Chapter 17
Mixed Models for Genetic Analysis

Julius van der Werf

Mixed Models for Genetic Analysis

Application of mixed models has become an attractive tool to evaluate animals in actual breeding programs of breeding organizations. The methodology cons ists of a framework with justifiable statistical and genetic properties and it potentially delivers the most accurate and least biased prediction of breeding values.
The quality of evaluations depends on

1. The data (recording of management groups, correct identification, correct parentage
2. The model.

The BLUP methodology has the property to account for selection of parents in a breeding population. Hence, it fairly accounts for the fact that some animals are from better parents than other animals. Note that a requirement is that the pedigree, and data on selected parents as well as non-selected contemporaries are included in the analysis.
Models can be extended to account for more complicated effects, such as

- different breeds (useful for an 'across-breed evaluation or when there are animals imported from other countries.
- maternal effects: important in all pre-weaning traits
- correlated traits: useful for higher accuracy or to account for selection on a second trait (e.g. first lactation versus. later lactation or weaning weight versus yearling weight)
- interactions between environment and genotype: Some sires may have a different effect in different environments
- heterogeneous variance: the differences in one herd may be much larger (on average) than the differences in another herd.

Some factors are more difficult to include in the model: e.g preferential treatment of some animals (e.g. with hormones), or serious illness at the time of measurement. It is up the herd-recording scheme to design rules for when measurements can be considered as 'valid'. It is important here that the herd recording is unbiased and non-selective.

The mixed model model: general form:

The Model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$

where b is the vector with fixed effects with design matrix X (relating obs'ns to fixed effects)
u is the vector with random effects with design matrix Z (relating obs'ns to random effects)

Model definition
$E(\mathbf{y}) = \mathbf{Xb}$
$var(\mathbf{u}) = \mathbf{G}$
$var(\mathbf{e}) = \mathbf{R}$
$var(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$

The equations

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

This structure is expandable in many ways. The vector **u** could contain more random effects (e.g. additive genetic, maternal genetic, permanent environmental, maternal environmental, etc. The effects in **u** determine the structure of G. The vector y could contain more traits, and consequently u would have (breeding) values referring to the different traits. The R matrix could have a structure is there are correlations between errors, e.g. with correlated observations (traits). Also, R could contain different error variances for different groups of observations. Therefore, in defing a mixed model, not only fixed effects have to be defined, but also the variance structure of the random effects (hence the terms G-structure and R-structure in ASREML.

Single Trait Animal Model

The simple mixed model used in animal breeding is a single trait animal model. It is an 'animal' model because we fit a breeding value for each animal. 'Single trait' refers to the fact that animals have only observations on one character (trait) and there are only fixed effects and additive genetic effects, and no other random effects such as maternal or dominance. It is important to understand the principles of the simplest model. Less simple models are based on the same principles, and therefore not really much more difficult to understand. Like in any other statistical model building more complicated models largely requires more knowledge of the data, and imagination of effects that could possibly be causing differences.

More detail

In the single trait animal model with breeding value as the only random effect, we assume often that the matrix R is s equal to $I\sigma^2_e$ and the matrix G is equal to $A\sigma^2_a$. The simple equations were therefore obtained by multiplying the equations with the factor σ^2_e.

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'Y \\ Z'Y \end{pmatrix} \qquad \text{where } \lambda = \sigma^2_e / \sigma^2_a .$$

IN ASREML:

```
Analysis of some kind
  anim !P   The variable 'anim' is related to a pedigree file
  dage 10 !A
  rt 6
  wwt
  grp 322 !A
example.ped
example.dat
wwt ~ mu rt dage !r anim  !f grp      #model definition
```

There is no need to define variance structures more specifically, as there is only one extra random effect (besides residual) so one component of variance. One could put a scalr behind the anim variable in the model statement to indicate a starting value (optional).
i.e.

```
      bwt ~ mu bt dage !r anim 0.5 !f grp      #model definition
```

Sire Model

In this model, only effects of sires are fitted on records of their progeny (those are 0.5 times their breeding values!), making for computational ease. We may have only 100 sires in a data set on 100.000 recorded animals, hence needing 0.1% of the number of equations of an animal model. The λ value represents the ratio of error variance (including ¾ of the add. genetic variance) and sire variance (1/4 of the add. genetic variance) and the solutions are sire effects, i.e. ½ breeding values.

Sire EBVs obtained from a sire model may be slightly less accurate both due to lower accuracy (in case of few progeny / sire) and potential bias, because there is no correction for differences between dams. The model basically assumes that all progeny of a sire are of a different dam and all dams are expected to be from the same homogeneous population all with the same expected mean. In reality, dams could be of different breeds and dams are selected over years making the younger dams probably better than older dams. Fitting an animal model would allow to fit genetic relationships among dams and accommodate trends in dams breeding values..

