Genetic properties of the animal model

Introduction

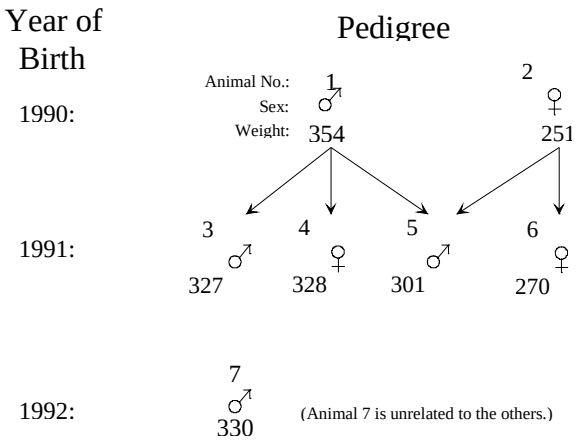
The important genetic properties of the BLUP and the animal model are due to the use of the Numerator Relationships Matrix (NRM), i.e. is the matrix with additive genetic relationships between animals. This matrix was originally used by Henderson to account for covariances between random effects, and therefore to use information from relatives in estimation of breeding value. However, important quantitative genetic properties such as accounting for selection over generations and accounting for inbreeding were revealed later. The essence is that the relationship matrix contains all information about the flow of genes through the population. It also allows an explicit dissection of genetic variation due to having different sires and/or dams, or due to differences from Mendelian sampling

This chapter will discuss some more genetic properties of BLUP EBV's, especially in relation to the NRM.

A BLUP example

The following example illustrates that BLUP selection uses selection index weights, based on the same principles as discussed in the previous paragraph.

Let us first ignore fixed effects, and write the BLUP equations for this example with observations taken as deviating from the overall mean (strictly giving BLP):



$$\hat{u} = [\quad Z'Z \quad + \quad A^{-1}\lambda \quad]^{-1} \quad (Y - \bar{Y})$$

$$\begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \end{pmatrix} = \left[\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 13/6 & 1/2 & -2/3 & -2/3 & -1 & 0 & 0 \\ 1/2 & 11/6 & 0 & 0 & -1 & -2/3 & 0 \\ -2/3 & 0 & 4/3 & 0 & 0 & 0 & 0 \\ -2/3 & 0 & 0 & 4/3 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & -2/3 & 0 & 0 & 0 & 4/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right]^{-1} \begin{pmatrix} 45.3 \\ -57.7 \\ 18.3 \\ 19.3 \\ -7.7 \\ -38.7 \\ 21.3 \end{pmatrix}$$

$$\begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \end{pmatrix} = \begin{pmatrix} .410 & -.030 & .117 & .117 & .127 & -.008 & 0 \\ -.030 & .435 & -.008 & -.008 & .135 & .124 & 0 \\ .117 & -.008 & .462 & .033 & .036 & -.002 & 0 \\ .117 & -.008 & .033 & .462 & .036 & -.002 & 0 \\ .127 & .135 & .036 & .036 & .421 & .039 & 0 \\ -.008 & .124 & -.002 & -.002 & .039 & .464 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .5 \end{pmatrix} \begin{pmatrix} 45.3 \\ -57.7 \\ 18.3 \\ 19.3 \\ -7.7 \\ -38.7 \\ 21.3 \end{pmatrix}$$

Note that BLP is the same as the classical selection index, except that there is a custom set of index weights for each candidate animal whose breeding value is to be estimated, depending on its particular set of information sources available.

A simplified versions of Henderson's proof why Mixed (or Random) model equations provide BLUP of breeding values goes as follows:

Start with the prediction of breeding values from regression, as cov(u,y)/var(y)

$$\hat{u} = G Z' (Z G Z' + R)^{-1} Y$$

dimensions: *a1 = aa ao (oa aa ao oo) o1* (*o*bservations and *a*nimals)

Note that $R = I\sigma^2_e$. "Divide" by G and rearranging gives:

$$\hat{u} = (Z'Z + G^{-1} I\sigma^2_e)^{-1} Z' Y$$

$$\hat{u} = (Z'Z + A^{-1} (\sigma^2_e/\sigma^2_g))^{-1} Z' Y \qquad \text{as } G^{-1} = A^{-1} \times \frac{1}{\sigma^2_g}$$

dimensions: *a1 = (ao oa aa) ao o1*

The BLUP equation for the same example are the BLP equations extended with equations for fixed effects (year effect in the example).

$$\begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{pmatrix}^{-1} \begin{pmatrix} X'Y \\ Z'Y \end{pmatrix}$$

$$\begin{pmatrix} \hat{b}_{mean} \\ \hat{b}_{1990} \\ \hat{b}_{1991} \\ \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \end{pmatrix} = \begin{bmatrix} 7 & 1 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 \\ 3 & 1 & 5 & 0 & 0 & 1 & 1 & 1 & 1 & -1 \\ \\ 1 & 1 & 0 & 19/6 & 1/2 & -2/3 & -2/3 & -1 & 0 & 0 \\ 1 & 1 & 0 & 1/2 & 17/6 & 0 & 0 & -1 & -2/3 & 0 \\ 1 & 0 & 1 & -2/3 & 0 & 7/3 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & -2/3 & 0 & 0 & 7/3 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & 3 & 0 & 0 \\ 1 & 0 & 1 & 0 & -2/3 & 0 & 0 & 0 & 7/3 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}^{-1} \begin{pmatrix} 2161 \\ 275 \\ 896 \\ \\ 354 \\ 251 \\ 327 \\ 328 \\ 301 \\ 270 \\ 330 \end{pmatrix} = \begin{pmatrix} 311.94 \\ -9.15 \\ -8.90 \\ \\ 28.26 \\ -28.85 \\ 18.34 \\ 18.77 \\ -0.87 \\ -22.40 \\ 0 \end{pmatrix}$$

We can understand more of the BLUP procedure for estimating breeding values if we write out individual equations. From the original MME equations we can look at particular rows for particular animals.

$$\begin{bmatrix} 7 & 1 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 \\ 3 & 1 & 5 & 0 & 0 & 1 & 1 & 1 & 1 & -1 \\ \\ 1 & 1 & 0 & 19/6 & 1/2 & -2/3 & -2/3 & -1 & 0 & 0 \\ 1 & 1 & 0 & 1/2 & 17/6 & 0 & 0 & -1 & -2/3 & 0 \\ 1 & 0 & 1 & -2/3 & 0 & 7/3 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & -2/3 & 0 & 0 & 7/3 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & 3 & 0 & 0 \\ 1 & 0 & 1 & 0 & -2/3 & 0 & 0 & 0 & 7/3 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}^{-1} \begin{pmatrix} \hat{b}_{mean} \\ \hat{b}_{1990} \\ \hat{b}_{1991} \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \end{pmatrix} = \begin{pmatrix} 2161 \\ 275 \\ 896 \\ \\ 354 \\ 251 \\ 327 \\ 328 \\ 301 \\ 270 \\ 330 \end{pmatrix}$$

