VARIANCE STRUCTURES AVAILABLE IN ASREML

A. R. Gilmour

NSW Agriculture, Orange Agricultural Institute, Forest Road, Orange, NSW 2800

SUMMARY

The paper explains why it is often difficult to estimate co-variance matrices from data and describes the variance structures available in ASREML to investigate the problems. Difficulty arises when the assumed model is not correct and because of sampling variation in the matrices.

Keywords: Genetic correlation estimation, average information, REML, variance components

INTRODUCTION

Animal breeders are often interested in estimating variance/co-variance matrices from data. These include genetic and residual matrices between traits and between times which are used to calculate heritabilities and genetic correlations. Restricted Maximum Likelihood (Patterson and Thompson 1971) is now the preferred method and is implemented in many programs including ASREML (Gilmour *et al.* 1997). ASREML uses a quadratic convergence method based on the Average Information (Gilmour *et al.* 1995) matrix and is generally quite efficient.

Assuming we have an appropriate variance model, there is a problem associated with estimating variance component matrices which is independent of the algorithm used to obtain the estimate. It is that sampling variation in the data often supports a non-positive definite matrix for some components, ie. the estimated matrix is outside the theoretical parameter space for the proposed model. Thus, strategies are required to produce estimates which are in the parameter space.

I will consider first why this problem occurs and then how this problem is addressed in ASREML.

HOW NON-POSITIVE DEFINITE MATRICES ARISE

Consider the simple sire model with 2 traits and balanced data. The Error Mean Square (Error MS) matrix can be calculated from sums of squares and cross products of the residuals. Similarly, the Sire Mean Square (Sire MS) matrix can be calculated from the sums of squares and cross-products (weighted) of the sire effects. Both matrices will always be non-negative semidefinite and will be positive definite unless all the residuals/effects are zero for one trait or the residuals/effects for the traits are perfectly correlated. When data is unbalanced, these mean square matrices are not necessarily but usually will be positive definite. However, an uneven distribution and amount of missing data may lead to non-positive definite mean square matrices.

Assuming these mean square matrices are positive definite, the problem arises when we impose an expectation model on them. We might suppose the Error MS matrix is a matrix sampled from a distribution with expected value Σ_E and that the Sire MS matrix is sampled from a distribution with expected value $\Sigma_E + k\Sigma_s$. Then we can estimate these component matrices (given we can calculate *k,* the number of progeny per sire in the balanced case) as simple functions of our observed matrices. While our estimates are unbiassed if the model is correct and the estimate of Σ_E is positive definite if the Error MS is positive definite, the estimate ofΣ_S will often not be positive definite. This arises because both mean square matrices are sampled with error even if our expectation model is correct.

Several factors contribute to this result. Sometimes there will be unidentified sources of variation in the data which inflate the Error MS and therefore reduce the estimate of Σ_S . Further, the sampling variation is increased if the sample size is small. Increasing the number of traits increases the chance that the estimate of Σ_S will not be positive definite. Finally, when correlations are high in the mean square matrices, correlations implicit in Σ_S are more likely to have magnitude greater than 1.

For the balanced model described above, the method of equating mean squares to their expectations produces REML estimates of the variance components when the estimates are in the presumed parameter space. However, we generally use specific REML software for estimating the components because it can handle the unbalance introduced by unequal replication, missing values and the fitting of other terms in the model to remove extraneous variation.

We are uncomfortable with negative variances (although they are not always invalid) and correlations with magnitude greater than 1. Furthermore, some REML algorithms fail when the matrix cannot be inverted so that we get no answer at all! Therefore we need a strategy to produce estimates within the parameter space or at least on the boundary. One strategy is to collect more data so that the sampling variance of the mean square matrices is reduced. Another strategy is to review the model being fitted in case there is a more appropriate model. ASREML provides two other strategies. One is to fit a more robust variance structure with fewer parameters. Some alternative parameterisations are discussed in the next section.

The other strategy is to restrict the parameter values to the parameter space: variances must be positive, correlations must be less than 1.0 in magnitude and matrices must be positive definite. It is relatively easy to fix individual variances and correlations near the boundary if the parameter updates would take them out of bounds. Forcing a matrix to be positive definite can be achieved by bending (drawing the eigen values towards their mean until all are positive) but this does not work as well as might be hoped when iterating. Thus, using more parsimonious variance matrix structures is often a better strategy. We now look at some alternative structures and comment on when they are useful.

