Mixed Models for Genetic Analysis

Julius van der Werf
University of New England
Armidale, NSW Australia

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 consists 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 animals 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: **y** = **Xb** + **Za** + **Zp** + **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

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

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:

$$\text{var}\begin{pmatrix} a \\ m \\ e \end{pmatrix} = \begin{pmatrix} As_a^2 & As_{am} & 0 \\ As_{am} & As_m^2 & 0 \\ 0 & 0 & Is_e^2 \end{pmatrix} = \begin{pmatrix} \mathrm{G} & 0 \\ 0 & \mathrm{R} \end{pmatrix} \qquad \mathrm{G} = \begin{pmatrix} \mathrm{A}\sigma_a^2 & \mathrm{A}\sigma_{am} \\ \mathrm{A}\sigma_{am} & \mathrm{A}\sigma_m^2 \end{pmatrix} = \mathrm{G}_0 \otimes \mathrm{A}$$

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 maternal 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 ASRTEML 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: $\begin{matrix} s^2_a & s_{am} & 0 \\ s_{am} & s^2_m & 0 \\ 0 & 0 & s^2_{me} \end{matrix}$

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
```

Multiple Trait Genetic Evaluation

Multi trait model

This is an extension of the single trait case. Data on a number of traits are available in Y, and EBV's are calculated for each trait. The results are generally different from what would be got from a number of separate single-trait BLUPs, because each trait is used to help give information about all other traits, much as with a selection index. In a later chapter, the multiple trait BLUP procedure will be worked out in more detail. The benefit form multiple trait models comes from

- more accuracy as information from correlated traits is used
- less bias as the analysis will take into account that for traits that are measured after sequential rounds of selection, only the better ones are evaluated.

An example of potential selection bias. Compare a good bull and a bad bull, each having 40 progeny at weaning. From the good bull, no progeny are culled, whereas from the bad bull 50% is culled. Comparing the progeny of these bulls at post-weaning will give a huge advantage to the bad bull, as his bad progeny have been removed. Multi-trait BLUP would correct for this bias.

For the genetic evaluation of the animals, we can use information which is available on all traits. Originally the main reason for using information on all traits was to obtain more accurate evaluations. With using information on correlated traits the accuracy of the estimated breeding value increases. A second advantage arose later, namely a multiple trait analysis is the only way to obtain unbiased estimates for a trait, which is observed only on animals selected based on values of a correlated trait. A model including information of the correlated trait, on which selection was based, is able to correct for this type of selection. An example of this is the evaluation of the second lactation productions of dairy cows where selection has been practised based on the first lactation. Only animals that survived the first lactation have a second lactation record, and those are usually only the better animals. Other examples are the analysis of piglets born in second litter, or the analyses of yearling weight after animals have been selected for weaning weight.

Multiple Trait Mixed Model

Taking again as a starting point the mixed model in its general form:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}.$$

With more traits we can now partition the observation vector **y** in a part for each trait. The same can be done with the associated environmental effects. The vector of breeding values is also partitioned for the different traits, so that each animal has a breeding value for each trait in the analysis.

For a 2-trait example, the vector $\mathbf{y}_1$ represents the n_1 observations for trait 1 and $\mathbf{y}_2$ represents n_2 observations for trait two. For each trait we can write a mixed model:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i ,$$

where there are p_i fixed effects associated with trait i so that $\mathbf{X}_i$ is an n_i x p_i matrix and $\mathbf{b}_i$ is a p_i x 1 dimensional column vector. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are incidence matrices for fixed effects and random effects for trait i, respectively.

The multiple trait model can be represented as follows:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

Notice that not all animals necessarily have an observation for both traits. Some animals may be represented in $\mathbf{y}_1$ but not in $\mathbf{y}_2$, or vice versa. All animals, however, are represented with a breeding value for each trait in the analysis, irrespective whether they had an observation for that trait. The vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ (and $\mathbf{e}_1$ and $\mathbf{e}_2$) are therefore not necessarily of the same length, but $\mathbf{u}_1$ and $\mathbf{u}_2$ are always equally long (with the number of elements equal to the number of animals in the analysis).

