Variances and Components

Understanding Linear Mixed Models

Arthur R Gilmour
former Principal Research Scientist (Biometrics), NSW Agriculture,
email:arthur.gilmour@cargovale.com.au

January 18, 2019

Contents

Preface

This is a user support document prepared to assist users of Echidna (and ASReml).

Echidna is available free for non-commercial use from `www.EchidnaMMS.org` where you can also download this document.

Please send feedback to Dr Gilmour (arthur.gilmour@cargovale.com.au).

It is the glory of God to conceal a matter,
But the glory of kings is to search out a matter. [1]

[1] Solomon, Proverbs 25:2

1 Introduction: Variances and Variance Components

1.1 Mean and Variance

The aim of this document is to provide a heuristic introduction to fitting linear mixed models in ASReml and Echidna. We will loosely define terms such as population, mean, variance, variance components, fixed and random effects.

When we collect data, we obtain a sample of values from a population. We may be interested in the particular individuals sampled, or in understanding something about the population. If our interest is the population, then we need to sample from the population in an unbiased way.

In linear mixed models, we are particularly interested in the mean and the variance. We will use minimal algebra but we need some. The mean, or expected value, will be represented by μ (MU) and tells us the most likely value (although this value will probably rarely be actually sampled). The variance tells us how variable the values are, the spread of the sample values around the mean value) and will be represented by σ^2 (Sigma squared).

These quantities will rarely ever be known but what we do know are a series of n sample values, $y_i, i = 1, n$.

The estimate of μ we typically use is $\hat{\mu} = \Sigma_{i=1}^{n} y_i/n$, the sum of our sample, divided by the number of observations. Other options are the median, the value which splits the sample into lower and upper halves, and the mode which is the most frequent value. These values will differ if the distribution is skewed, as say for household income.

The estimate of σ^2 we typically like to use is the averaged squared deviation from the mean given by $\hat{\sigma^2} = \Sigma_{i=1}^{n}(Y_i - \mu)^2/n$. But since we do not know μ, we need to use $\hat{\sigma^2} = \Sigma_{i=1}^{n}(Y_i - \hat{\mu})^2/(n-1)$.

1.2 Sources of Variation

The variation in our sample values has many causes. We cannot discern most of them, but we can identify some, and this leads us to partitioning the variance into components. This in turn leads us to the question which sources are important. Sample characteristics like age, sex, ethnicity, education level are obviously important for some characters. When you see data in a text book, you only have the values given you, but in the research world, you will collect the data yourself, and an important issue is then to identify as many potentially important covariables as possible. In some cases, you will impose treatments to see if they effect the response. In other cases, you will control the experiment so as to minimise extraneous sources of variation.

1.2.1 Sire model

The usual motivation for this model is to ascertain how much of the variation is due to genetics. The HARVEY dataset has calf weights of 65 calves sired by 9 bulls. The data set also records information on the age of the cow (DamAge) and the Line (maybe of the cow). We can therefore calculate an analysis of variance table

Source of Variation	DF	Sum Squares	Mean Square	F-Ratio	Expected Mean Squ
Mean	1		2028339.5	15262.1	
Line	2	4454.6	2227.3	16.76	$\sigma_e^2 + k\sigma_s^2 + L$
DamAge	1	46.5	46.5	0.35	$\sigma_e^2 + A$
Sire	6	2033.4	338.9	2.55	$\sigma_e^2 + k\sigma_s^2$
Residual	55	7309.5	132.9		σ_e^2

Since this is not a balanced set of data, we leave calculation to the software. This model was fitted as a fixed model and the software just reports the F-Ratios and the residual variance. Hand calculation would proceed by calculating the Sum of Squares first. The model was written as

```
ADG ~ mu Line DamAge Sire
```

and the model fit summarized as

```
Model_Term                          Gamma         Sigma   Sigma/SE   % C
Residual               SCA_V   65   1.00000       132.895       5.24   0 P

                                Wald F statistics
   Source of Variation          NumDF     DenDF    F-inc            P-inc
 7 mu                               1      55.0 15262.07            <.001
 4 Line                             2      55.0    16.76            <.001
 5 DamAge                           1      55.0     0.35            0.557
 2 Sire                             6      55.0     2.55            0.030
```

1.2 Sources of Variation

In this case, the Variance attributed to each term is termed 'Incremental'. The process involves first sweeping out the variation attributed to the mean ($176.65^2 \times 65$), then Line, then DamAge, then Sire leaving 7309.5 unexplained. Since the terms or not orthogonal, changing the order would change the variance explained by each term. This issue will be discussed later.

