:323

Lund, M.S., Sahana, G., de Koning, D-J., Su G. and Carlborg, O. (2009) *BMC Proc.* **3**(Suppl 1):S1

Meuwissen, T.H.E., Solberg T.R., Shepherd R., Woolliams J.A. (2009) *Genet. Sel. Evol.* **41:**2

Shepherd R.K., Meuwissen, T.H.E. and Woolliams J.A. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**: 84.

Yi, N. and Xu, S. (2008) *Genetics* **179**:1045