For animal 6:

$$\hat{\mu} + \hat{b}_{1991} - \tfrac{2}{3}\hat{u}_2 + \tfrac{7}{3}\hat{u}_6 = 270$$

$$\rightarrow \hat{u}_6 = \tfrac{3}{7}(270 - \hat{\mu} - \hat{b}_{1991}) + \tfrac{2}{7}\hat{u}_2$$

$$\rightarrow \hat{u}_6 = \tfrac{3}{7}(270 - \hat{\mu} - \hat{b}_{1991} - \tfrac{1}{2}\hat{u}_2) + \tfrac{1}{2}\hat{u}_2$$

Therefore, the breeding value of animal 6 is estimated as a deviation of her phenotypic record from the expected mean and from her dams breeding value. The expected mean is the sum of all fixed effects plus the family mean. Since animal 6 has only one parent known, we take the deviation from one parent only (half sib family mean) and the weighting factor is the regression of the within half sib family deviation on the within half sib family breeding value.
The variance of the within half sib breeding value is 0.75 times the genetic variance. The weight is therefore equal to the following regression coefficient

$$\frac{\text{cov}(u_{whsf}, (y - y_{hsfmean}))}{\text{var}(y - y_{hsfmean})} = \frac{\tfrac{3}{4}\sigma_g^2}{\sigma_e^2 + \tfrac{3}{4}\sigma_g^2} = \frac{\tfrac{3}{4}h^2}{1 - h^2 + \tfrac{3}{4}h^2} = \tfrac{3}{7}$$

We can do the same thing for **animal 5:**

$$\mu + b_{1991} - \hat{u}_1 - \hat{u}_2 + 3\hat{u}_5 = 301$$

$$\rightarrow \hat{u}_5 = \tfrac{1}{3}(301 - \mu - b_{1991}) + \tfrac{1}{3}(\hat{u}_1 + \hat{u}_2)$$

$$\rightarrow \hat{u}_5 = \tfrac{1}{3}(301 - \mu - b_{1991} - \tfrac{1}{2}(\hat{u}_1 + \hat{u}_2)) + \tfrac{1}{2}(\hat{u}_1 + \hat{u}_2)$$

The breeding value of animal 5 is also estimated as a deviation of her phenotypic record from the expected mean and from the mean of his parents' breeding values. The expected mean is again the sum of all fixed effects plus the family mean. Since animal 5 has both parents known, we take the deviation from the full sib family mean and the weighting factor is the regression of the within full sib family deviation on the within full sib family breeding value.
The variance of the within full sib breeding value is 0.5 times the genetic variance. The weight is therefore equal to the following regression coefficient

$$\frac{\text{cov}(u_{whsf}, (y - y_{fsfmean}))}{\text{var}(y - y_{fsfmean})} = \frac{\frac{1}{2}\sigma_g^2}{\sigma_e^2 + \frac{1}{2}\sigma_g^2} = \frac{\frac{1}{2}h^2}{1 - h^2 + \frac{1}{2}h^2} = \frac{1}{3}$$

Writing out the equation for **animal 2** who is a parents with progeny:

$$\mu + b_{1990} + \frac{1}{2}\hat{u}_1 + \frac{17}{6}\hat{u}_2 - \hat{u}_5 - \frac{2}{3}\hat{u}_6 = 251$$

$$\rightarrow \hat{u}_2 = \frac{6}{17}(251 - \mu - b_{1990}) - \frac{3}{17}\hat{u}_1 + \frac{6}{17}\hat{u}_5 + \frac{4}{17}\hat{u}_6$$

$$\rightarrow \hat{u}_2 = \frac{6}{17}(251 - \mu - b_{1990}) + \frac{6}{17}(\hat{u}_5 - \frac{1}{2}\hat{u}_1) + \frac{4}{17}\hat{u}_6$$

Hence we see that the breeding value for animal 2 is estimated from her own record as deviation from the fixed effects (we have no family mean since she has no parents known), and from the estimated breeding values of her progeny.
Notice that the breeding values are corrected for the other parent (i.e. there's a correction for the mate), if the mate is known. In this case, the EBV of animal 5 is corrected for the contribution of his sire.
The weights for animal 2 are not very easy to recognize, but they are the same as selection index weights. We can check this by simplifying a bit the example, and ignore animal 5 as a progeny.

If animal 2 had only one progeny (animal 6), than her BLUP equation would look like

$$\hat{u}_2 = {}^{6}\!/_{14}(251 - \mu - b_{1990}) + {}^{4}\!/_{14}\hat{u}_6$$

but we saw earlier how the breeding value of animal 6 is estimated, therefore:

$$\rightarrow \hat{u}_2 = {}^{6}\!/_{14}(251 - \mu - b_{1990}) + {}^{4}\!/_{14}[{}^{3}\!/_{7}(270 - \mu - b_{1991} - {}^{1}\!/_{2}\hat{u}_2) + {}^{1}\!/_{2}\hat{u}_2]$$

$$\rightarrow \hat{u}_2 = {}^{6}\!/_{14}(251 - \mu - b_{1990}) + {}^{6}\!/_{49}(270 - \mu - b_{1991}) + {}^{4}\!/_{49}\hat{u}_2$$

$$\rightarrow \hat{u}_2 = {}^{7}\!/_{15}(251 - \mu - b_{1990}) + {}^{2}\!/_{15}(270 - \mu - b_{1991})$$

and the two weights can be found exactly by selection index to estimate the breeding value of an animal based on her own and her progeny's phenotypic record using a heritability of 0.5. This again shows that BLUP uses the same weights as selection index. The difference is that the means (fixed effects) are more properly corrected for.

BLUP accounts also for possible genetic differences. It takes deviations from expected genetic means, and the EBV is regressed towards the expected genetic mean. This expected genetic mean is in most cases the family mean. This is an illustration of a important properties of BLUP:

- It corrects for selection.
  Since the better parents have usually (more) offspring, we expect that the average breeding value goes up in later generations (in case of selection).
- The genetic trend is estimated from the average EBVs over time, i.e. the EBVs are plotted against the birth year of the animals.
- There is a correction a correction for assortative mating

As an illustration, compare the EBV with and without accounting for relationships across years

Offspring 1 and 2 are from sire 1 and offspring 3 is from sire 2. A heritability of 0,25 is assumed here.

1) No relationships across years: EBV's within a year sum to zero.