The final column gives 'Expected Mean Squares' in terms of variance components (σ_e^2 and σ_s^2 and fixed effect deviations given simply as A and L. If there is no variance attributable to sires, (that is, σ_s^2 has a true value of zero), then the Sire Mean Square is simply another estimate of the residual variance and an F-test will show it is not significant. Indeed an F-ratio of 2.55 tested with 6,55 degrees of freedom has a probability 0.03 so is significant at the 5% level.

Our interest in this model was to calculate heritability. For that, we need the value of k so that we can calculate $\hat{\sigma_s^2}$ =(SMS-RMS)/k. k is the effective number of progeny per sire (roughly 65/9) but since the actual numbers vary between 5 and 9, we let the software calculate it: $k = 7.2$, $\hat{\sigma_s^2} = 28.9$. Heritability is therefore calculated as $4\hat{\sigma_s^2}/(\hat{\sigma_e^2} + \hat{\sigma_s^2}) = 4 \times 28.9/(132.9+28.9) = 0.71$. The 4 comes in because sires represent only half the genetic variance.

The REML method provides an alternative way to calculate things. For this, we indicate that sire is to be fitted as random effects. The model is written as

```
ADG ~ mu Line DamAge !r Sire
```

and the model fit summarized as

```
        Approximate stratum variance decomposition
Stratum      Degrees-Freedom   Variance      Component Coefficients
Sire                    5.93    341.035          7.2     1.0
Residual Variance      55.07    132.756          0.0     1.0

Model_Term                          Gamma          Sigma   Sigma/SE   % C
Sire                  IDV_V     9  0.217651        28.8946       1.04   0 P
Residual              SCA_V    65   1.00000        132.756       5.25   0 P

                                  Wald F statistics
   Source of Variation            NumDF     DenDF    F-inc            P-inc
 7 mu                                 1       5.9  5906.96            <.001
 4 Line                               2       5.9     6.19            0.035
 5 DamAge                             1      57.8     0.62            0.435
```

Notice that this reports the k coefficient in the stratum variance decomposition and regenerates the sire mean Square using that coefficient.

The other thing to note is the Denominator DF for Line is reported as 5.9. This is because Sires are nested within Lines so that if we wanted to test Lines, it would need to be relative to the Sire Variance (which has 6 DF). This is why the F-ratio for LINES is now 6.2 rather than the 16.8 reported in the fixed analysis. If this was your data, you had designed the trial, you would/should have known that sires were nested in Lines, and had you wanted to test for Line differences, would recalculate the F statistic for Lines in the fixed analysis as 6.57 and tested it with 2,6 degrees of freedom. We do not know whether Lines is a classification of the sires and/or of the Dams. However, structurally, lines are a classification of sires because each sire is associated with only one Line.

Even so, there are small differences between the two analyses. This is because the REML analysis fits the sire effects better accounting for the unequal numbers of calves per sire.

1.3 Fixed vs Random

REML estimation of Variance components is based on the linear mixed model written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\eta}$$

and by solving the mixed model equations given as

$$\begin{pmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z + G^{-1}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}y} \end{pmatrix}$$

where $\boldsymbol{y}$ is the known vector of n observations, $\boldsymbol{X}$ is the design matrix for fixed effects $\boldsymbol{\beta}$, $\boldsymbol{Z}$ is the design matrix for random effects $\boldsymbol{u}$, $\boldsymbol{R}$ is $\mathtt{Var}(\boldsymbol{\eta})$ and $\boldsymbol{G}$ is $\mathtt{Var}(\boldsymbol{u})$.

That is, with the assumption of normality, $\boldsymbol{y} \sim \mathtt{N}(\boldsymbol{X\beta},\ \boldsymbol{R + ZGZ'})$.

In the simplest case, $\boldsymbol{R}$ and $\boldsymbol{G}$ are simply scaled Identity matrices, scaled by σ_e^2 and σ_u^s respecively, as in the Sire Model above. However, for many models fitted in AS-Reml/Echidna, they can be quite complex structures.