To obtain the mixed model equations for estimating fixed effects **b** and breeding values **u**, we need to specify the covariance matrices **R** and **G** associated with the vector $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2)'$ of residual errors and the vector $\mathbf{u} = (\mathbf{u}_1 , \mathbf{u}_2)'$ of random effects.

For the breeding values we can write

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \text{ and } \text{var}(u) = G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

If $\sigma_g{}^2_{ii}$ is the genetic variance of trait i, and σ_{gij} is the genetic covariance between the two traits (within one animal), we can define a 2 by 2 genetic covariance matrix

$$G_0 = \begin{pmatrix} \sigma^2_{g11} & \sigma_{g12} \\ \sigma_{g21} & \sigma^2_{g22} \end{pmatrix}$$

Each part of G is obtained by multiplying the relationships matrix with either the variance of a trait (diagonal blocks $g_{ii.}A$) or the covariance between the traits (off diagonal blocks $g_{ij.}A$) where g_j is an element of $G_{0.}$. The covariance between the breeding value of trait *i* on individual *k* and the breeding value of trait *j* in individual *l* is the additive genetic covariance between traits *i* and *j* multiplied by the additive genetic relationship between individuals *k* and *l.*

Multiple Trait Mixed Model Equations

The mixed model equations for a multiple trait model can be written according to the general principle of setting up mixed model equations. However, they are extended for the **G-** and the **R**-matrices.

For the mixed model equations we will need the inverse of **G**. In te multiple trait mixed model this becomes:

$$G^{-1} = \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix} \text{ where } G^{ij} = g^{ij}A^{-1},$$

and where g^{ij} is (i,j)-element of the inverse of the 2 by 2 genetic covariances matrix G_0 and A^{-1} is the inverse of the relationships matrix as it can be setup directly.

The residual covariance matrix **R** has the same form, but **A** is replaced by an identity matrix **I** assuming there are no correlations between the residuals of different animals. While residual deviations for a given trait measured on different individuals are often assumed to be uncorrelated, this is not necessarily the case for different traits measured on the same individual. The phenotypic correlation between traits is often the result of correlation between genetic as well as environmental effects. When all traits are measured on all individuals ($n_1=n_2=n$), the covariance matrix between $\mathbf{e}_i$ and $\mathbf{e}_j$ can be written as $\sigma(\mathbf{e}_i, \mathbf{e}_j)=r_{ij}\mathbf{I}$, where $r_{ij} = \sigma_e(i,j)$ is the environmental covariance between traits *i* and *j* as expressed in the same individual. The resulting *n*.2 x *n*.2 variance-covariance matrix for the total error vector $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2)'$ becomes:

$$\mathbf{R} = \begin{bmatrix} \sigma(e_1, e_1) & \sigma(e_1, e_2) \\ \sigma(e_2, e_1) & \sigma(e_2, e_2) \end{bmatrix} = \begin{bmatrix} I\, r_{11} & Ir_{12} \\ I\, r_{21} & I\, r_{22} \end{bmatrix}$$

and the inverse is $R^{-1} = \begin{bmatrix} Ir^{11} & Ir^{12} \\ Ir^{21} & Ir^{22} \end{bmatrix}$

where r^{ij} is i-j element of the inverse of the 2 by 2 environmental covariances matrix between the two traits: R_0.

The set of multiple trait mixed model equations are given in the next figure. It is not the idea to memorise these equations, but to give you a how single trait mixed model equations are 'blown up' to multiple trait mixed model equations. This has a rather large impact on the number of equations that has to be solved. Roughly, computer time for solving multiple trait mixed models goes up quadratically with the number of traits!