Reduced animal model (RAM)

In this model, breeding values are only fitted for animals that have progeny records. This makes for faster computing (only equations for anima ls that are parents), and the EBV's for all other animals are simply derived from those of their parents, plus their own corrected phenotypes. The results are the same as for a full animal model. Less computing time at the cost of some extra computer programming time is needed.

Repeated records model

This is used where animals can have more than one record, such as multiple fleece weight records in sheep. The phenotypic correlation between recordings is equal to repeatability, and genetic correlation between recordings is assumed one (if the genetic correlation was less than one, then the multi-trait approach outline above is applicable!).
The approach is to invent a permanent environmental effect for each animal, i.e. when the animal has a second record, not only his breeding value but also part the environmental effects are repeated. This can represent effects of raising of the animal (a good 'development' guarantees a consistently good performance later on), or the occurrence of a disease that happened to a particular animal, with permanent effects.

The mixed model with repeated records can look like: $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Zp} + \mathbf{e}$
where **y** is the vector of the observations, **b** is the vector of fixed effects, **a** is a vector of additive genetic effects, **p** is a vector of permanent environmental effects and **e** is a vector of residual effects. The matrix **X** is the incidence matrix for the fixed effects and **Z** is the incidence matrix relating observations to animals. Each animal has an additive genetic as well as a permanent environmental effect, hence both effects have the same design matrix.
The three random effects have the following distribution

$$\text{var}\begin{pmatrix} a \\ p \\ e \end{pmatrix} = \begin{pmatrix} As_a^2 & 0 & 0 \\ 0 & Is_c^2 & 0 \\ 0 & 0 & Is_e^2 \end{pmatrix} = \begin{pmatrix} \mathrm{G} & 0 \\ 0 & \mathrm{R} \end{pmatrix} \qquad \mathrm{G} = \begin{pmatrix} \mathrm{A}\sigma_a^2 & 0 \\ 0 & \mathrm{I}\sigma_c^2 \end{pmatrix}$$

where σ^2_a is the direct additive genetic variance and σ^2_c is the variance due to permanent environmental effects. The model shows that those permanent environmental effects for different animals are uncorrelated, and within an animal there is no correlation between its additive and its permanent environmental effect. The total phenotypic variance is the sum of the three variance components.
The mixed model equations for a model with repeated records look like:

$$\begin{pmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + lA^{-1} & Z'Z \\ Z'X & Z'Z & Z'Z + gI \end{pmatrix} \begin{pmatrix} b \\ a \\ p \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \\ Z'y \end{pmatrix}$$

where now $\lambda = \sigma^2_e / \sigma^2_a$ and $\gamma = \sigma^2_e / \sigma^2_c$

IN ASREML:

```
Analysis of some kind
  anim !P   The variable 'anim' is related to a pedigree file
  dage 10 !A
  rt 6
  wwt
  grp 322 !A
example.ped
example.dat
wwt ~ mu rt dage !r anim  ide(anim) !f grp      #model definition
```

Maternal effects model

Some traits such a survival of piglets or early growth in beef cattle and meat sheep are influenced by maternal effects. The mother has an influence on the performance of her offspring over and above that of her direct additive genetic contribution, i.e. through maternal effects. These maternal effects are strictly environmental for the offspring, but can have both a genetic and environmental component. In selection of animals, and especially in dam lines, it is important to consider the maternal genetic effects. Beef cattle producers are interested in animals which have a high breeding value for growth (direct genetic effect) but also in cows with good mothering abilities (milk production). Including maternal effects in the model allows to estimate maternal effects and to correct for possible biases in genetic evaluation of the growing animal. It is usually assumed that maternal effects are genetic, although part of it might also be a permanent environmental effect (e.g. a beef cow with only three teats).

Maternal Effects Model

In the following model the direct genetic and maternal genetic effect are considered:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{m} + \mathbf{e}$$

where **y** is the vector of the observations, **b** is a vector of fixed effects, **a** is a vector of additive genetic effects, **m** is a vector of maternal genetic effects and **e** is a vector of residual effects. **X** is the incidence matrix for the fixed effects and $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are incidence matrices relating observations to random effects of animal (additive genetic) and dam (maternal genetic), respectively. The random effects have the following distribution:

$$\text{var}\begin{pmatrix} a \\ m \\ e \end{pmatrix} = \begin{pmatrix} As_a^2 & As_{am} & 0 \\ As_{am} & As_m^2 & 0 \\ 0 & 0 & Is_e^2 \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \qquad G = \begin{pmatrix} A\sigma_a^2 & A\sigma_{am} \\ A\sigma_{am} & A\sigma_m^2 \end{pmatrix} = G_0 \otimes A$$

where G_0 is a 2 by 2 matrix: $\begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix}$ and $\otimes$ is a direct product (it 'blows up' a matrix)