This is not the place to delve into these details. We simply note that we observe $\boldsymbol{y}$ and collect other information from which we hypothesize the linear model given above. Implicitly we assume we know the correct $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{R}$ and $\boldsymbol{G}$ so that we can solve the mixed model equations for $\boldsymbol{\beta}$ and $\boldsymbol{u}$.

In fact, we hypothesize a model which defines $\boldsymbol{X}$ and $\boldsymbol{Z}$ and the form of $\boldsymbol{R}$ and $\boldsymbol{G}$ and use the REML algorithm to find suitable parameter values for $\boldsymbol{R}$ and $\boldsymbol{G}$. The final solutions for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are often called *empirical* because they are based on estimated $\boldsymbol{R}$ and $\boldsymbol{G}$ matrices.

Since there are so many assumptions tied up in this process, it is likely that different

statisticians will end up with different models for the same data, or similar data. We will consider comparison among models latter.

1.3.1 Random effects are shrunken

Going back to our sire model, we could define a sire effect as (the sum of observations for that sire) divided by (the number of those observations) less the mean, and could write that as $\hat{\beta}_i = \Sigma_{j=1}^{n+i}(y_{ij}-\mu)/n_i$. This is known as the Best Linear Unbiased Estimate (BLUE) of the sire effect and it uses all the information on the sire equally. However, if our purpose is to predict future performance of progeny from this sire, and we want to select among several sires which have unequal amounts of information (say one is estimated from 10 progeny and another from 1000 progeny) then a better selection criterion is the Best Linear Unbiased Predictor (BLUP) of the sire effect given by $\tilde{u}_i = \Sigma_{j=1}^{n+i}(y_{ij} - \mu)/(n_i + \sigma_e^2/\sigma_u^s)$. This is a shrunken value (compared to $\hat{\beta}_i$) because we have added a function of the heritability to the divisor (assuming the variance components are both positive). If the genetic variance is small, there is more shrinkage than if it is large. If n_i is relatively small, there is more shrinkage than if it is large. So, if the sire variance is 10% of the residual variance, the divisors for samples of size 10 and 1000 change from 10 and 1000 for the BLUE to 20 and 1010 for the BLUP. That is, the sire effect based on 10 progeny is reduced 50% while the sire effect based on 1000 progeny is reduced by 1%.

1.3.2 Fixed or Random

One criterion for choosing whether to fit a factor as fixed or random is the role of a term in the model, whether you are interested in the effects and whether that interest relates to describing the past or predicting the future, whether it relates to the individual or to a class the individual represents.

1.3.3 Recovery of interblock information

The REML algorithm was first published (Thompson 1971) as a general method for recovering interblock information.

In agricultural field experiments, it was early recognised that two plots sown to the same cultivar did not give the same yield. So, to estimate cultivar/treatment effects, multiple plots of each could be sown in a Completely Random Design. The cultivars need to be allocated randomly to the plots to avoid bias in the estimates.

However, part of this discepancy is associated with patches in the field; some parts tend to yield more than other parts. To try and manage this, the field was divided into Blocks and plots within plots. Then, the trial was sown so that each block contained 1 plot of each cultivar/variety/treatment. This is know as a Randomised Complete Block Design.

However, it is nearly impossible to predict where to best place the blocks so that plots within blocks were most similar. Therefore, use of smaller blocks was proposed. These

were known as Incomplete Block designs and the whole discipline of design efficiency was developed.

The problem remained that if you followed a sweep algorithm to estimate treatment effects, you would adjust the data first for the block effects, then estimate treatment effects from the block adjusted data. However, the block effects, being incomplete blocks, contained some of the treatment effect. They over adjusted. The REML method was then devised to estimate the appropriate shrinkage of the block effects so as to most efficiently estimate the treatment effects. The best shrinkage turned out to be the BLUP of the block effects utilising the ratio of residual variance component to block variance component.

1.4 General Variance Models

Mixed models can be viewed as a linear model with correlated residuals. Here we could fit a model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\texttt{Var}(\boldsymbol{\epsilon}) = \boldsymbol{R} + \boldsymbol{ZGZ}'$. In this context the random terms in the model simply define a correlation structure among residuals. We still need to estimate the variance components because they define the correlations.

1.4.1 Animal model

The sire model has the disadvantage that it only considers one genetic relationship, a covariance due to having the same sire. However, we expect there will also be correlations due to a common dam, a common environment and other genetic relationships (common maternal grandsire). This leads us to the genetic relationship matrix, $\boldsymbol{A}$, derived from a pedigree of animals. This matrix, often called the Numerator relationship matrix (NRM) is relatively easy to compute by building the line for an individual by averaging the lines for its parents and adjusting for any inbreeding.

