*Statistical methods I*

# MIXED MODELS IN ANIMAL BREEDING: WHERE TO NOW?

**A.R. Gilmour**

Cargo Vale, CARGO, NSW 2800,
formerly Orange Agricultural Institute, NSW Department of Primary Industries

## SUMMARY

Over the past 60 years, mixed models have underpinned huge gains in plant and animal production through genetic improvement. Charles Henderson (1912-1989) established mixed models for estimating breeding values (BLUP) using the popularly called Henderson's Mixed Model and provided early methods (Henderson's Methods I, II and III) for estimating variance parameters. Robin Thompson then published the widely acclaimed REML method for variance component estimation in 1971. These two innovators, along with the development of computing power, have spawned national and international breeding programs in almost all animal species used for human food and fibre.

Our ability to generate data is outstripping our ability to analyse data and this will lead to mixed models playing new roles in genetic estimation. The focus is changing from simply describing the relationship between variables through a correlation, to modelling the relationship based on knowledge of the Genome.

## INTRODUCTION

Selective breeding goes back at least to Jacob (1800 BC, Genesis 30) who selected the fitter rams for his own flock. Traditional breeding has largely relied on visual assessment with many such classers having considerable skill in recognising genetic potential with respect to their objective, whether breeding war horses, dogs or pigeons. What characterises modern breeding though is the extensive use of objective measurement and adjustment for environmental effects. The digital age has seen a rapid increase in the number of traits included in a breeding objective or selection criterion, as well as use of data on relatives to improve the separation of genetic from environmental differences. Charles Henderson (1912-1989) *et al.* (1949, 1959) developed and popularised the mixed model equations which underpin the BLUP estimation of breeding values. His development of these equations included use of the additive genetic relationship matrix, showing how it accommodates selection as well as their primary role of adjusting for nuisance environmental effects.

Charles Henderson

Robin Thompson

However, the mixed model equations used for evaluation assume knowledge of variance parameters. Henderson (1953) defined the main methods used to estimate these until Robin Thompson (Patterson and Thompson 1971) presented the Residual Maximum Likelihood (REML) method. Karin Meyer and Dorothy Robinson produced software to implement REML methods (in animal breeding and more generally respectively). However analysis was difficult until Robin presented the Average Information method (Johnson and Thompson 1995; Gilmour *et al*. 1995) underpinning ASReml (1997, 2002, 2006, 2009) which become generally available in 1997.

The promise of the genomic revolution is that we may be able to select directly for specific combinations of genes based on reading an individual's genetic code and having good information on the phenotypic and pleiotropic effects of genes/alleles.

## MIXED MODEL EQUATIONS AND BLUP

The linear mixed model is written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{e}$$

where ***X*** is the design matrix for fixed effects, $\boldsymbol{\tau}$ , ***Z*** is the design matrix for random effects, ***u***, ***y*** is the vector of phenotypic measurements and ***e*** is the vector of model residuals. The mixed model equations (MME) are conveniently represented in matrix form by

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\tau} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

where var$\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$ . Given ***R*** and ***G***, the solution for the fixed effects given by the mixed model equations is the same as given by solving $X'V^{-1}X\hat{\tau} = X'V^{-1}$ where $V = R + ZGZ'$. The solutions for the random effects are the Best Linear Unbiased Predictors of those effects and as such are ideal for selecting breeding stock.

The power of this system lies in the structure that can be incorporated into ***X, Z, R*** and ***G***. It is is not unusual for ***u*** to include sub-vectors for various traits and various 'strata' such as direct genetic, maternal genetic, maternal environment, dominance and nuisance blocking effects. This can lead to a fairly complex structure to ***G*** involving relationship matrices and variance matrices of various sorts. The main advantage of the mixed model equations is that the left hand side matrix is typically fairly sparse so that large systems of equations can be solved quite efficiently. This arises because matrices ***X*** and ***Z,*** and inverses of ***R*** and ***G*** are typically sparse.

## RESIDUAL MAXIMUM LIKELIHOOD

Without going into the detail, suffice to say that if we assume ***u*** and ***e*** (and therefore ***y***) are normally distributed (given ***R*** and ***G***), we obtain an expression involving ***y, X, Z, R*** and ***G*** which is called the likelihood. This expression can be partitioned into two parts; one providing information on $\boldsymbol{\tau}$ conditional on ***G*** and ***R*** leading to the mixed model equations, the other providing information on ***R*** and ***G*** conditional on $\boldsymbol{\tau}$. Residual Maximum Likelihood seeks to find the parameter values for ***R*** and ***G*** that are most likely because they maximise this second part (rather than the whole likelihood). This maximisation exercise though was not trivial when ***R*** and ***G*** involved more than a few parameters and the problem was large. Consequently, REML estimation was restricted in application to small problems or well structured standard animal breeding models until Thompson presented the Average Information procedure (Gilmour *et al.* 1995, Johnson and Thompson, 1995) which is also centred around the mixed model equations. The implementation in ASReml exploits the sparsity of the mixed model equations though judicious ordering of the equations, avoiding the need to obtain the complete inverse of the left hand side matrix. Now REML can be applied to large problems (with several hundred variance parameters).