Further σ^2_a is direct genetic variance, σ^2_m the maternal genetic variance, σ_{am} the covariance between direct and maternal genetic effects and σ^2_e the error variance. The model shows that both random effects have a covariance structure depending on the genetic relationships. Related dams have related maternal genetic effects, and there is a correlation between a dam's direct additive genetic effect and her maternal genetic effect. The total phenotypic variance is equal to $\sigma^2_p = \sigma^2_a + \sigma^2_m + \sigma_{am} + \sigma^2_e$

The mixed model equations are:

$$\begin{pmatrix} X'X & X'Z_1 & X'Z_2 \\ Z_1'X & Z_1'Z_1 + \alpha_{11}A^{-1} & Z_1'Z_2 + \alpha_{12}A^{-1} \\ Z_2'X & Z_2'Z_1 + \alpha_{21}A^{-1} & Z_2'Z_2 + \alpha_{22}A^{-1} \end{pmatrix} \begin{pmatrix} b \\ u \\ m \end{pmatrix} = \begin{pmatrix} X'y \\ Z_1'y \\ Z_2'y \end{pmatrix} \quad \text{where} \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} = G_0^{-1}.\sigma^2_e$$

Since the full (inverse) relationships matrix is used with relationships between all animals

(progeny as well as dams), estimates will be obtained for additive effects for progeny (with records) as well as for dams (with possibly no own record). Equally, estimates for maternal effects will be obtained for dams (with progeny) as well as for progeny (which may not have expressed their maternal ability yet).

IN ASREML:

```
Analysis of some kind
  anim !P   The variable 'anim' is related to a pedigree fil
  dam !P    The variable 'dam' is related to a pedigree file
  dage 10 !A
  rt 6
  wwt
  grp 322 !A
example.ped
example.dat
wwt ~ mu rt dage !r anim  dam !f grp       #model definition
0 0 1                                     #R struc: # sites, dim Ro, #G struct
anim 2                                    #G structure: model term, dimensions
2 0 US !GP                                 #order Go, 0, model
.2 0 .15                            starting_values
anim  o AINV                                      #inner dimension of G structure
```

Note that this model has an 'US' structure for the G matrix, i.e. all 3 variances of G_o will be estimated. It is possible to ignore the covariance between direct and ma ternal effects. In that case the G_0 is diagonal and the last lines of the as file are:

```
wwt ~ mu rt dage !r anim  dam !f grp       #model definition
0 0 1                                     #R struc: # sites, dim Ro, #G struct
anim 2                                    #G structure: model term, dimensions
2 0 DIAG .2 .15                            #order Go, 0, model  starting_values
anim                                      #inner dimension of G structure
```

In maternal effects models, it is also possible to fit the dam effect as environmental effect (me), in ASREML: ide(dam), i.e. the dam effect is fitted with an identity structure rather than with a relationships structure (A-structure).. In that case, no genetic relationships among dams are considered in estimating its effects. The me effect is considered uncorrelated with direct effects.

G_0 is a 2 by 2 matrix: $\begin{pmatrix} s_a^2 & 0 \\ 0 & s_{me}^2 \end{pmatrix}$

And the ASReml model statement reads like:

```
wwt ~ mu rt dage !r anim  ide(dam) !f grp      #model definition
```

Another model fits both genetic and environmental dam effects. Note that the number of genetic maternal effects estimated is equal to the number in the pedigree, while the number of perm. env. dam effects is equal to the number of dams that have progeny with data.

G_0 is a 3 by 3 matrix: $\begin{matrix} \sigma^2_a & \sigma_{am} & 0 \\ \sigma_{am} & \sigma^2_m & 0 \\ 0 & 0 & \sigma^2_{me} \end{matrix}$

And the ASREML model statement reads like:

```
wwt ~ mu rt dage !r anim  dam ide(dam) !f grp      #model definition
0 0 1                                     #R struc: # sites, dim Ro, #G struct
anim 2                                    #G structure: model term, dimensions
2 0 US .2  0 .10                           #order Go, 0, model  starting_values
anim
```

Notice that we don’t need to define the whole of G. We only define the first 2 terms (US structure). The last term is ‘left over’ but independent and only one variance component need to be estimated for that term (no need to define a structure.

```
wwt ~ mu rt dage !r anim  dam ide(dam) !f grp      #model definition
0 0 2                                     #R struc: # sites, dim Ro, #G struct
anim 2                                    #G structure: model term, dimensions
2 0 US .2  0 .10                           #order Go, 0, model  starting_values
anim
```

Exercise

Given are 6 records on yearling weight in beef cattle. The animals were kept in two paddocks. Also given is the pedigree.

- Calculate EBV's for the animals involved, ignore maternal effects.
- Calculate EBV's with accounting for maternal effects.

data

paddock	animal	yearling weight
1	4	354
1	5	251
2	6	327
2	8	328
2	9	301
2	10	270

pedigree

animal	sire	dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	3
6	1	2
7	0	0
8	7	4
9	7	4
10	7	5

Parameters:
Phenotypic Variance: Vp = 900
Heritability: h2 = 0.30
Maternal Heritability: m2 = 0.15

Assume a zero correlation between the direct and the maternal effect