Year effect is overestimated

year		sire 1	sire 2	sire 3
1	300	350	300	250
		13	**0**	**-13**

		offspr. 1	offspr. 2	offspr. 3
2	333	365	325	310
		8	**-2**	**-6**

2)BLUP with relationships across years: EBV's in subsequent year are above zero.

Year effect estimate now only reflects environmental trend.

year		sire 1	sire 2	sire 3
1	300	350	300	250
		14	**-2**	**-13**

		offspr. 1	offspr. 2	offspr. 3
2	329	365	325	310
		13	**5**	**-4**

The Numerator Relationship Matrix (A)

Additive relationships are a measure of the proportion of genes, which are identical by descent, which are expected to be shared by two animals. Sewall Wright (1921) was responsible for the idea of tracing paths to establish the relationships among animals, although Malecot (1948) is given credit for the definition of relationships based on probabilities of individual genes at a locus being identical by descent.

The NRM is needed to account for the additive genetic covariances between records of related animals. In applications for breeding value estimation, use of the relationships matrix implies that information of related individuals will be used.

Accounting for genetic relationships also has proven to be very useful to account for selection. It is able to account for changes of means and variance after selection. The theory of deriving the relationships matrix can also be used for computing inbreeding coefficients for members of a population. This makes sense when we realize that inbreeding is computed from the additive genetic relationship between parents. The matrix is therefore essential for appropriate evaluation with data on complex pedigrees (more than one generation). Furthermore, the NRM can be used to optimize short-term selection decisions and predictions of response and inbreeding, since it can also in this respect give an account of the relationships among the current selection candidates. Understanding the structure of the relationships matrix helps in understanding genetic properties of the animal model.

The numerator relationships are equal to twice the coancestry, and they express the proportion of additive genetic variance that two individuals have in common. Ignoring epistatic effects and letting *a* equal the fraction of additive genetic variance, and *d* that of dominance variance, the generalized covariance for any sort of relationship is (Falconer, 1981):

$$Cov = aV_A + dV_D$$

For the covariance between P and Q (with parents A,B and C,D respectively), values for *a* and *d*

are given by:

$$a_{PQ} = 2f_{PQ}$$
$$d_{PQ} = f_{AC}f_{BD} + f_{AD}f_{BC}$$ where *f* is the coancestry

but we'll ignore *d (dominance)* hereafter

The coefficients of coancestry of two individuals reflects the probability that two gametes taken at random, one from each, carry alleles that are identical by descent (= inbreeding coefficient of their progeny should they be mated together).
Furthermore, at the diagonals, the NRM contains the coefficient of inbreeding. That is the probability that two genes at any locus in an individual are identical by descent.
The following equivalencies hold (see also page 4.3):

$$a_{ii} = (1 + F_i) = 2f_{ii}$$
$$a_{ij} = 0.5(a_{i,sireofj}, a_{i,damofj}) = 2f_{ij} \text{ for } i \neq j$$
$$F_i = 0.5a_{sd} = f_{sd}$$

The additive relationship (a_{ij}) is used as a measure of the covariance of breeding values between relatives. Wright's coefficient of relationship (*R*) is equivalent to the correlation between breeding values of two animals. For non-inbred animals these two measures are identical. *R* is given by:

$$R_{ij} = a_{ij} / \sqrt{a_{ii}a_{jj}}$$
$$= a_{ij} / \sqrt{(1 + F_i)(1 + F_j)}$$

Elements of *A* are the numerator of Wright's coefficient of relationship!

<u>The construction of the relationships matrix</u>

Systematic recurrent rules that are based on the flow of genes from generation to generation with individual animals being specified.

- Path coefficient method

  suitable for small pedigrees with few generations and little inbreeding

- Genomic table of probabilities

  is useful if dominance genetic effects are to be included in the animal model

- Recursively using the tabular method.

  where the pedigree is large and/or complicated, and/or where high levels of inbreeding are evident

For very large pedigrees it may be impossible to calculate *A* at all!

The tabular method

Step 1: Order pedigree list chronologically so that parents precede offspring. Base parents are considered unrelated and non-inbred.

Step 2: Working one row at a time, compute elements of A using the following relationships: For individuals *i* or *j*, elements of *A* are:

$$a_{ii} = (1 + F_i)$$
$$a_{ij} = 0.5(a_{ij'}, a_{ij''}) \text{ for } i \neq j$$
$$F_i = 0.5a_{sd}$$

where F_i is the inbreeding coefficient of animal *i*, *j'* = sire of *j* and *j''* = dam of *j*, and *s* and *d* refer to sire and dam. By knowing previous relationships, it is possible to calculate future relationships (given no missing information). The inbreeding coefficients (*F*) for base animals are zero. The following recursive function would calculate the additive genetic relationship between all animals.

```
 RECURSIVE function xnumrelmat(i,j) RESULT (value)
INTEGER PED
common PED(10000,2)
   if(i.eq.0 .or. j.eq.0) then
    value=0.0
    return
   endif
  IF(i.eq.j)then
   value=1+.5*xnumrelmat(PED(i,1),PED(i,2))
  elseIF(i.lt.j)then
   value=.5*(xnumrelmat(PED(j,1),i) + xnumrelmat(PED(j,2),i))
  elseIF(j.lt.i)then
   value=.5*(xnumrelmat(ped(i,1),j) + xnumrelmat(ped(i,2),j))
  endif
end function xnumrelmat
```

The algorithm to compute *A* using the tabular method is easy to implement. However, with a large number of animals, *A* may also be large and time consuming to compute (depending on the efficiency of your algorithm). Further, large *A* excludes direct inversion of *A* to obtain A^{-1} (needed for the MME) as a feasible option. Henderson (1976) developed a set of rules by

which A^{-1} can be built directly. Quaas (1976) has generalized these rules for large pedigrees and inbreeding. These findings contributed to routine use of *A* in animal breeding applications.

Construction of the inverse of the relationships matrix

Consider the case that u is a vector with additive genetic values (breeding values) of animals in a population. The variance of u can be given as var(u) = $A\sigma^2_a$ where A is the matrix with additive genetic relationships between animals and σ^2_a is the additive genetic variance. Now let us order the animals in u according to age, such that the oldest animals are first. A breeding value of an animal from which both parents are known in u can be written as $u_p = .5u_s + .5u_d + ms$, where u_p, u_s and u_d are breeding values of progeny, sire and dam, and ms is a part of the breeding value due to Mendelian sampling (within full-sib family variance).