$$\begin{bmatrix} X_1' r^{11} X_1 & X_1' r^{12} X_2 & X_1' r^{11} Z_1 & X_1' r^{12} Z_2 \\ X_2' r^{21} X_1 & X_2' r^{22} X_2 & X_2' r^{21} Z_1 & X_2' r^{22} Z_2 \\ Z_1' r^{11} X_1 & Z_1' r^{12} X_2 & Z_1' r^{11} Z_1 + g^{11} A^{-1} & Z_1' r^{12} Z_2 + g^{12} A^{-1} \\ Z_2' r^{21} X_1 & Z_2' r^{22} X_2 & Z_2' r^{21} Z_1 + g^{21} A^{-1} & Z_2' r^{22} Z_2 + g^{22} A^{-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} X_1'(r^{11} y_1 + r^{12} y_2) \\ X_2'(r^{21} y_1 + r^{22} y_2) \\ Z_1'(r^{11} y_1 + r^{12} y_2) \\ Z_2'(r^{21} y_1 + r^{22} y_2) \end{bmatrix}$$

If not all traits are recorded for all animals, the inverse of the residual covariance matrix ***R*** becomes a bit trickier. The reason is that animals with one record only do not have a

residual covariance with another trait. The covariance matrix between the residuals of the different traits ($\sigma(e_1, e_2)$) can no longer be written as a diagonal matrix (a multiple of $\mathbf{I}$). When some observations are missing, the matrix $\mathbf{X}_1'$ can not be directly multiplied with $\mathbf{X}_2$, i.e. the number of columns (= n_1) does not correspond with the number of rows (= n_2). This can be solved by using $\mathbf{X}_1' r^{12} \mathbf{I}_{12} \mathbf{X}_2$ where $\mathbf{I}_{12}$ identifies when two observations are on the same individual (only in those case we have an environmental covariance). When both traits are measured on all animals $\mathbf{I}_{12} = \mathbf{I}$ and $\mathbf{X}_1 r^{12} \mathbf{I}_{12} \mathbf{X}_2$ reduces to $\mathbf{X}_1' r^{12} \mathbf{X}_2$.
The rules for building up multiple trait mixed model equations are outlined hereafter, as a reference for the further interested reader.

Rules for building mixed model equations:
(*this section is only for reference)*

For small examples the mixed model equations can be build from the corresponding design matrices. For larger data sets, however, this becomes complicated. Rules have been developed to build the mixed model equation without explicitly setting up the design matrices. These rules for building the mixed model equations with a multiple trait model (per animal) are:

1**)** If both y_1 and y_2 are known for an animal;
The values for r^{11}, r^{12}, r^{21}, and r^{22} are added to the particular sections for each trait in the fixed and random part of the coefficient matrix. For instance for 2 effects (herd and animal), we have to add these four numbers to each of **X'X**, **X.Z,** , **Z'X** and **Z'Z** . To the vectors with the totals (right hand sides) we add $r^{11}y_1 + r^{12}y_2$ and $r^{12}y_1 + r^{22}y_2$ to each trait partition of the two vectors (**X**'y and **Z**'y).
In a single trait model, we would have added only one figure to the 4 partial matrices for a trait. For the totals (right hand sights), we would add only y to the partial vectors for herd and animal.

2) When only one observation for one trait on the animal is available;
The values $(r_{11})^{-1}$ or $(r_{22})^{-1}$ are added to each of the relevant partial matrices in the coefficient matrix, while $y_1(r_{11})^{-1}$ or $y_2(r_{22})^{-1}$ are added to the relevant parts of the right hand sides.

3) Independent of the pattern of traits measured, we add the relationships matrix multiplied by g^{ij} to the i-j block of the random effects of the coefficient matrix.

Example of a Multiple Trait Model

Consider a situation where we have the following measurements on 6 unrelated and non-inbred individuals, performing in two different herds. Both traits on an animal are measured in the same herd.