In fact, for REML we need the inverse of this matrix and that turns out to be a sparse matrix so it is relatively easy to analyse data with a large pedigree. Again, our purpose may be to estimate the genetic value of the individuals, or to estimate the variance components.

Any such analysis needs to also adjust for known sources of information which contribute to the variation in the response variable. For example, when analysing birth weight or weaning weight, we may include the parity of the dam, a 'Herd-Year-Season' effect since season conditions and management effects will affect the responses, maternal genetic and non-genetic effects, litter effects, birth/rearing rank.

This kind of genetic model can be applied to trees but the population structure is different. Trees will often be planted in say 5 full-sib trees per plot. With trees though, there can be spatial plot effects (elevation, aspect, changes in soil structure or moisture) which need to be adjusted for. Fitting a tree model with spatial variation also requires the fitting of a non-genetic tree effect (equivalently an uncorrelated tree residual effect) so that the genetic effect is not overstated. A reduced animal model is a good alternative to an animal

model for tree data.

1.4.2 GBLUP

We have discussed use of a pedigree based relationship matrix. However, now we have the possibility of assessing the genetic relationship based on analysis of the genome. There are many Single Neucleotide Polymorphisms (SNPs, or markers) in the genome, that is an identifiable string of genetic letters (A,C,G,T) in which just 1 letter is altered. We do not need to know the parentage of an individual to read its SNP sequence. Therefore, for any pair of individuals, we cannot tell whether that are identical because they have the same parents or not. However, closely related individuals will tend to be identical by descent relative to unrelated individuals. So the average degree of identity across many SNPs predicts genetic relatedness. Given a matrix $\boldsymbol{M}$ coded 0/1/2 with individuls in rows and markers in columns, we can calculate a marker based relationship as $\boldsymbol{K} = \boldsymbol{M}\boldsymbol{M}'/s$ where s scales the matrix to a scale similar to the $\boldsymbol{A}$ matrix. Thus, if we do not have a pedigree but do have marker information, we can still do a genetic analysis either to estimate heritability, or to predict genetic potential. `GBLUP` is the name commonly used for such an analysis.

1.5 Direct Product Structures

In introductory statistics, the variance is a simple quantity quantifying the spread of observations about a mean. But this description assumes the observations are independent (uncorrelated) and have the same variance. In fact, measurements of any kind are rarely uncorrelated and often do not have the same variance. Nevertheless, independent with equal variance is a convenient starting point and we can write $\text{Var}(\boldsymbol{y}_1)$ as the matrix $\sigma_1^2\boldsymbol{I}$. Now consider another set of measurements on the same individuals listed in the same order also with equal (but different) variance $\text{Var}(\boldsymbol{y}_2) = \sigma_2^2\boldsymbol{I}$. If we plotted $\boldsymbol{y}_1$ against $\boldsymbol{y}_2$, we will likely find an association which we can call the covariance $\sigma_{12}\boldsymbol{I}$. If we wrote the numbers out, we would probably use a table where we could label the rows by the names of the subjects/units/patients/animals/plots/whatever and the columns by some label that described the measurement/trait. Or we could stack the variables into a single column and have two labels for each row (the unit and the trait). The mathematical function that converts a matrix to a column by stacking the rows below each other is *vec()*. So we can write $\text{Var}\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix} = \begin{pmatrix}\sigma_1^2\boldsymbol{I} & \sigma_{12}\boldsymbol{I}\\ \sigma_{12}\boldsymbol{I} & \sigma_2^2\boldsymbol{I}\end{pmatrix} = \boldsymbol{\Sigma}\otimes\boldsymbol{I}$ where $\boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}$. Thus, the variance of the combined vector can be written as a direct product of a variance matrix pertaining to the columns ($\boldsymbol{\Sigma}$) and a variance matrix pertaining to the rows ($\boldsymbol{I}$).

This principle can be extended to any matrix of effects and is widely used in mixed models. We use it for plots laid out in a grid, for different traits measured on the same individuals, for repeated measures (the same measurement taken at different times) etc. In fact, whenever a set of effects naturally form a table, it is likely that the variance matrix for those effects can be written as a direct product of a structure related to columns and a structure related to rows.