We can define now a matrix P represents the transmission of genes. The vector of breeding values can be written as

$$u = P\,u + \phi$$

This can easiest be illustrated by dividing vector u into two parts, one being a part with breeding values of 'base' parents, which we define here as animals that have unknown pedigree, and the other part referring to animals that have both parents known. For simplicity we assume there are no animals with one parent known, but this could also be implemented. The vector of breeding values can be written as

$$\begin{pmatrix} u_b \\ u_p \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ P_{21} & P_{22} \end{pmatrix} \begin{pmatrix} u_b \\ u_p \end{pmatrix} + \begin{pmatrix} u_b \\ \phi \end{pmatrix}$$

where P_{21} refers to the flow from base animals in u_b to progeny in u_p and P_{22} refers to the passing of genes to progeny from parents that are also progeny themselves. Each row of P has therefore at most two nonzero elements: a .5 for each parent.

We can rewrite (x) as

$$(I-P)\,u = \phi$$

$$u = (I-P)\phi$$

The vector with breeding values u is now presented as a linear function of independent genetic values, which are in vector ϕ. This vector contains breeding values for base animals and Mendelian sampling values for offspring with two parents known. The additive genetic variance can be written as

$$var(u)= A = (I-P)^{-1} var(\phi) (I-P')^{-1} = (I-P)^{-1} D (I-P')^{-1} = TDT'$$

The matrix D is a diagonal matrix. If there is no inbreeding, then the values for each diagonal d_i are

d_i=1 if no parents are known,

d_i= .75 if one of the parents is known

d_i= .5 if both parents are known.

With d_i referring to the variance (proportional to σ_a^2) that has not been explained by known parents. We will first derive rules for A^{-1} because these rules are easiest and most often used in genetic evaluations. The inverse of the relationships matrix is

$$A^{-1} = [TDT']^{-1} = (I-P')D^{-1}(I-P)$$

P is a matrix with at most two elements of 0.5 in each row, one in the column for each parent, assuming there are no animals that have grandparents known but not their parents. This simple structure provides also simple rules to construct A^{-1}

The matrix (I-P) is lower triangular with all diagonals equal to 1 and, if parents are known, for each row only two parent-progeny off-diagonals equal to –0.5. Multiplication of (I-P') and (I-P) gives a symmetric matrix. It is constructed as: (I-P')(I-P)= I – (P + P') + P'*P.

This results is an identity matrix with a –0.5 on each known parent-progeny off-diagonal (created by subtraction of (P'+P), and a value of 0.25 added to each parental diagonal as well as the off

diagonal between 2 mates. In the A-inverse, these values are weighted by the diagonal matrix D^{-1}. The matrix D^{-1} contains diagonals being (with no inbreeding) 1, 4/3, and 2 for animals with none, one, and two parents known, respectively.

The rules for constructing A^{-1} for this non-inbreeding case are then based on the following table (where the i, j and k refer to row and column numbers of an animal, its sire and its dam)

pedigree knowledge:		element added to	value
no parents known			
(i,0,0)	animals' diagonal	i,i	1.0
one parent known			
(i,j,0) or (i,0,k)	animals' diagonal	i,i	1.0 * 4/3
	parent-offspring off-diagonal	i,j and j,i	
		or i,k and k,i	-0.5* 4/3
	parents' diagonal	j,j or k,k	0.25* 4/3
two parents known			
(i,j,k)	animals' diagonal	i,i	1.0 * 2
	parent-offspring off-diagonal	i,j; j,i; i,k and k,i	-0.5* 2
	parents' diagonal	j,j and k,k	0.25* 2
	mates' off-diagonal	j,k and k,j	0.25 * 2

Example:

pedigree

1	0	0
2	0	0
3	0	0
4	0	0
5	1	0
6	3	4
7	5	6

P =

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
0	0	0.5	0.5	0	0	0
0	0	0	0	0.5	0.5	0

(I-P) =

1.	0	0	0	0	0	0
0	1.	0	0	0	0	0
0	0	1.	0	0	0	0
0	0	0	1.	0	0	0
-0.5	0	0	0	1.	0	0
0	0	-0.5	-0.5	0	1.	0
0	0	0	0	-0.5	-0.5	1.

P+P' =

0	0	0	0	0.5	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0.5	0
0	0	0	0	0	0.5	0
0.5	0	0	0	0	0	0.5
0	0	0.5	0.5	0	0	0.5
0	0	0	0	0.5	0.5	0

P'*P =

0.25	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0.25	0.25	0	0	0
0	0	0.25	0.25	0	0	0
0	0	0	0	0.25	0.25	0
0	0	0	0	0.25	0.25	0
0	0	0	0	0	0	0

(I-P')(I-P)=

1.25	0	0	0	-0.5	0	0
0	1.	0	0	0	0	0
0	0	1.25	0.25	0	-0.5	0
0	0	0.25	1.25	0	-0.5	0
-0.5	0	0	0	1.25	0.25	-0.5
0	0	-0.5	-0.5	0.25	1.25	-0.5
0	0	0	0	-0.5	-0.5	1.

D =

1.	0	0	0	0	0	0
0	1.	0	0	0	0	0
0	0	1.	0	0	0	0
0	0	0	1.	0	0	0
0	0	0	0	0.75	0	0
0	0	0	0	0	0.5	0
0	0	0	0	0	0	0.5

$A^{-1} = (I-P')\, D^{-1}(I-P)$

1.33	0	0	0	-0.67	0	0
0	1.	0	0	0	0	0
0	0	1.5	0.5	0	-1.	0
0	0	0.5	1.5	0	-1.	0
-0.67	0	0	0	1.83	0.5	-1.
0	0	-1.	-1.	0.5	2.5	-1.
0	0	0	0	-1.	-1.	

The P matrix has another nice property P identifies in each row i the parents of i, P^2 identifies in each row of i the grandparents of i, etc.

Therefore, P^2 contains in each row 4 nonzero elements, each equal to 0.25 When this does not hold, some grandparents have to be missing from the pedigree, or there has been inbreeding (e.g. when there are only three elements, and one of the elements is 0.5, then that grandparent has been used twice in the pedigree of the same animal). Note that the sum of each row of P is equal 0, 0.5 or 1 for none, one or two parents known.

In general P^t identifies ancestors back to t generations, hence P^3 identifies great grand parents etc. If Pt= 0, than the first ancestors date back to less than t generations.

Building the NRM while accounting for inbreeding

If there is inbreeding, there is only one element of the previous derivation that changes, which is the diagonal D matrix. The diagonal represents the variance due to Mendelian sampling within family. The within family variance is equal to $\frac{1}{2}\sigma_a^2$ with no inbreeding, but this variance is smaller if the parents are inbred. In practice, it becomes much more difficult to derive the appropriate coefficients of A when there is inbreeding, because to know the inbreeding, one should know the additive genetic relationship between parents. In the previous section we only derived simple rules for the inverse of A. Obtaining coefficients A for two individuals (e.g. two parents), which is more difficult.

