Chapter 9
Modeling and Variance structures

Julius van der Werf

In this lecture we will look at the different sources of variation, how they are associated with model factors and how these together determine the variance structure of the data. Understanding various ways to define a variance structure is important for model building and also for understanding how the design of the data might or might not allow the estimation of variance components. Working with mixed models requires in the first place that you have a good understanding of the variance structure that is imposed on the data. Software used for mixed model analysis will require the user to specify the variance structure. In this chapter we will discuss some simple variance structures with some more sophisticated extensions.

The Equation

The equation of a model defines the factors that will or could have an effect on an observed trait. The general linear model equation in matrix form is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad ...(1)$$

where

y is an $n \times 1$ vector of n observed records
b is a $p \times 1$ vector of p levels of fixed effects
u is a $q \times 1$ vector of q levels of random effects
e is an $n \times 1$ vector of random, residual terms
X is a known *design matrix* of order $n \times p$, which relates the records in **y** to the fixed effects in **b**
Z is a known *design matrix* of order $n \times q$, which relates the records in **y** to the random effects in **u**

Expectations and Variance Covariance (VCV) Matrices

In general the expectation of the model parameters is

$$E\begin{pmatrix} \text{y} \\ \text{u} \\ \text{e} \end{pmatrix} = \begin{pmatrix} \text{Xb} \\ 0 \\ 0 \end{pmatrix} \qquad ...(2)$$

which is also known as the 1st moment. The 2nd moments describe the variance-covariance structure of y:

$$V\begin{pmatrix} \text{u} \\ \text{e} \end{pmatrix} = \begin{pmatrix} \text{G} & 0 \\ 0 & \text{R} \end{pmatrix} \qquad ...(3)$$

where **G** is a dispersion matrix for random effects other than errors and **R** is the dispersion matrix of error terms, for which both are general square matrices assumed to be non-singular and positive definite, with elements that are assumed known.

We usually write

$$\text{Var}(y) = \mathbf{V = ZGZ' + R} \qquad \text{.....(4)}$$

Hence, the variance structure among y is only determined by the random effects.

Structure among data

Take a simple model with 5 observations

Animal 1	y1
Animal 1	y2
Animal 2	y3
Animal 2	y4
Animal 3	y5

The model is y = µ + a + e, with a being the animal effect.

We have 5 observations and 3 animals, hence the Z matrix has5 rows and 3 columns. The variance-covariance (VCV) matrix among animals is **G** is a 3 x 3 matrix.

We could write var(a) = **G** = $I_3 s_a^2$, assuming there is no covariance among the animal effects. The VCV among residuals is **R** is a 5 x 5 matrix.

Would we expect covariances among the residuals? There would certainly be a covariance between two observations on the same animal, but this is assumed to be covered by the common effect of animals. This is easy to see by working out the elements of (4). We assume first all random residual effects are uncorrelated and have equal variance, i.e. R = $I s_e^2$.

The Z matrix looks like:

$$Z = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } ZGZ' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} s_a^2$$

Note that the dimensions need to be 5 by 5 as it is a description of the variance structure of the data, this part due to the animal component. We see a block diagonal with blocks for each animal and the dimension of each block equal to the number of observations for each animal.

The total variance structure of the data as described in (4) by ZGZ' + R looks like

$$V = ZGZ' + R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} s_a^2 + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} s_e^2 =$$

$$\begin{pmatrix} s_a^2+s_e^2 & s_a^2 & 0 & 0 & 0 \\ s_a^2 & s_a^2+s_e^2 & 0 & 0 & 0 \\ 0 & 0 & s_a^2+s_e^2 & s_a^2 & 0 \\ 0 & 0 & s_a^2 & s_a^2+s_e^2 & 0 \\ 0 & 0 & 0 & 0 & s_a^2+s_e^2 \end{pmatrix}$$

Hence, we see a variance structure cause by two variance components such that

Var(y_i) = $s_a^2 + s_e^2$

cov(y_i, $y_{i'}$) = s_a^2 if the two observations are measured on the same animal and

cov(y_i, $y_{i'}$) = 0 otherwise.