There are a few considerations. The main one is that only one dimension can be defined in terms of variances, the others must be in terms of correlations or some other known relationship (as in marker based or pedigree based relationship matrices).

1.5.1 Multivariate Analysis

As discussed above, for multivariate data, the residual variance is $\boldsymbol{I} \otimes \boldsymbol{\Sigma}_E$. Usually $\boldsymbol{\Sigma}_E$ is a general (unstructured) symmetric variance matrix. Similarly, for random genetic effects fitted say as `Trait.animal`, the variance matrix is $\boldsymbol{\Sigma}_G \otimes \boldsymbol{A}$ where $\boldsymbol{\Sigma}_G$ is the genetic variance matrix and $\boldsymbol{A}$ is the relationship matrix for the individuals. While ideally we would like $\boldsymbol{\Sigma}_G$ to have the same form as $\boldsymbol{\Sigma}_E$, an unstructured symmetric variance matrix, practically, such a matrix cannot be reliably estimated because of sampling variation, especially when there are many traits. We can then use a form with fewer parameters such as the factor analytic structure discussed below.

The code for a model of this type is

```
wwt yyt gfw ~ Trait Trait.Flock !r us(Trait).nrm(animal)
residual units.us(Trait}
```

1.5.2 Spatial Model

Ultimately, this problem led to the so called spatial modelling of field effects. Instead of using a block based error variance, we use a proximity based correlation structure for the residuals. While there are many possible forms for this, a practical model is the separable autoregressive model. It is computationally efficient compared to many others and accommodates many forms of spatial variation. In particular, it does not require the artificial delineation of plots into blocks in the field. However, it is not recommended we discard all the design considerations of the past. Indeed block designs and row/column designs are still used for determining field layout, although fitting an autoregressive residual correlation structure usually sweeps out the block effects that might be fitted.

In field trials, row effects, column effects, changes in soil type, cultural operations (sowing, plot trimming, harvesting), previous land use all contribute to correlation among plots. Often, we cannot identify the source and that is when the autoregeressive spatial model becomes useful. It simply assumes plots closer together will be more highly correlated. More formally, the yield of a plot is modelled as a proportion of the residual from the immediate neighbour plus a new independent effect. However, as fitted, we do not need to assume one plot was first and its neighbour second. We just estimate the correlation of neighbours raised to the distance between plots. Since plots are usually rectangular and cultural operations are usually applied row-wise or columnwise, we estimate a separate correlation for rows and for columns.

Spatial analysis (Gilmour *et al.*, 1997) uses $\boldsymbol{R}$ to account for correlation among residuals. Since field plots are laid out in a grid, they have a tabular structure and an appropriate

variance structure can be defined as a variance (σ^2) times a row structure ($\boldsymbol{C}_R$) times a column structure ($\boldsymbol{C}_C$) where $\boldsymbol{C}$ is a correlation structure. A separate correlation structure is used for rows and columns because plots are typically rectangular so the distance from plot midpoints is greater in one dimension than the other. A simple form of the correlation structure is first degree autoregressive. This is a process where the yield of a plot is a proprtion of its immediate neighbour plus some extra variation. The process as implemented is not directional in that the same correlation structure is produced regardless of whether we start from the left or the right. Given a lag 1 correlation coefficient of ϕ, the correlation between any to points is ϕ^d where d is the distance. While this structure is dense, its inverse is tridiagonal (i.e. quite sparse). So the spatial residual model for a field trial is typically $\sigma^2\boldsymbol{C}_R \otimes \boldsymbol{C}_C$ and is specified to ASReml/Echidna as

```
yield ~ mu + variety
residual ar1(Row):ar1(Column)
```

more formally specified as

```
yield ~ mu + variety
residual ar1v(Row):ar1(Column)
```

1.5.3 Genotype by Environment

In plant breeding, we need to evaluate genotypes across environments and are looking/hoping for a consensus of genotype rankings across the environments. Here we have separate $\boldsymbol{R}_i$ matrices for each trial and complex genetic relationships differentially expressed across the environments modelled as a direct product of an across environment covariance matrix and a genetic relationship matrix. For example, given 18,432 yields from 64 experiments across 12 YrLoc representing a pedigree of 5132 genotypes.

```
!ASSIGN ExSet 1 7 9 11
!ASSIGN BlSet 2 3 4 5 7 8 9 10 11 12
!ASSIGN RwSet 1 4 7 9 11
!ASSIGN ClSet 1 2 3 4 5 6 7 8 10 11 12

yield ~ mu YrLoc mv !r xfa1(YrLoc).nrm(Geno)  xfa1(YrLoc).ide(Geno) +
  at(YrLoc $ExSet):expt + at(YrLoc $BlSet):expt:Block +
  at(YrLoc $ClSet):expt:Col + at(YrLoc $RwSet):expt:Row
residual at(expt).ar1(Col).ar1(Row)
```