When an animal is inbred, it has an inbreeding coefficient F_i with $F_i = \frac{1}{2}a_{sd}$ where a_{sd} is equal to the additive genetic relationship between sire and dam. The variance of a breeding value of such an animal is then $var(u) = (1+F)\sigma_a^2 = \frac{1}{4}\, var(u_s) + \frac{1}{4} var(u_d) + \frac{1}{2}\, cov(u_s,u_d) + var(\phi)$

so that the within family variance is equal to

$$var(\phi) = (1+F_i)\sigma_a^2 - \frac{1}{4}(1+F_s)\sigma_a^2 - \frac{1}{4}(1+F_d)\sigma_a^2 - F_i\sigma_a^2 = \frac{1}{2}(1-\frac{1}{2}(F_s+F_d))\sigma_a^2$$

Hence, the within family variance is not reduced if the parents are related (and the full sibs are inbred), but only when the parents are inbred. The variance of an inbred animal is (1+F) and therefore the i^{th} diagonal of the relationships matrix is 1+F.

The diagonal for D becomes $d_i = 1 - 0.25a_{ss} - 0.25a_{dd} = \frac{1}{2}(1-\frac{1}{2}(F_s+F_d))$

And if only one parent known: $d_{ii} = 1 - 0.25a_{jj} = 3/4 - 0.25F_j$

Hence, to determine rules for A accounting for inbreeding, the diagonal elements of A have to be changed compared to the situation without inbreeding. It is a lot of work to determine A for large populations, but Quaas (1976) has given some rules to determine the diagonals of A efficiently, so that inbreeding coefficients and elements of A^{-1} with account for inbreeding can be derived from this.

We use again the structure of the relationships matrix and write

$A = TDT = LL'$

where $L = TD^{0.5}$ is a lower triangular matrix.

$T = (I-P)^{-1}$ is also lower triangular and it describes the flow of genes through the population. The T(i,j) element of T indicates the fraction of the genes that animal i has received from animal j. Diagonals of T are 1 and off-diagonals are nonzero between animals and their ancestors: equal to 0.5^n, where n is the number of generations between animal and ancestor.

This can be best seen in a numerical example.

Pedigree with inbreeding (animal, sire, dam)

1	0	0
2	0	0
3	1	0
4	1	2
5	3	2
6	3	5
7	3	6

$T = (I-P)^{-1}$

1.	0	0	0	0	0	0
0	1.	0	0	0	0	0
0.5	0	1.	0	0	0	0
0.5	0.5	0	1.	0	0	0
0.25	0.5	0.5	0	1.	0	0
0.3750	0.25	0.75	0	0.5	1.	0
0.4375	0.1250	0.8750	0	0.25	0.5	1.

D =

1.	0	0	0	0	0	0
0	1.	0	0	0	0	0
0	0	0.75	0	0	0	0
0	0	0	0.5	0	0	0
0	0	0	0	0.5	0	0
0	0	0	0	0	0.5	0
0	0	0	0	0	0	0.4375

$L = = TD^{.5}$

1.0000	0	0	0	0	0	0
0	1.0000	0	0	0	0	0
0.5000	0	0.8660	0	0	0	0
0.5000	0.5000	0	0.7071	0	0	0
0.2500	0.5000	0.4330	0	0.7071	0	0
0.3750	0.2500	0.6495	0	0.3536	0.7071	0
0.4375	0.1250	0.7578	0	0.1768	0.3536	0.6614

$A=T*D*T' = LL'$

1.	0	0.5	0.5	0.25	0.375	0.43
0	1.	0	0.5	0.5	0.25	0.125
0.5	0	1.	0.25	0.5	0.75	.875
0.5	0.5	0.25	1.	0.375	0.31	0.281
0.25	0.5	0.5	0.3750	1.	0.75	0.625
0.3750	0.25	0.75	0.3125	0.75	1.25	1.
0.4375	0.1250	0.8750	0.2813	0.6250	1.	1.375

An algorithm to determine inbreeding coefficients for large pedigrees can be based on this principle (Quaas, 1976). Later, algorithms have been described, e.g. by Meuwissen and Luo (1992) as described in the book by Mrode (1996)

Quaas algorithm

The (i,j) element of A is now obtained as the multiplication of the i^{th} row of L^{-1} with the j^{th} row of L^{-1} The inbreeding coefficient of an animal can be computed as Fi= $0.5a_{sd}$ = where s and d are the rows in L of the sire and the dam. We do not have to store all rows of L to compute inbreeding coefficients, which would require a lot of memory space for large populations. Quaas (1976) has shown that the triangular structure of L can be efficiently exploited to calculate F efficiently. Therefore, each column of L will be computed and only one column at a time will be kept in memory (say in array V). The squared elements will be accumulated in another array (say array U) and working from columns 1 to N (N= number of animals), the accumulations in U will contain the sum of the squared elements for each row. Therefore, U contains after completion the diagonal elements of the relationship matrix A and the inbreeding coefficient for animal i is then

The i-th column of L can be made for each animal, and the sum of the squared elements can be accumulated giving each time that li is formed, the complete pedigree list has to be read, which is most of the work of this method. Even if a new generation of progeny is added to the pedigree list, we have to read also the old list of pedigree to create the columns of L for the new animals.

After the inbreeding coefficients have been determined, the elements of D can be computed as di- = 1 / [1-(1+Fs)-(1+Fd)]. We use then the rules of [1-5] to create the elements of the A matrix.

The algorithm to determine inbreeding coefficients the routine builds a column of L-matrix, the routine runs for an animal a column in array v and accumulates the sum of squared elements in array U= diagonals of A

```
!N=total number of animals
DO 10 I=1,N        !  for each animal
s=sire of i
d= dam of i
IF s NE 0 and d NE 0 THEN
XX= 1-0.25*U(s)-0.25*U(d))
IF s NE 0 and d = 0 THEN
XX= (1-0.25*U(s))
IF s = 0 and d NE 0 THEN
XX= (1-0.25*U(d))
IF s = 0 and d = 0 THEN
XX= 1
END IF
V(I)= sqrt(XX)    !V is column of L matrix
U(I)= U(I)+ XX*XX        !U accumulates
DO 2 J=I+1, N
s= sire of J¸
then other off diagonals
d= dam of J
IF s GE I and d GE I THEN
XX=  .5*(V(s)+ V(d))
IF s GE I and d LT I THEN
XX= .5*V(s)
IF s LT I and d GE I THEN
XX=  .5*V(d)
IF s LT I and d LT I Then
XX= 0
END IF
V(J)=XX
U(J)= U(J)+ XX*XX
2   CONTINUE
10  CONTINUE
```

BLUP accounting for changes of variance due to selection

It has been shown that variance in a population change drastically after one round of selection. Since this change is large and immediate, it is important that the models that we use for genetic evaluation and for prediction of response account for the Bulmer effect. This section describes how BLUP and REML account for the Bulmer effect using the mixed model.