Individual	Herd	Trait 1	Trait 2
1	1	160	-
2	1	180	320
3	1	210	330
4	2	190	-
5	2	228	360
6	2	210	350

The phenotypic standard deviations for weaning weight and yearling are 20 and 40 kg, respectively, the heritabilities are 0.42 and 0.39 and the genetic correlation is 0.769. The 2

$$G_0 = \begin{bmatrix} 169 & 250 \\ 250 & 625 \end{bmatrix} \text{ which corresponds with } G_0^{-1} = \begin{bmatrix} .0145 & -.058 \\ -.058 & .0039 \end{bmatrix}$$

by 2 matrix with additive genetic covariances (within an individual) are:

The within-individual environmental covariances are:

$$R_0 = \begin{bmatrix} 231 & 285 \\ 285 & 975 \end{bmatrix} \text{ which corresponds with } R_0^{-1} = \begin{bmatrix} .0068 & -.0020 \\ -.0020 & .0016 \end{bmatrix}$$

The design matrices for the first trait are straightforward. $\mathbf{Z}_1$ is an identity matrix since all animals have a record for the first trait. For the second trait, however, more attention is needed. The matrix $\mathbf{Z}_2$ has one column for each breeding value (i.e. 6 columns) and one row for each observation (i.e. 4 rows) which results in:

$$Z_2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The right hand side (RHS) for the example are (transposed):

[2.05 2.38 0.271 0.272 0.69 0.58 0.77 0.82 0.83 0.73 0 0.16 0.11 0 0.13 0.15]

The first 4 elements are for the fixed effects (2 herds for 2 traits).
The values are all scaled by multiplying with residual (co)variances. For example: the RHS-value for the second animal for the first trait (6^{th} element) is obtained as:
$r^{11}y_{12} + r^{12}y_{22} = 0.0068*180 + (-0.002)*320 = 0.584$, where y_{12} (=180) and y_{22} (=320) are the record for the first and second trait for the second animal. The first animal has only the first trait measured, and its RHS value (5^{th} element) becomes
$(r_{11})^{-1} y_{11} = 0.0043*160 = 0.69$. Notice that when an animal has only one trait recorded, we multiply it by $(r_{11})^{-1}$ (the inverse of the 1-1 element of the residual covariance matrix) and not by r^{11} (the (1,1)-element of the inverse of the residual covariance matrix). Notice also that animals with no record for a given trait have a zero in the RHS.

The solutions for the fixed effects and the <u>multiple trait BLUP</u> EBV's are:

$$b_{1'} = \begin{bmatrix} 183 & 209 \end{bmatrix} \quad b_{2'} = \begin{bmatrix} 309 & 342 \end{bmatrix}$$

$$u_1 = \begin{bmatrix} -9.86 \\ -1.00 \\ 10.86 \\ -8.17 \\ 7.70 \\ 0.47 \end{bmatrix} \quad u_2 = \begin{bmatrix} -14.58 \\ 2.87 \\ 11.72 \\ -12.08 \\ 9.35 \\ 2.73 \end{bmatrix}$$

where b1 are the solutions for the herd effects for weaning weight, and u_1 are EBV's for

weaning weight, and b_2 and u_2 refer to yearling weight.

When performing two single trait evaluations for the two traits the following solutions were found for Single Trait BLUP:

$$\textit{Single trait}: \quad b_{1'} = \begin{bmatrix} 183 & 209 \end{bmatrix} \quad b_{2'} = \begin{bmatrix} 325 & 355 \end{bmatrix}$$

$$u_1 = \begin{bmatrix} -9.86 \\ -1.41 \\ 11.27 \\ -8.17 \\ 7.89 \\ 0.28 \end{bmatrix} \quad u_2 = \begin{bmatrix} 0 \\ -1.95 \\ 1.95 \\ 0 \\ 1.95 \\ -1.95 \end{bmatrix}$$

Notes to the solutions:

1) The average breeding value for both traits is equal to zero within herd. This is to be expected because animals are assumed to be unrelated. This illustrates that it is impossible to make a fair comparison of the average breeding value of animals in herds when there are no genetic ties (e.g. offspring from a common sire).