SOME VARIANCE MATRIX STRUCTURES

For the residuals ordered traits within units, the typical variance structure for the residuals is given by the direct product $\mathbf{I} \square \Sigma_E$ of an identity matrix for independent units by an unstructured variance/co-variance matrix across traits. For the sire effects, assuming sire.trait (traits nested within sires) and a genetic relationship matrix among sires of A, the variance structure would be $\mathrm{A} \square \Sigma_S$. ASREML can form the inverse relationship matrix if a pedigree file is provided or it can read one which the user might create assuming some other relationship model, for example a gametic relationship matrix.

Unstructured (US). The basic form of a variance matrix is the unstructured matrix (US) which is symmetric but each variance and co-variance is a separate parameter. Therefore it has $n(n+1)/2$ parameters for a matrix of order n. This is usually reasonable for n small but quickly seems over-parameterised as n increases. ASREML typically assumes Σ_E (referred to as an R structure) has this form in multi-variate (and repeated measures fitted as multi-variate) analyses. When used as a model for Σ_S (referred to as a G structure), ASREML sometimes fails when the estimate becomes non-positive definite because the algorithm depends on being able to invert the matrix. If requested, ASREML will bend the matrix to make it positive definite. Otherwise, it might converge to a negative definite solution.

Homogeneous variance correlation models σ^2C. For these models, we specify a correlation structure **C** and a scale parameter (σ^2) for the whole matrix. The most commonly used forms of **C** are the Identity (zero correlation), the uniform correlation model and various low-order auto-regressive moving average correlation models. These are often plausible models which usually produce solutions in the parameter space.

Heterogeneous variance correlation models SCS. For these models, we specify a correlation structure **C** as before and a diagonal variance matrix. The correlation matrix is then pre and post multiplied by the square root (**S**) of the variance matrix to generate a matrix with the specified variances. Sometimes, it is helpful to fit this model to obtain starting values for fitting the US model.

Factor analytic models. ASREML has two forms of the factor analytic model. The standard form is a matrix **SCS** where **S** is a scaling matrix as before, **C** = **FF'+E** is a correlation matrix, **F** is a matrix defining the 'factors' (similar in concept to principal components) and **E** is a diagonal matrix such that the diagonal of **FF'+E** is 1. Conceptually, for **E** to have positive values (representing trait specific independent variation), the diagonal of **FF'** must be constrained to be less than 1. Typically **F** will have 1 or 2, maybe 3 columns which will soak up most of the correlation structure.

The second form of the factor analytic model omits **S** so that the scale components are built in to **F** and **E**. In this model, **E** and **F** are estimated rather than **S** and **F**. The constraint is then to require the elements of **E** to be positive.

The latest model considered for ASREML is the reduced rank model **FF'** which is related to the factor analytic models with **S=I** and **E=0**.

Cholesky decomposition LDL'. The Cholesky decomposition as implemented in ASREML has not proved to be a useful parameterisation. **L** is a lower triangle matrix with 1 on the diagonal and **D** is diagonal. In its full form, it has as many parameters as the **US** structure but they are progressively row-wise dependent which leads to convergence problems. ASREML allows for high lag elements to be zero but such a banded variance matrix is rarely required. Another parameter reduction is to fix elements of **D** to be very small and positive (almost zero) and the corresponding columns of **L** to be zero. This mimics the reduced rank model described in the preceding paragraph.

Antedependence decomposition UDU'. This model is parametized in terms of the inverse of the Choleski decomposition and can be viewed as a generalization of the heterogeneous autoregressive (AR) model. ASREML allows bands of **U** to be set to zero. The model is particularly appropriate in a growth model where each 'trait' represents a random addition to the previous one(s). Unlike the AR model, it does not assume the 'traits' are equally spaced or have constant variance of the additional growth.

Covariance functions and splines. Another approach to variance modelling is to think of transforming the design matrix (**Z**) for the random effects. This is behind the use of co-variance functions and cubic splines, particularly in a repeated measures context. A change of the design matrix implies a complementary change to the G matrix as it changes the meaning of the random effects fitted in the model.

CONCLUSIONS

The problem of parameter estimates being outside the assumed parameter space may arise from specifying inappropriate variance models but usually is caused by sampling variation. The three strategies available in ASREML are to accept the estimates anyway, to fix offending parameters close to the boundary and to fit a less parameterised structure which may overcome the problem. By providing a hierarchy of variance model structures, ASREML allows plausible 'smoothed' models to be fitted in the general multi-variate variance component estimation situation. The uniform correlation and factor analytic models have proved very useful in a variety of contexts, especially where the dominant components of the structure are of interest.

REFERENCES

Gilmour, A.R., Cullis, B.R., Welham, S., and Thompson, R. (1997) *ASREML*, NSW Agriculture
Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995) *Biometrics* **51:** 1440
Patterson H.D. and Thompson, R. (1971) *Biometrika* **58**: 545