A BAYES-A LIKE METHOD IN ASREML

Arthur R Gilmour

School of Mathematics and Applied Statistics, University of Wollongong, NSW, 2522
***Cargo Vale* Cargo NSW 2800**

SUMMARY

The paper describes a method implemented in **ASReml** for estimating genomic breeding values and marker effects distributed according to a *t* distribution from a large panel of SNP markers. The method is similar to the MCMC Bayes-A method. It estimates effects in the individual animal space and back-solves to obtain the marker effects.

INTRODUCTION

With the increasing availability of SNP panels for genotype characterization comes the challenge of how best, or at least effectively, to utilize them. Two emphases are common; first to predict breeding values, using the markers to define genetic relatedness more accurately than by using expected average relatedness as predicted from a pedigree, and second to identify loci (markers) of large effect hoping that the large effect is due to a nearby major gene (QTL).

Meuwissen *et al.* (2001) proposed several approaches including methods called GBLUP and Bayes-A. The basic marker model is $\boldsymbol{y} = \mathbf{X}\boldsymbol{b} + \mathbf{M}\boldsymbol{g} + \boldsymbol{e}$ where marker scores (**M**) are used to fit random marker allele effects ($\boldsymbol{g}$) with a common variance σ^2_{g}. This is equivalent to using the genomic relationship matrix we write as $\mathbf{G} = \mathbf{MDM}'$ used in place of the Numerator Relationship matrix in the animal model to produce GBLUP ($\boldsymbol{u}$) where **M** is the matrix of (centred) marker scores (0/1/2) and $\boldsymbol{D}$=diag(1/*s*), $s = \Sigma\ 2p_i(1\text{-}p_i)$ and p_i is the proportion of the minor allele for marker *i* (Stranden and Garrick, 2009). The animal model formulation is generally more tractable because the number of markers typically far exceeds the number of animals. The link is that $\boldsymbol{u} = \mathbf{M}\boldsymbol{g}$ and $\boldsymbol{g} = \mathbf{M}'\mathbf{G}^{-1}\boldsymbol{u}$. In the mixed model equations, **G** is scaled by a variance parameter which is related to the marker variance component: $\sigma^2_{\mathrm{u}}=s\sigma^2_{\mathrm{g}}$.

However, it is likely that markers are not equally informative, that they have diverse variances. The Bayes-A model assumes a scaled inverse Chi-square distribution for the individual marker variances implying a *t* distribution for the marker effects and uses Markov chain Monte Carlo techniques to estimate the marker effects. Sun *et al.* (2012) propose an EM method based on the GBLUP model but where **D**, initialized at diag(1/s), is updated each iteration using the estimated marker effects as $\mathbf{D} = \mathrm{diag}(\boldsymbol{g}{*}\boldsymbol{g} + (k\text{-}2)\sigma^2_g)/(k{+}1)$ where k is the degrees of freedom of the Chi-square distribution and σ^2_g is the marker variance assumed known. The idea here is that if we assume a scaled inverse Chi-square distribution for the marker variances, we can estimate those variances using the estimated marker effects and the scale parameter. We then use the estimated marker variances as weights for the marker covariables and reform **G**. That is then used in the mixed model equations to re-estimate the marker effects. This paper describes an implementation of this method in **ASReml** (Gilmour 2013). The method is called Fast Bayes-A (FBA).

MATERIALS AND METHODS

Data Set The method is demonstrated on a simulated data set (Szydlowski and Paczyńska, 2010) comprising marker scores for 10031 markers on 3226 animals, 2326 of which have phenotype and all of which have 'true' breeding values. The data is supplied in a marker file and a phenotype file. The phenotypic variance for this data is 100.6.

Three Models

For **M**, $\sigma^2_u = s\sigma^2_{\mathrm{g}}$, **g** and k as defined above, we define three forms of the genomic relationship matrix **G**=**MDM**' where **D** is a diagonal matrix of relative variances used to weight each marker:
GBLUP: **D** is diag(1/s), **G** formed once.
FBA-F: **D** is diag($\boldsymbol{g}*\boldsymbol{g}/\sigma^2_u + (k\text{-}2)/s)/(k+1)$, **G** reformed each iteration with σ^2_u fixed
FBA-G: **D** is diag($\boldsymbol{g}*\boldsymbol{g}/\sigma^2_u + k/s)/(k+1)$, reformed each iteration with σ^2_u updated.