2) Animal 1 has no observation for trait 2. Consequently its breeding value is entirely based on the information from the correlated trait 1. Animal 1 has a value for trait 1 which lies below the herd average and as a result of the positive genetic correlation between traits its breeding value is also below average, i.e. negative. The same is true for animal 4 in herd 2.

3) The single trait breeding values (and fixed effect solutions) deviate from the multiple trait solutions. As to be expected the single trait breeding values of animal 1 and 6 for yearling weight are equal to zero. There is no information to estimate the breeding value and consequently the animals get the average breeding value. The difference in breeding value for yearling weight

between animal 2 and 3 (and between 5 and 6) is larger in the mu ltiple trait case. The reason is that the information from weaning weight (the correlated trait) gives additional evidence that these animals are different in breeding value.

4) The difference between the average herd effect for weaning weight and yearling weight is larger in the single trait analysis. In fact, this difference is overestimated, since is it biased by the fact that for yearling weight we only recorded the best animals (=selection). The multiple trait evaluation takes this into account. From using the information on the first trait, the model knows that only the better animals had a yearling weight measured.

5) In the multiple trait EBV's we see that the animals that were not culled have an average EBV's above zero. This makes sense, because from the information on trait 1 we know that these are actually the better animals. Single trait evaluation would not use information on weaning weight, and consider the yearlings that were weighted as average animals. This shows that multiple trait evaluation is able to correct for sequential selection.

Advantages of Multiple Trait BLUP evaluation

In general, using the multiple trait model gives an increase in accuracy of estimated breeding values. Furthermore, in many cases it is the only way to correct for selection on correlated trait.
The importance of increase of accuracy by using extra information, i.e. the importance of using a multiple trait (MT) model, depends on several aspects:

- *the information available on each animal*

If few or no observations are available for a particular trait, using observations on another trait when both traits are genetically correlated can increase the accuracy.

- *parameter structure*

If genetic and environmental correlations are small, the multiple trait model has few advantages. Furthermore, in a situation with a high h^2, only a few observations are needed for an accurate estimate of the breeding value. In other words, information of other traits is less important in that case. Besides, the difference between r_g and r_e is important; the larger the difference, the larger is the contribution of a correlated trait to the reduction of the Prediction Error Variance. The contribution of correlated traits to the accuracy of estimating breeding values can be examined with the selection index method.

- *correctness of parameters;*

In multiple trait model we make use of estimated values of the genetic parameters (heritabilities, correlations). This variance-covariance (VCV) matrix has to be checked on incorrectness (or consistency).

Schaeffer (1984) discussed the effects of incorrect estimated parameters. He distinguished two kinds of mistakes. First, the VCV matrices may not be valid, i.e. within the parameter space. A valid VCV matrix, by definition, is a positive definite matrix. This can be checked by looking at all the eigenvalues of the matrix. Eigenvalues of covariance matrices all have to be positive, making the matrix "positive definite". The second and most common mistake, mentioned by Schaeffer, is that estimates used in the model, could be greatly different from the underlying true values. Assume that the true parameters give the maximum response of selection. The realised response then depends on the difference with the parameters used, namely $(r_g - \hat{r}_g)$ and $(r_e - \hat{r}_e)$.

In this respect, it is good to realise that single trait models are MT models with the assumption that $\hat{r}_e = \hat{r}_g = 0$. Therefore, inaccurate correlations are often still closer to the true values than zero correlations!

- *Correction for selection*

The example illustrated selection on sequentially recorded traits leads to culling and missing records for traits that are recorded in a later stage. Multiple trait evaluation was able to avoid selection bias.