The genetic model assumed for showing these properties is (again) the infinitesimal model, i.e. the only changes in genetic variance due to selection are due the gametic disequilibrium and inbreeding, but not to change of gene frequencies. This assumption is reasonable when short-term responses are considered (Bulmer 1971).

Kennedy and Sorensen (1988) have given an excellent explanation of the genetic properties of the mixed model, and showed that the mixed model accounts for changes of genetic variance after selection. They refer to Henderson's (1975) paper on properties of BLUP under a selection model, where he used an argument of Pearson (1903) to show that the mixed model equations yield BLUP in case of selection, given that all the data used in making selection decisions are included in the analysis. Pearson's result gives the variance of a variable x after selection on a correlated variable y as

$$\sigma^2_{sx} = \sigma^2_x - \mathrm{cov}(x,y)^2 \sigma_y^{-2}(1-(\sigma^2_{ys}/\sigma^2_y)) = \sigma^2_x - \mathrm{cov}(x,y)^2 \sigma_y^{-2}\sigma_x^{-2}(1-(1-k))\sigma^2_x = \quad (1\text{-}r^2k)\,\sigma^2_x$$

where r is the correlation between x and y and k is the proportional reduction in phenotypic variance in the selected group. Using Pearson's result, we can write the genetic variance in a selected group (after mass selection) as $(1\text{-}h^2k)\,\sigma^2_A$, since the correlation between additive genetic value and phenotype is equal to h (the square root of the heritability). Using formula (x] and assuming equal selection intensities gives the genetic variance in the progeny of selected parents as $(1\text{-}\tfrac{1}{2}h^2k)\,\sigma^2_A$, a result also found by Bulmer (1971).

Kennedy and Sorensen (1988) point out that Henderson (1975) has shown that BLUP correctly accounts for selection because Pearson's rules for conditional variances apply. That is, if the base population is unselected, and the next generation descends from the best parents of the

previous generation, than both generations can be evaluated in an unbiased fashion. Although the 2nd generation animals are not random animals (but from selected parents), they are unselected within their families. In statistical terms, the distribution (i.e. mean as well as variance) of the second generation conditional on the first generation is not affected by selection. In quantitative genetic terminology, this conditional variance is equal to the within family variance, or the Mendelian sampling variance. The parental contributions are evaluated unbiasedly if their contemporaries that they were selected from are in the model

In an earlier chapter we have written A as TDT' where T describes the flow of genes over the generations and D described the variance of the part of the breeding value of each animal that is not explained by its ancestors. The breeding values can be written as a = Tϕ, i.e. they are a linear combination of values in ϕ, which are not affected by selection. The matrix T describes how a given breeding value is a linear function of effects of ancestors that each was not affected by selection. This assumes that unknown ancestors not in the model (not in a) were unselected. For our example, T was

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 1 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0.5 & 0 & 1 & 0 & 0 \\ .375 & .25 & .75 & 0 & .5 & 1 & 0 \\ .4375 & 0.125 & 0.875 & 0 & 0.25 & 0.5 & 1 \end{pmatrix}$$

Therefore, as long as the base population is unselected, and the Mendelian sampling terms are not affected by selection, as is the case with an infinitesimal model, then var(a)= $A\sigma_A^2$, even though selection may have operated in subsequent generations, and the effects of linkage disequilibrium and inbreeding are accommodated (Kennedy and Sorensen, 1988).

An example might illustrate that Pearson's rules for conditional variance lead to the same results as Bulmer (1971). The example is (slightly) adapted from Henderson (1982).

Consider a model y = a + e, being phenotype explained by additive genetic and environmental effects. Heritability is 0.30. In the genetic part of the model we can replace the additive genetic

value by the value of the 2 gametes: $y_i = g_i^p + g_i^m + e$.

Now consider 2 sires and we are interested in selecting one of them (based on its phenotype)

The variance of the gametic effects before selection is:

$$G = var(g) = var\begin{pmatrix} g_1^p \\ g_1^m \\ g_2^p \\ g_2^m \end{pmatrix} = \begin{pmatrix} 0.15 & 0 & 0 & 0 \\ 0 & 0.15 & 0 & 0 \\ 0 & 0 & 0.15 & 0 \\ 0 & 0 & 0 & 0.15 \end{pmatrix}$$

Now suppose we select on the phenotypes. After selection among the 2, the variance of the phenotypes after selection is (based on order statistics) equal to $var(y_s) = V_s = \begin{pmatrix} 0.6817 & 0.3183 \\ 0.3183 & 0.6817 \end{pmatrix}$.

Using Pearson's rule, the variance among the gametes after selection is equal to

$$G_s = G - cov(g,y).var(y)^{-1}.(var(y)-var(y_s))var(y)^{-1}cov(y,g) =$$

$$= G - BV^{-1}(V-V_sV^{-1}B \qquad \text{where } B = cov(y,g)$$

$$= \begin{pmatrix} .1428 & -.0072 & .0072 & .0072 \\ -.0072 & .1428 & .0072 & .0072 \\ .0072 & .0072 & .1428 & 0.0072 \\ .0072 & .0072 & -.0072 & .1428 \end{pmatrix}$$

Note that in this result, the covariance between gametes on the same animals is negative after selection. Furthermore, the variance of breeding values is $var(a_i) = var(g_i^p) + var(g_i^m) + 2cov(g_i^p, g_i^m) = .1428+.1428+2*(-.0072)=0.2712$. This agrees with the variance reduction expected from Bulmer's formula. For selecting 1 out of 2, i= .564 and x=0, therefore k= 0.318. The variance of a_i after selection on phenotype is $(1-h^2k)\sigma_A^2 = (1-.3*.318)*.3 = 0.271$. The additive genetic variance in the progeny (a_p) of this selected animal (giving only the paternal gamete, to mate with an unselected maternal gamete) is $var(a_p) = var(g_i^p) + var(g_i^m) = 0.2928$. Hence, 50% of the loss in genetic variance due to selection is recovered in the next generation.

Which variance should be used in BLUP?

In the previous we showed that the 'base population' variance is not the same as the variance in later generations. The genetic variance in a given generation can be divided in a part coming from variation between base parents and a part resulting from Mendelian sampling. This causes a somewhat complicated dilemma: what is the relevant genetic variance in a selection program. Is it the variance of the base population, or the variance at the current generation? The answer is that both are relevant. The base population variance remains important, because the Mendelian Sampling component of the variance (within FS family variance) remains to be 50% of the base population variance. The sire variance as well as the dam variance (i.e. the between family variance) are significantly reduced in a selection program. Estimates of genetic variance based on between family variance components are therefore not appropriate for populations under selection. Instead, and animal model is used, and variance components are estimated with REML. In principle, if all genetic relationships are included back to the base population, and if all data is used in the analysis that was included in the selection decisions, REML will provide estimates of the base populations variance (Sorensen and Kennedy, 1984). It should be noted that including relationships does not only account for gametic disequilibrium due to selection, buy also for reduction of variance due to inbreeding and the buildup of covariances of related animals. Of course a debate exists on what is the base population (how many generations do we need to go back), and we should realize that practically it is usually not possible to include relationships and data since the start of selection. The conclusion is that genetic variances that are estimated can be expected to be somehow biased by selection, generally more if more of the selection history is omitted from the analysis (Van der Werf and De Boer, 1990).