## WHERE TO NOW?

One thing programming has taught me is that no matter how big you allow, someone will want bigger. While computing technology has helped with the more traits, more records issue of modern animal breeding based on BLUP technology, we are now faced with genome level data of a higher magnitude and methodologies which do not have the statistical and mathematical rigor that supports conventional quantitative genetics. Three problem areas come to mind. The first is the well established variance estimation problem (Hill and Thompson 1978) that when estimating a variance matrix, the probability that the maximum value of the REML likelihood occurs outside the imposed parameter space increases with the matrix size. The second is the application of mixed models to genomic data. The third is how to effectively combine specific genomic data into the BLUP evaluation process.

**Structured Variance models.** The more traits involved in a REML analysis, the more likely there will be difficulties with the estimation of all the variances and co-variances involved. ASReml will estimate a negative definite matrix if permitted, or attempt to estimate a positive definite matrix which is almost singular. But this raises the issue of whether a reduced parameterization within the parameter space will be preferable. It is not uncommon to find that a matrix can be reduced by use of principal components to a more parsimonious form. That is, the first 1, 2 or 3 principal components will contain the big bulk of the information contained in the matrix. The remaining variation is noise and is often associated with negative eigen values. Therefore it makes sense to estimate the matrix based on some underlying structure. Three structures are common in ASReml. For variates that have no intrinsic ordering, the principal component/factor analytic models allow more parsimonious modelling. For measurements repeated at irregular intervals, the random regression models are often applied but these may produce unreasonable estimates at the ends of the time range. For regular repeated traits, for example weights at successive ages, the expected structure is an autoregressive one for which the Antedependence (Generalised auto regressive) models apply. Jaffrèzic *et al.* (2002) has extended the Antedependence model to a Structured antedependence where a model is imposed on the regression and innovation parameters. Meyer and Kirkpatrick (2009) have investigated a reduced parameterization based on assuming common eigen vectors across strata which is another proposal within this framework. To my mind, this leads to a general area of writing models for the variance parameters, and is the next logical step when it comes to fitting models with hundreds of variance parameters. The question will always be whether a reduced parameterization has adequately captured the real variation without imposing a structure unsupported by the data.

**Mixed models for genomic data.** There is a huge literature on analysing the huge amount of genomic data that is being presented and little consensus on the best approach. One issue is the diversity of kinds of data available and the other is the sheer volume of data and the knowledge that meaningful/useful variation is present in only a small proportion of it. The issue here is then to separate signal from noise. I believe mixed models could have a bigger role here because signal will represent a covariance (or inflated variance) over the noise (base variance). Mixed models have been successfully used to adjust for spatial variation in genomic slides. They have been used to locate QTL in back-cross/F2 experiments (Gilmour 2007; Verbyla *et al.* 2007) and in association studies where there are often more 'markers' than experimental units. Thomson *et al.* (2009) use mixed models as part of their procedure to combined cattle and sheep genomic data to look for differentially expressed genes. The new outlier method in ASReml 3 may help in this regard.

**Incorporating genomic markers in BLUP evaluation.** Scientists are an optimistic group when it comes to incorporating genetic markers into BLUP evaluation. I suspect there is a lot of detailed work required before this becomes standard procedure across the industries.

## DISCUSSION

Linear Mixed Models have underpinned a revolution in livestock breeding in the last 50 years and despite the huge investment in genomic research and Bayesian methods, there remains a continuing major role for them in the foreseeable future. However, the general model needs adaption for the specifics of each particular species and application. By this I mean, identification of the principle sources of variation, whether they should be accommodated as fixed or random effects, appropriate variance structures and extending the analyses to larger populations and with more traits.

While a bivariate analysis is now readily performed, larger multivariate analyses for the estimation of positive definite variance matrices are often difficult requiring use of structured matrices and raising the issue of whether the structure is adequate. There will undoubtedly be further developments in this area.

The literature on analysis of genomic data reports a wide range of methods as people have hurried to analyse their large amounts of newly acquired data. Some of these analyses have demonstrated the utility of mixed models in this area, but have also shown up limitations due to the amount and structure of the new data. This also will need more attention.

## ACKNOWLEDGEMENTS

Upon my retirement, I am grateful to all my colleagues, to NSW DPI, and to the gracious Lord God, creator, for the opportunity given me to contribute to plant and animal breeding through my part in ASReml, in LAMBPLAN (through the program BVEST), through mentoring and assisting other researchers and in the analysis of breeding data, especially for sheep.

## REFERENCES

Gilmour, A. R. (2007) *Comp Stats and Data Analysis* **51**:3749

Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995) *Biometrics* **51**:1440

Gilmour, A. R., Gogel, B. G., Cullis, B. R. and Thompson, R. (2009) ASReml 3. www.vsni.co.uk

Henderson, C. R. (1953) *Biometrics* **9**:226

Henderson, C. R., Kempthorne, O., Searle, S. R. And Von Krosigk, C. N. (1959) *Biometrics* **15**:192

Hill, W. G. and Thompson, R. (1978) *Biometrics* **34**: 429.

Jaffrѐzic, F., White, I. M. S., Thompson, R., Visscher, P. M. and Hill, W. G. (2002) *7WCALGP*

Johnson, D. L. and Thompson, R. (1995) *J. Dairy Sci.,* **78**:449

Meyer, K., and Kirkpatrick, M. (2009) AAABG

Patterson, and Thompson, R. (1971) *Biometrika* **58**:545

Thomson, P. C., Singh, M., and Raadsma, H. W. (2009) AAABG

Verbyla, A. P., Cullis, B. R. and Thompson, R. (2007) *Theor. And Appl. Genet* **116**:95