Note that the structure is here defined by the design matrix Z. The random effects themselves do actually not have a structure among themselves, they are independent. Note also that the covariance among observations on the same animal is not invoked by correlations among the residuals, but by fitting an animal effect, which is a common effect to all observations for that animals, hence determining covariance. The residual effects are all equally distributed and independent (often called ‘measurement error’, which is a convenient term for interpretation purposes)

Structure among random effects

Now, what if the animals are genetically related? Suppose animals 1 and 3 are two half sibs? It seems simple to accommodate this by defining a covariance amongst these animals due to their additive genetic relationship.

We could write var(a) = **G** = $\begin{pmatrix} 1 & 0 & .25 \\ 0 & 1 & 0 \\ .25 & 0 & 1 \end{pmatrix} s_a^2 = \mathrm{A} s_a^2$, where A is a matrix with

additive genetic relationships among animals (usually called the numerator relationship matrix: NRM, see later).
The total variance structure of the data as described by ZGZ’ + R is now

$$V = ZGZ' +R = \begin{pmatrix} 1 & 1 & 0 & 0 & .25 \\ 1 & 1 & 0 & 0 & .25 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ .25 & .25 & 0 & 0 & 1 \end{pmatrix} s_a^2 + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} s_e^2 =$$

$$\begin{pmatrix} s_a^2+s_e^2 & s_a^2 & 0 & 0 & \tfrac{1}{4}s_a^2 \\ s_a^2 & s_a^2+s_e^2 & 0 & 0 & \tfrac{1}{4}s_a^2 \\ 0 & 0 & s_a^2+s_e^2 & s_a^2 & 0 \\ 0 & 0 & s_a^2 & s_a^2+s_e^2 & 0 \\ \tfrac{1}{4}s_a^2 & \tfrac{1}{4}s_a^2 & 0 & 0 & s_a^2+s_e^2 \end{pmatrix}$$

Hence, we see a variance structure caused by two variance components such that

Var(y_i) = s_a^2 + s_e^2; cov(y_i, $y_{i'}$) = s_a^2 if measured on the same animal; cov(y_i, $y_{i'}$) = $\tfrac{1}{4}s_a^2$ if measured on genetically related (half sib) animals and cov(y_i, $y_{i'}$) = 0 otherwise.

Note that the structure is now defined not only by the design matrix Z, but also by a structure among random effects themselves (in ASReml, this is indicated as ‘internal structure’).

Structure due to different random effects

Although the last model has already a non-trivial variance structure implemented, the question is whether the model is complete. One observation is that the covariance among two records of half sibs is one quarter of the covariance among two records on the same animal. The covariance among half sibs is, presumably, only genetic in origin (unless the half sibs have a common environment, usually not). This implies that the covariance among two records on the same animal also has to be additive genetic. However, usually, there are other reasons for the latter to be similar, due to common effects usually indicated as ‘permanent environmental (PE) effect’. Hence, we would expect records on the same animals to covary due to a complete similarity of genetic effect as well as the same PE effect. It is therefore expected that the covariance cov(y_1, y_2) > 4.cov(y_1, y_5). We can model this as follows:

$$y = µ + Z_1a + Z_2? + e.$$

Note that the design matrix is actually the same for a and ?, (Z_1 = Z_2) as this it relates observations to animals and one dose of a is always accompanied by one dose of ?. Now, relating this to the notation of the general mixed model: y = µ + Zu + e,