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:

Bulmer, M.G. 1971. The effect of selection on genetic variability. American Nat. 105:201.

Bulmer, M.G. 1980. The mathematical theory of quantitative genetics. Clarendon Press. Oxford.

Henderson, C.R. 1975. Best Linear Unbiased estimation and prediction under a selection model. Biometrics 31:423-447.

Henderson, C.R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69-83.

Kennedy, B.W. and D. Sorensen. 1988. Proc. of the Conference on Quantitative Genetics.

Meuwissen, T.H.E., 1991. Reduction of selection differentials in finite populations with a nested full-half sib family structure. Biometrics 32: 875-887.

Meuwissen, T.H.E. and Luo.Z. 1992. Computing inbreeding coefficients in large populations. Genet. Select. Evol. 24:305-313

Mrode, R.A. 1996. Linear models for the prediction of animal breeding values. CAB Int.

Kennedy B.W. and D. Sorensen. 1987. Properties of mixed model methods for prediction of genetic merit. In: Proc. Of the second. Int. Confer. On Quant. Genet. Eds. Weir, B.S. et al. Sinauer Publ. Sunderland, Mass.

Rawlings, J.O., 1976. Order statistics for a special class of unequally correlated multinormal variates. Biometrics 32: 875-887.

Quaas, R.L.,1976 Computing the diagonal elements and inverse of a large numerator Relationship Matrix. Biometrics 32:949-953

Quaas, R.L.,1988 Additive genetic model with groups and relationships. J.Dairy Sci. 71:1338-1345.

Sorensen, D.A., and B.W. Kennedy. 1984. Estimation of genetic variance from selected and unselected populations. J. Anim. Sci. 59:1213.

Van der Werf, J.H.J., and I.J.M. de Boer. 1990. Estimation of additive genetic variance when base populations are selected. J. Animal Sci. 68: 3124-3132.

Genetic Evaluation across Breeds

Application of mixed models has become an attractive tool to evaluate animals in actual breeding programs of breeding organisations. 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. It has to be recalled that the quality of evaluations always depends on the model that is used. Models can be extended to account for more complicated effects, such as different breeds.

The mixed model equations in a general form look like:

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

where b is the vector with fixed effects with design matrix X

u is the vector with random effects with design matrix Z

Note that the expectation for the variables are

$E(\mathbf{y}) = \mathbf{Xb}$

$E(\mathbf{u}) = 0$

$E(\mathbf{e}) = 0$

the variances are: $var(\mathbf{u}) = \mathbf{G}$

$var(\mathbf{e}) = \mathbf{R}$

$var(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$

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

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z+\lambda A^{-1} \end{pmatrix}\begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

Accounting for genetic group effects

A model for genetic evaluation needs to account for genetic groups when the animals in the data set come from widely divergent sources. The mixed model assumes that the breeding values to be estimated come from a homogeneous population (E(u) = 0), and all have the same expected mean, that is for the animals with unknown parents (the expectation of animals with parents known is equal to the parental average EBV). Animals without parents are called ‘base animals’, and if they are not from a homogeneous population, genetic groups are needed to distinguish between different genetic levels of base animals.
Notice that the relationships matrix takes care of all genetic differences due to selection since the base population. For example, in analyzing data of a selection experiment with a high an low line, but both stemming from the same base population, genetic groups are not needed as long as pedigree and data since the start of selection is included in the analysis. Genetic groups are therefore needed for those cases where we can’t explain genetic differences between animals by pedigree and data. This is typically the case if animals arise from different breeds or populations.

Consider Finnsheep (F, average litter size about 3) mixed in with Merinos (M, lucky to get one). Litter size is a lowly heritable trait, and so any genetic evaluation ignoring breed will regress all EBV's to close to the average - clearly wrong, as the breed effect on litter size is strong and reliable.
The solution is to fit animal source as a fixed effect. With ongoing breeding, individual animals can be a mixture of sources - but this is not a problem. Here is an example of entries in the X matrix for the F(inn) and M(erino) fixed effects:

Type of animal	F effect	M effect
Finn	1	0
Merino	0	1
F x M	½	1/2
M x (FxM)	¼	3/4

Other examples of genetic groupings are:

"animals imported from Canada"

“animals born before 1980"

The EBV of an animal is now the sum of it’s EBV (random effect) estimate within the group, with added to that the genetic group effect. For example, if the fixed effect estimate of F is +0.7 compared to M, animals fully belonging to the Finn breed get 0.7 added to their random within breed breeding value, so that EBV’s of Finns and Merino’s can be directly compared to each other.

Additive genetic models with groups: Modified equations

The outline with genetic groups as fixed effects as outlined above is straightforward if all animals belonged only to one genetic group. However, often they belong to two or more genetic groups, since the parents can be from different origin. In a crossbreeding context, an animal can have a Merino dam, and his sire can be a cross of Border Leicester * Poll Dorset.

Quaas (1988) has presented the basic structure of additive genetic relationships within a population. Based on this structure, rules for creating the relationships matrix were derived. This theory can be extended to the situation of having different means for different groups of base animals, leading to a coherent and operationally simple approach to the problem of genetic grouping in animal evaluations.

The problem to be dealt with is that not all base animals have equal means or, in other words, equal expectation. Realize that usually in mixed models the expectations of the random effects is equal to 0. When breeding values of sires did not have the same expectation, e.g. because sires

were from different breeds, the problem was solved by incorporating genetic groups in the model. Hence, instead of the model y=Xb + Zu + e, we used the model y= Xb + ZQg + Zs + e. The vector g referred to fixed group effects and the vector s referred to random sire effects within genetic groups. The matrix Q related sires to groups. The breeding value was u= Qg+ s, and the mixed model was well defined again because the expectation of the vector of random effects was equal to 0. In fact, records were linked to fixed group effects, and random effects were predicted after correction for fixed groups.