This reflects a more general rule, also applicable in single trait genetic evaluation, that to

avoid selection bias, all information that was used to base selection decisions on, should be included in the analysis. This is not only the case with missing records in sequentially recorded traits. Assume the situation when two traits are recorded simultaneously, and all animals have records for all trait, but selection is only for one of the traits. Single trait evaluation of one trait only would lead to biased EBV's and generally to an underestimate of the genetic trend for the correlated trait (although this depends on the genetic and environmental correlation between the traits). Since selection is usually on an index (a linear combination on all traits), single trait evaluation leads to incorrect estimates of the genetic trend in most of the cases!

Computational considerations

Computer requirements quickly increase with the application of multiple trait BLUP genetic evaluation procedures. Suppose we want to carry out a 5-trait BLUP analysis. The multiple trait mixed model equations require nearly 25 times more coefficient to be handled compared to single trait BLUP. Solving the mixed model equations when multiple traits are present can be greatly simplified by constructing a transformation for the traits being considered (this is called 'canonical transformation'). This transformation constructs a new set of uncorrelated variables, which can be analysed in independent single trait evaluations. Such a transformation is possible when all animals had observations for all traits. Recently, algorithms have been developed to handle transformations also for the case of missing observations on some traits. Multiple trait models can still be quite cumbersome if more random effects are included (e.g. maternal effects for some traits. However, The combination of more efficient computing algorithms with the rapid increase of computing power has lead to a situation that multiple trait BLUP is the method of choice for more and more genetic evaluation systems.

Software

There are software packages available that can be used to implement multiple trait genetic evaluations. A commonly used package for breeding value estimation is PEST (Prediction and ESTimation) written by Groeneveld et al. (1994). A more versatile and increasingly used package is ASREML (Gilmour et al., 1996: This package is most suitable for estimation of genetic parameters in animal breeding data for a wide variety of models. There are also genetic evaluation services around that provide the whole package of delivering multiple trait EBV's.

An ASREML example:

```
bwt wwt~ Trait at(Trait,1).bt at(Trait,2).rt Trait.dage !r Trait.anim!f Trait.grp
1 2 1                                #R struct: 1 site, dimension Ro, 1 G structure
0                                    #order R (?), ASREML figures out if put to zero
2 0 US 12 0 14 !GP                   # order Ro, 0, model, starting_values
Trait.anim 2                         #G structure: model term, dimension
2 0 US 4.9 0 4.5 !GP                 #order Go, 0, model starting_values
anim
```

In the model statement, some effects are fitted for both traits: Trait.dage

Other effects are fitted for one trait only at(Trait,2).rt

!GP means that the matrix (R of G) has to be positive definite

A multi-trait model can also have more random effects, e.g. a maternal effect:

```
bwt wwt~ Trait at(Trait,1).bt at(Trait,2).rt Trait.dage !r Trait.anim Trait.dam !f
Trait.grp
1 2 1                                #R struct: 1 site, dimension Ro, 1 G structure
0 0 ID                          #nrec (= outer dim. Of R), ASREML figures out if put to zero
Trait 0 US 12 0 14 !GPUP        # order Ro (equal to nr. of traits), 0, model, start_values
Trait.anim 2                         #G structure: model term, dimension
4 0 US !GP                           #order Go, 0, model starting_values
4.9
2 9.5
0 0    4.5
0 0    2   4.2
anim 0 AINV
```

The G_0 has now dimension 4. The definition of the G_0 can be spelled out in some more detail:

```
4 0 US 4.9 2 9.5 0 0 4.5 0 0 2 4.2 !GPUPFFPFFUP
```

4 order of Go
0 always a zero here
US unstructured Go
4.9 following is lower Go starting values
2 9.5
0 0 4.5
0 0 2 4.2
!GPUPFFPFFUP indicating whether the components should be **P**ositive, **U**nstructured, or **F**ixed at the starting value

the same line could be replaced by:

```
4 0 US !+10 !GPUPFFPFFUP
4.9         following is lower Go starting values
2 9.5
0 0 4.5
0 0 2 4.2
```

!GPUPFFPFFUP could be replaced by !GP if we simply want Go to be positive definite

Random Regression Models

Random regression models can typically be used when a trait is expressed repeatedly, e.g. over time or in different environments. In that case, the effect changes gradually along a trajectory of time, or of some other continuous variable (temperature, elevation, rainfall. For simplicity, we think of the expression of body weight as a function of time. If the random effects are modeled as a function of time, then both the variance as the covariance between expression at different times are modeled as a continuous function. Note that previously we often modeled repeated measures of weight as multiple traits, e.g. wwt, pwwt, ywt.. The advantage of random regression is that traits can be measured at any point along a trajectory, i.e. at any age, and we do not have to chop this up in distinct traits.