Since there are 12 YrLoc levels, an unstructured genetic variance would have 78 parameters to estimate for each of the additive genetic and nonadditive strata. This will rarely be feasible because of sampling variation. We have therefore used a factor analytic 1 structure (`xfa1()` discussed below) which estimates just 24 parameters for each strata.

1.5.4 Factor analytic models

Estimation of an unstructured (fully parameterised) variance matrix constrained to be positive (semi-) definite often presents problems. A Parameter Expanded EM algorithm (Lui et al., 1998) is implemented in ASReml for this case but does not quickly produce an acceptable result when the maximum likelihood corresponds to parameter values outside the parameter space. The factor analytic variance structure is given by $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi}$ where the columns of $\mathbf{\Gamma}$ are referred to as loadings which relate the k variables to f latent factors; $\mathbf{\Psi}$ is a diagonal matrix of k *specific variances*. The factors are analagous to principal components typically constrained to be orthogonal.

This model can be understood as an extension of the compound symmetry model fitted say as `Genotype + Site:Genotype`. The factor analytic model is equivalent to compound symmetry when $\mathbf{\Gamma}$ has one column with all loadings equal (being the square-root of the `Genotype` variance component) and all specific variance are equal (being the `Site:Genotype` variance component). This structure has proved a useful alternative to the fully parameterised formulation which is often over parameterized. The factor analytic structure is easily made positive (semi-) definite by restraining the specific variances to be zero or positive.

Thompson et al. (2003) present a sparse implementation of this model writing the variance structure and its inverse as $\begin{pmatrix} \mathbf{\Psi} + \mathbf{\Gamma}\mathbf{\Gamma}' & -\mathbf{\Gamma} \\ -\mathbf{\Gamma}' & \boldsymbol{I} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{\Psi}^{-1} & \mathbf{\Psi}^{-1}\mathbf{\Gamma} \\ \mathbf{\Gamma}'\mathbf{\Psi}^{-1} & \boldsymbol{I} + \mathbf{\Gamma}'\mathbf{\Psi}^{-1}\mathbf{\Gamma} \end{pmatrix}$. To use this extended formulation, the design matrix with k columns needs to be extended to include f zero columns for the factors. When some specific variances are zero, Thompson et al. (2003) show how the zero Ψ rows can be collapsed into the factor rows so that all loadings can be estimated.

The advantages of this model are: 1) that it can be built up with first 1 factor, then 2, and more if needed; 2) that the elements of $\mathbf{\Psi}$ may be zero leading to a possibly singular matrix (correlation of 1); 3) while sometimes fragile, the factor analytic model is more robust than the fully parameterised model in those cases; 4) in most cases, the bulk of the covariance is picked up with just a few factors leading to a parsimonious model when k is large; 5) the sparse formulation runs substantially faster than the variance structures with a dense inverse. More recently, it has been appreciated that fitting this model as a completely reduced rank FA plus a diagonal term (`rr1(Site):Genotype + diag(Site):Genotype`) is sometimes even faster (more sparse) formulation. However, it all depends on the ultimate sparsity pattern.

Table 3 reports some recent timing comparisons in ASReml 4.2 for an analysis performed by Alison Smith. The model fits AR×AR spatial variation to 123 experiments evaluating 3755 genotypes in 59 environments. Additional blocking factors `RowBlock`, `ColBlock`, `expt`, `expt:Row`, `expt:Col` were fitted in environments as needed. The genetic model as fitted as `rr`k`(Env).name + diag(Env).name` for the RR+D runs and `xfa`k`(Env).name` for the XFA runs with k taking values 1:6. The XFA1 model reports 37,608 data records,