The animal model is then fitted using the **G** matrix and marker effects (**g**) are predicted from the animal effects. For GBLUP and FBA-G, the genetic variance (σ^2_u) is estimated; for FBA-F it must be held fixed. FBA-F is the model described by Sun *et al.* (2012) but with **D** multiplied by the σ^2_u rather than applying σ^2_u as a scale factor for **G**. Further, Sun *et al.* (2012) used the σ^2_u estimated from the GBLUP (equal marker variances) model as the known prior variance.

The difference between FBA-F and FBA-G is that σ^2_u is estimated in the latter, and is scaled according to the RHS constant in the expression **D**. So, if *s* is set to one, the variance parameter is related to the marker variance σ^2_g, not the genetic variance; using *k-2* instead of *k* results in a value *k/(k-2)* larger; use of *k/s* results in a variance estimate comparable to the GBLUP value.

Meuwissen *et al.* (2001) used a value of *k* close to 4 which pulls the marker variances toward *(k-2)/(k+1)*=0.4 of the average value under GBLUP. The distribution is less skewed under the FBA-G model and so it does not follow the nominal inverse Chi-square distribution.

Models fitted. These three models were fitted to the simulated data and the FBA models evaluated with *k* at 4.2, 3.8 and 3.5. The FBA-F model was fitted assuming the variance ratio (Genetic/Residual) obtained from the GBLUP fit, although it could have easily been evaluated with σ^2_u =44.0. The FBA models identify a few markers of large effect and we examine the impact of fitting 4 of these as fixed covariates (putative QTL).

RESULTS

The primary results are summarized in Table 1. For the GBLUP model, the genetic variance (ratio) was estimated at 44.03 (0.808) corresponding to a marker variance component of 0.01177 and the Log Likelihood was -6077.5. The largest marker effect was -0.165 for marker 4480.

The number of markers having a large effect was strongly influenced by the value of *k*, with large consequent jumps in the Log Likelihood. However, further reducing *k* to 3.2 gave a poorer fit, especially for the FBA-F model (values not given). From these and other models fitted, we see a jump in Log likelihood for each marker of large effect identified: -6049, -6033, -6012, -6007 for 1, 2, 3 and 4 markers with large effect. Each large marker is effectively fitted as a fixed effect (having a relatively large individual effect variance).

The marker variances are less skewed under FBA-G than FBA-F and so fewer large markers are detected for a given value of *k*. Indeed a plot of marker effects with variance fixed and *k*=3.8 against marker effects estimated when the variance is estimated and *k=3.5* shows very close agreement except for the 2 largest effects which are 20% larger under the latter model (Figure 1).

The accuracy is the correlation between the BLUP values predicted for the 900 individuals without data and the 'true' breeding values of these individuals. It increases with increasing Log Likelihood.

Having identified markers of large effect, these can be fitted as separate covariates. Table 2 shows the Wald F statistics and effects of the 4 markers having largest effect; they explain 40% of the genetic variance. Markers 952 and 954 are neighbours and each is as effective as the other when fitted singly but they also complement each other.

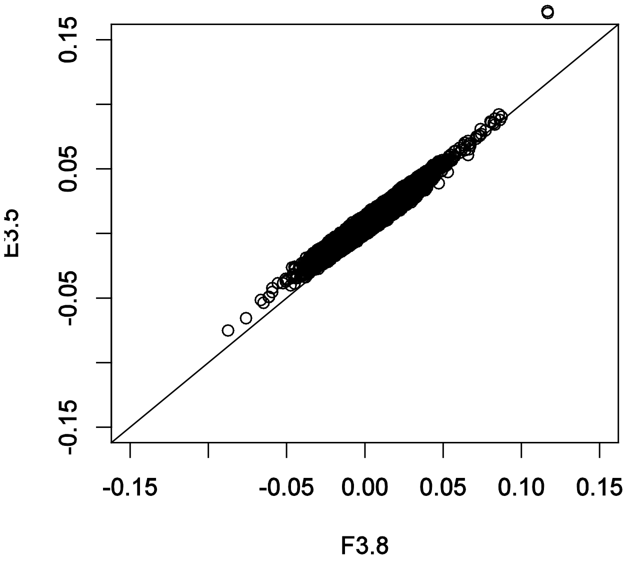

Figure 1. Comparison of marker effects estimated with k=3.8 and fixed known variance with values estimated with k=3.5 and genetic variance estimated (ignoring 4 markers of large effect).