In linear models we are used to fitting weight as a regression of age. This is often a fix3ed regression, indicating that for each animal that is a certain amount of time younger or older than an average age there will be a weight correction. This correction is the same for all animals, hence a fixed regression. In random regression models, we estimate a different regression coefficient for each animal. Hence, each animal has his/her own slope (some grow faster than others) and we estimate te variance of all slop parameters. An animal individual's slope is estimated as a BLUP, depending on the variance of slopes (like the breeding value is derived from the variance of breeding values.

Hence, each animal may have 3 breeding value for weight, if we fit a three order regression. The first is an intercept, the 'average weight', the second is a slop, 'the growth', and the third is a quadratic term

The regression coefficients are not the same for each animal, but they are drawn from a population of regression coefficients. In other words, regression coefficients in $\boldsymbol{a}$ and $\boldsymbol{p}$ are *random regression coefficients* with var($\boldsymbol{a}$)= $\boldsymbol{K}_a$ and var($\boldsymbol{p}$)= $\boldsymbol{K}_p$, where a is additive genetic effect and p is permanent environmental effect.

In fact, we have rewritten a multivariate mixed model to a mixed model in a format of a univariate random regression model, with each random effect having *k* random regression coefficients. A model for n observations on q animals can then be

written as

$$\mathbf{y} = \mathbf{Xb} + \sum_{j=0}^{k-1} \mathbf{Z}_j\mathbf{a}_j + \sum_{i=0}^{k-1} \mathbf{Z}_j\mathbf{p}_j + \mathbf{e}, \qquad [4\text{-}6]$$

where $\mathbf{Z}_j$ are n by q matrices for the ith polynomial, and $\mathbf{a}_j$ and $\mathbf{p}_j$ are vectors with random regression coefficients for all animals for additive genetic and permanent environmental effects. The matrix **Z** contains the regression variables, i.e. the coefficients are those of the polynomials in F **(i.e. rather than a 1's, Z contains 1, x, x^2, etc.**. We can order the data vector by sorting records by animal, and we can stack the $\mathbf{a}_j$ and $\mathbf{p}_j$ vectors and sort them by animal, each animal having *k* coefficients in **a** and *k* coefficients in **p** (to simplify notation, we assume equal order of fit for CF's for both random effects, therefore having equal incidence matrices). We can then write $\mathbf{Z}^*$ as a block diagonal matrix of order *n by k*q*, with for each animal *i* block $\mathbf{Z}_i^* = F_i$.

The mixed model can be written as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}^*\mathbf{a} + \mathbf{Z}^*\mathbf{p} + \mathbf{e},$$

with $\mathbf{a}' = \{\mathbf{a}_1', ...\mathbf{a}_q'\}$ and $\mathbf{p}' = \{\mathbf{p}_1', ...\mathbf{p}_q'\}$, with $\mathbf{a}_i$ and $\mathbf{p}_i$ being the sets of random regression coefficients for animal i for the additive genetic and the permanent environmental effects, respectively. If all animals have measurements on the same age points, all $\mathbf{Z}_i^*$ are equal and $\mathbf{Z}^* = \mathbf{I_q} \otimes F$;

The variances and covariances of the random effects can be written as:

$$\text{var}(\mathbf{a}) = \mathbf{A} \otimes \mathbf{K}_a$$

$$\text{var}(\mathbf{p}) = \mathbf{I} \otimes \mathbf{K}_p$$

and $\text{cov}(\mathbf{a},\mathbf{p}) = 0$.