We have Z = [Z_1 Z_2] and u = $\begin{pmatrix} a \\ g \end{pmatrix}$ and var(u) = G = $\begin{pmatrix} As_a^2 & 0 \\ 0 & Is_g^2 \end{pmatrix}$

and var(y) = V = ZGZ’ = Z_1’A$Z_1 s_a^2$ + Z_2’$Z_2 s_g^2$ + R (each term being an n x n matrix) therefore

$$V = \begin{pmatrix} s_a^2+s_g^2+s_e^2 & s_a^2+s_g^2 & 0 & 0 & \frac{1}{4}s_a^2 \\ s_a^2+s_g^2 & s_a^2+s_g^2+s_e^2 & 0 & 0 & \frac{1}{4}s_a^2 \\ 0 & 0 & s_a^2+s_g^2+s_e^2 & s_a^2+s_g^2 & 0 \\ 0 & 0 & s_a^2+s_g^2 & s_a^2+s_g^2+s_e^2 & 0 \\ \frac{1}{4}s_a^2 & \frac{1}{4}s_a^2 & 0 & 0 & s_a^2+s_g^2+s_e^2 \end{pmatrix}$$

Compared to the previous model, the half sib covariances are smaller. As the total variance should be the same, the extra component s_g^2 has been taken out of one of the other. In this case, it has been fully taken out of s_a^2 as $a + ?$ in the latter model replaces a in the previous model. In other words, the animal effect has now been split into an additive genetic component and a permanent environmental component. The residual component is therefore the same in both models, i.e. s_e^2 is unchanged.

In the last model, we see a variance structure caused by tree variance components such that

Var(y$_i$) = $s_a^2 + s_g^2 + s_e^2$; cov(y$_i$, y$_{i'}$) = $s_a^2 + s_g^2$ if measured on the same animal; cov(y$_i$, y$_{i'}$) = $\frac{1}{4}s_a^2$ if measured on genetically related (half sib) animals and cov(y$_i$, y$_{i'}$) = 0 otherwise.

In complicated models, more than two random effects might be fitted, and these might have a structure among them e.g. maternal genetic effects, dominance effects, QTL effects etc. or uncorrelated (litter effects, maternal environmental effects), and different random effects may have a correlation among them (e.g. maternal and direct additive genetic effects, or additive genetic effects for different traits). Recently, more complicated models have been proposed to model the change of variance depending on a continuous variable, e.g. of time or environment. Hence, the variance of a breeding value (a) could depend on the age at measurement of the phenotype. These are random regression models, allowing implementation of covariances functions to fit the data (see again a later Chapter). In multivariate analysis, a range of structured or unstructured covariance patterns can be imposed

It is not difficult to formulate a mixed model with many random effects, however, it will be more difficult to estimate these effects if there are many. This will discussed further in the chapter on variance component estimation.

Variance structure for errors

In the previous, the error terms were assumed uncorrelated and had equal variance. Often, a more sophisticated model for the error terms is more appropriate.

1) The error variances may not be the same for the different observations. *Heterogeneous variances* are often modelled for different herds or flocks, different environment or more general for different subsets of the data.

2) There may be covariances between error terms.

Note that correlations between error terms can often be accommodated by an additional random effect, such as a permanent environmental effect. In the previous example, we could have use the model

$$y = \mu + Z_1 a + e.$$

and var(y) = V = ZGZ' = $Z_1'AZ_1 s_a^2$ + R

where

$$R = \begin{pmatrix} 1 & r & 0 & 0 & 0 \\ r & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & r & 0 \\ 0 & 0 & r & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} s_e^2, \text{ where } s_e^2 = s_g^2 + s_e^2 \text{ and } ? = s_g^2 / s_e^2$$

However, an additional fixed effect is usually easier to implement in estimation software than complicated covariance patterns among errors. Some error structures, however, are systematic and can be handled by some software programs. For example, if correlations are functions of distances or time, we can impose systematic correlation patterns such as an autocorrelation structure, where the correlation between observations in two periods that are t units apart is equal to $?^t$. In a following chapter on longitudinal data, more discussion on correlation patterns in longitudinal data will be presented.

Exercises:

1. Write the error covariance structure for repeated records from 4 consecutive years.

2. Write the covariance structure of the data for 6 records on 3 animals, each with 2 records, where animals 1 and 3 are full sibs.

3. Repeat 2) with the additional knowledge that animals 1 and 2 are from the same litter.