Table 1.1: Timing comparison for 6 models fitted to

Model	Degree	Size of $\boldsymbol{C}$	Size of $\boldsymbol{C}^{-1}$	Order(secs)	Iteration (secs)
RR+D	1	3,723645	7,158653	20	18(13)
RR+D	2	5,032754	11,928955	37	45(32)
RR+D	3	6,532998	19,243205	61	122(80)
RR+D	4	8,224377	26,794285	90	218(135)
RR+D	5	10,106891	32,994921	98	277(180)
RR+D	6	12,180540	43,982611	129	461(287)
XFA	1	1,72921	8,527122	50	37(25)
XFA	2	1,962101	9,211045	72	44(30)
XFA	3	2,206574	9,965900	75	55(40)
XFA	4	2,483050	10.825425	79	67(50)
XFA	5	2,802036	11,796371	84	85(64)
XFA	6	3,131117	12,272034	80	93(73)

The times are CPU time with Wall Clock times in parentheses.

544,598 equations and 482 variance components. In this instance, the RR+D formulation is decidedly slower than the XFA formulation for cases with more than 2 factors.

There are two issues with FA models. 1) When there are multiple loadings, singularities appear in the $\boldsymbol{\mathcal{I}}$ matrix. Historically we set some loadings to Zero but in fact it is apparently better to keep the loadings close to orthogonal and just not update loadings that would have been set to zero; I have had less convergence issues with these models since adopting this approach, 2) The $\boldsymbol{\mathcal{I}}$ matrix rows related to loadings can be ill-conditioned, the loading parameters sometimes being strongly (negatively) correlated. This is overcome by using a ridge regression technique, testing for poor condition and inflating the diagonal elements of $\boldsymbol{\mathcal{I}}$ pertaining to the loadings by 1% or more.

1.6 Hypothesis testing

1.6.1 Wald F statistics

Testing for fixed effects is sometimes required but is not straight-forward in mixed models. The change in residual sum of squares due to adding a fixed term in the model is a sum of squares that can be divided by its degrees of freedom and the residual variance to produce a Wald F statistic. The issue is that the value of this statistic depends on the order that terms are fitted in the model. Typically a test for a model term is not considered valid if it is marginal to a significant higher order term. To test all possibilities using only simple incremental F statistics will often require judicious ordering and several runs.

ASReml reports two Wald F statistics. The first is the simple incremental F value based on the sequential adding of model terms in the order specified in the model formula. However, this is really only valid for testing the final terms in the list that are not significant. The second F statistic is described as the Maximal Conditional Incremental (MCI) F statistic.

1.6 Hypothesis testing

Model terms are classified into groups such that all the terms in a group are marginal to a term in a higher group, but not to any terms in the same group or a lower group. For example, fitting model terms `mu + A + B + C + D + A:B + A:C + A:D + B:C + B:D + A:B:C + A:B:D` results in 3 groups (main factors, first order interactions and second order interactions). The MCI F statistic for terms in the last group tests these terms as if they were fitted last. The MCI tests for `A:C` and `B:C` are valid for testing these terms if the term it is marginal to (`A:B:C`) is not significant. Similarly, the MCI tests for `A:D` and `B:D` are valid for testing these terms if `A:B:D` is not significant, and the MCI test for `A:B` is valid if neither of the three-way interations are significant. Using the two F-statistics together, the number of runs required to thoroughly test all terms is greatly reduced. The formation of groups is not just based on the base factors but also whether a term removes degrees of freedom from another term. For example, in a model `mu Region Location`, `Region` cannot be in the same group as `Location` because locations are nested in regions.

Formal testing of Wald F statistics in a mixed model depends on the calculation of the denominator degrees of freedom. ASReml uses numerical derivatives to calculate the denominator degrees of freedom using the methodology of Kenward and Roger (1997). However, this is expensive if there are many variance parameters as an extra half iteration (steps 1-5) are required for each parameter.

Bibliography

Gilmour, A. R., Cullis, B. R., & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics* **2**, 269–293.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2006). *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, www.vsni.co.uk.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J., & Thompson, R. (2015). *ASReml User Guide Release 4.1 Functional Specification*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, www.vsni.co.uk.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information reml, an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450.

Kenward, M. G. & Roger, J. H. (1997). The precision of fixed effects estimates from restricted maximum likelihood. *Biometrics* **53**, 983–997.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–54.

Thompson, R., Cullis, B. R., Smith, A., & Gilmour, A. R. (2003). A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Australian and New Zealand Journal of Statistics* **45**, 445–459.