where $\mathbf{K_a}$ and $\mathbf{K}_p$ are the coefficients for the CF for a additive genetic and permanent environmental effects, respectively. The mixed model equations for the random

regression model with covariance functions (RR-CF-model) have a similar structure as a repeatability model, except that more coefficients are generated through the polynomic regression variables from Φ which are incorporated in **Z**. In the additive genetic effects part of te equations there is for each animal a diagonal block $\Phi_i'\Phi_i + a^{ii}\sigma_\varepsilon^2 K_a^{-1}$, and there are off diagonal blocks $a^{ij}\sigma_\varepsilon^2 K_a^{-1}$ with a^{ij} the $(i,j)^{th}$ element of the inverse of the numerator relationships matrix ($\mathbf{A}^{-1}$). The part for the permanent environmental effects is block diagonal with diagonal blocks equal to $\Phi_i'\Phi_i + \sigma_\varepsilon^2 K_p^{-1}$. Schematically, the mixed model equations will be like

$$\begin{bmatrix} X_i'X_i & \cdots & X_i'\Phi_i & \cdots & X_i'\Phi_i & \cdots \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\ \Phi_i'X_i & \cdots & \Phi_i'\Phi_i + a^{ii}s_e^2K_a^{-1} & \cdots & \Phi_i'\Phi_i & \cdots \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\ \Phi_i'X_i & \cdots & \Phi_i'\Phi_i & \cdots & \Phi_i'\Phi_i + s_e^2K_p^{-1} & \cdots \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} b \\ \vdots \\ a_i \\ \vdots \\ p_i \\ \vdots \end{bmatrix} = \begin{bmatrix} X_i'y_i \\ \vdots \\ \Phi_i'y \\ \vdots \\ \Phi_i'y \\ \vdots \end{bmatrix}$$

where the subscript i refers to those part of the equations for animal i. For the earlier example, we a 3-order CF with measurements at standardized ages [-1 0 1], $\Phi'\Phi$ is

The ASREML package can be used for random regression analysis. The latter package requires the user to define a regression model (e.g. a 3rd order polynomial regression on 'days in milk', and random regression is achieved by defining a random interaction term between animal and this polynomial regression term.

weight = herd poly(dim,2) !r poly(dim,3).animal

The first term is a polynomial regression of milk on days in milk (*dim*) as a fixed effect. This basically fits an average lactation curve equal for all animals. The random term indicates individual animal variation around this mean curve.

Alternatively, in ASREML, the regression coefficients (e.g. the Legendre regression on age as in the $\mathbf{F}$ matrix for each animal) can be constructed 'by hand' based on the age of the measurement and provided in a data file. ASREML allows estimation of variances and covariance components between these regression coefficients when they are taken as random. This covariance matrix should be equal to the **K**-matrix.

References

Belonsky, G.M. and Kennedy, B.W. 1988. Selection on individual phenotype and best linear unbiased prediction of breeding value in a closed swine herd. J. Animal Sci. 66:1124:

Falconer, D.S. and McKay, T.F.C. 1996. Introduction to quantitative genetics. 4th ed. Longman. Essex, England.

Gilmour, A.R. R Thompson, B.R. Cullis. and S. Welham 1998. ASREML – Biometrics Bulletin 3, NSW Agriculture 90pp.

Groeneveld, E., M. Kovac and T. Wang. 1990. PEST, a general purpose BLUP package for multivariate prediction and estimation. Proc. 4th World Congr. Genet. Applied Livest. Prod. 13:488.

Meuwissen, T.H.E. 1997. Maximizing the response of selection with a predefined rate of inbreeding. J. Animal Sci. 75:934.

Wray, N.R., and M.E. Goddard. 1994. Increasing long term response to selection. Genet. Sel. Evol. 26:431.