Table 1. Comparison of model statistics from the GBLUP model and for the Fast Bayes-A like model for 3 levels of degrees of freedom and holding the variance ratio fixed, or estimating it.

Degrees of freedom (*k*)	LogL	Residual Variance	Genetic Variance	Accuracy	Large Markers
GBLUP	-6077.5	54.5	44.0	0.611	
FBA-F Genetic Variance ratio fixed at 0.808					
4.2	-6042.6	57.9	46.8	0.635	954/4480
3.8	-6008.1	52.5	42.4	0.656	952/954/4480/5488
3.5	-5995.9	52.0	42.0	0.659	145/952/954/ 2719/4480/5488
FBA-G Genetic Variance estimated					
4.2	-6050.3	53.7	49.7	0.636	4480
3.8	-6033.0	53.9	43.4	0.645	954/4480
3.5	-6004.3	53.8	44.6	0.655	952/954/4480/5488

Table 2. Wald F statistics and fixed effects for 4 markers of large effect in a GBLUP model where the genetic variance was estimated at 26.35 and residual variance at 54.17. The Incremental (Marginal) F reflects the variation explained as markers are added in order (after all others).

Source	Incremental F	Marginal F	Fixed Effect
snp(952)	55.84	12.80	2.24 ± 0.63
snp(954)	14.42	12.90	-2.19 ± 0.61
snp(4480)	60.69	61.51	3.50 ± 0.45
snp(5488)	46.57	46.57	-3.33 ± 0.49

DISCUSSION

ASReml has been widely used for fitting GBLUP models where users have supplied the **G** matrix. Now it can directly make a common form of the **G** matrix, and report marker effects as well as animal effects.

The Bayes-A like models give a better fit to the genetic relationship matrix than the GBLUP model, as indicated by the Log likelihoods, and identify markers of large effect. The number of large effects identified is related to the peakedness of the *t* distribution which is controlled by the degrees of freedom, *k*. There is currently no formal method to choose a value for *k* in this implementation. Sun *et al.* (2012) used 4.2 but 3.5-3.8 seems more appropriate here.

The GBLUP model runs much faster than the FBA model because the **G** matrix is only formed and inverted once whereas in the FBA model it must be formed and inverted each iteration, increasing the time in this example from 40s to 180s per iteration. Therefore, it will generally be more efficient to follow the path Sun took and estimate the genetic variance under the GBLUP model and then use that value as the fixed prior for the FBA-F model. Furthermore, the FBA runs typically required from 20 to 40 iterations for the marker effects to stabilize while the GBLUP run took about 8 to 10 iterations to estimate the variance parameters.

ASReml can fit identified markers as separate (fixed or random) effects. Including the 4 markers identified in the FBA-F model with *k=3.8* as fixed effects and estimating the remaining genetic variance under the GBLUP model produced an estimate 40% lower than obtained in the original GBLUP model.

The FBA implementation is restricted to a single marker matrix on a single trait but the **G** matrix formed can be saved for use in more complex models.

REFERENCES

Gilmour A.R. (2013) *Pers. Comm.* Technical Report, University of Wollongong.

Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001) *Genetics* **157**: 1819.

Stranden I. and Garrick D.J. (2009) *J Dairy Sci* **92**: 2971.

Sun X., Qu L., Garrick D.J., Dekkers J.C.M., and Fernando R.L. (2012). *PLoS ONE* **7(11)**, e49157.

Szydlowski M., and Paczyńska P. (2010) http://www.biomedcentral.com/1753-6561/5/S3/S3.