Now we could again define genetic groups, and random effects of base animals within groups. Groups could then be defined e.g. according to the breed and/or the birth year of the base animal. The problem with such a model would be to define the incidence matrix for groups, i.e. how observations on animals are related to groups. For example, an animal could have ancestors (base animals) from different breeds and these ancestors could be born in different years. The breeding value (and the record) of such an animal would then be linked for say 0.25 to the mean of breed 1 in year 1978 and for 0.25 to breed 2 in year 1982, and for 0.50 to breed 2 in year 1984. Because we basically want to derive the contribution of each group to the genome of each animal that we want to evaluate, it would be an advantage if we could make use of rules for defining these coefficients systematically. From the relationship of a certain animal to the groups we want to derive the relationship of its progeny to these groups. The procedure developed by Quaas (1988) shows such a systematical approach in a very elegant way.

For the base animals, where in principle, we can not determine their pedigree, we can often figure which breeds or groups are on the origin. We can therefore assign unknown parents to such animals, arising from genetic groups. Such parents are indicated as phantom parents. For example, an average milking cow could have assigned a “phantom” sire to the group “sires born between 1985 and 1990”, whereas its dam would be assigned to “cows born in 1992”. If we assign all such phantom parents to a genetic group, equal to their expectations, than descendent are linked to genetic groups through the pedigree. In fact, we can treat phantom parents as normal part of the pedigree (i.e. using Henderson’s ruls fro the coefficients). This creates a very flexible framework to assign animals to genetic groups.

The model is written as:

$$\begin{bmatrix} u_b \\ u \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ P_b & P \end{bmatrix} \begin{bmatrix} u_b \\ u \end{bmatrix} + \begin{bmatrix} u_b \\ w \end{bmatrix} \qquad \mathbf{1}$$

where $\mathbf{P}_b$ relates the animals to the unknown parents and **P** relates the known animals as in 1. Furthermore, the expectation of unknown parents is $E(\mathbf{u_b})=\mathbf{Q_b g}$, where **g** is a vector with genetic group effects and $\mathbf{Q^b}$ assign base animals to genetic groups. The expectation of **u**, i.e. the vector of breeding values of known animals, is then:

$E(\mathbf{u})= (\mathbf{I-P})^{-1}\mathbf{P_b Q_b g}= \mathbf{Qg}$ (Quaas, 1988).

Quaas shows with numerical examples that the matrix **Q** exactly relates the breeding values of all known animals to the genetic group effects. Hence, if there are n animals and p genetic groups, than **Q** is a n x p matrix and the $(i,j)^{th}$ element of **Q** reflects the fraction of the genes of animal i are originating from group j.

The mixed model including grouping would now be

$$\mathbf{y= Xb + ZQg} + \mathbf{Za^* + e}$$

where **ZQ** relates records to groups. The vector $\mathbf{a}^*$ is a random vector with breeding values corrected for group effects (hence, as if base animals all had equal expectation). The expectation is $E\mathbf{a}^*= \mathbf{0}$ and $var(\mathbf{a}^*)= \mathbf{A}\sigma^2_a$, and the vector of breeding values for animals is $\mathbf{a}= \mathbf{Qg}+ \mathbf{a}^*$.

Because in this model **g** is just a common fixed effect, the mixed model equations would be:

$$\begin{bmatrix} X'X & X'ZQ & X'Z \\ Q'Z'X & Q'Z'ZQ & Q'Z'Z \\ Z'X & Z'ZQ & Z'Z + \alpha A^{-1} \end{bmatrix} \begin{bmatrix} b \\ g \\ a^* \end{bmatrix} = \begin{bmatrix} X'y \\ Q'Z'y \\ Z'y \end{bmatrix} \qquad \mathbf{2}$$

These equations are in principle correct in the sense that it takes into account that all animals are in different ways related to the genetic groups. In practice such equation would cause problems, unless a systematic way is found to create the **Q** matrix. This was solved by Quaas by 1) writing the above equations in another way which he calls 'modified equations' and 2) by realizing that

modified equations can be set up by simple rules.

$$\begin{bmatrix} X'X & 0 & X'Z \\ 0 & \alpha Q'A^{-1}Q & -\alpha Q'A^{-1} \\ Z'X & -\alpha A^{-1}Q & Z'Z + \alpha A^{-1} \end{bmatrix} \begin{bmatrix} b \\ g \\ a \end{bmatrix} = \begin{bmatrix} X'y \\ 0 \\ Z'y \end{bmatrix}$$

Hence, the grouping equations are now basically and extension of the relationships matrix.
The rules to create grouping equations are summarized as

- Assign phantom parents to base animals
  (if only one parent known, assign another phantom parent
- Determine for each phantom parent to which genetic group it belongs
- Build the mixed model equations using the pedigree, including phantom parents

  The matrix A^{-1} is obtained by the usual rules for obtaining the inverse of the relationship matrix. A list of pedigrees, consisting of only actual animals, but with unknown ancestors assigned to groups is set up. For the ith animal, calculate the inverse (b_i) of the variance of Mendelian sampling as:

  $b_i = 4/(2 +$ number of parents of animal i assigned to groups)

  Then add:

  b_i to the (i,i) element of A^{-1}

  $-b_i/2$ to the (i,s), (i,d), (s,i) and (d,i) elements of A^{-1}

  $b_i/4$ to the (s,s), (s,d), (d,s) and (d,d) elements of A^{-1}

Example (from Mrode, 1996)

By way of example the modifications of a pedigree structure needed to set up the above NRM is shown.

Calf	**Sire**	**Dam**
1	unknown	unknown
2	unknown	unknown
3	unknown	unknown
4	1	unknown
5	3	2
6	1	2
7	4	5
8	3	6

This can be rewritten assigning unknown sires to one group and unknown dams to another group.

Calf	Sire	Dam
1	G1	G2
2	G1	G2
3	G1	G2
4	1	G2
5	3	2
6	1	2
7	4	5
8	3	6

The NRM is then constructed using the above rules, in this case n = 8 animals and p = 2 groups. The solutions to the modified MME have a problem in that the genetic group effects are still fixed effects and some restrictions on their solutions may be needed.

Non additive genetic effects

In the analysis of data across populations, one might expect non-additive effects. Depending on the crossbreeding group, different coefficients for dominance (or heterosis) and epistatic (or recombination) effects are expected. A straightforward way to account for such effects is to include them in the model as linear regression coefficients (Van der Werf and De Boer, 1989). The additive genetic breed effects will be a regression of phenotype on proportion of genes of a particular breed in the animal making the record. Similarly, dominance is related to heterozygosity of the animals' genome, and epistasis is related to heterozygosity of the parents. A problem is often that not all crossbreeding types are evenly (or even at all) represented. The regression model is not very robust to such sub-optimal designs. Depending on the dataset, one might 'pre-estimate- crossbreeding effects and pre-correct the data. In estimating crossbreeding effects, is useful to check the estimability of the crossbreeding parameters (often A, D and E have a quite high sampling correlation). It is also useful to compare a regression model with a model with each crossbreeding type as a fixed effect. The latter model does not rely on any assumptions of genetic effect in the model. If the expected mean for a particular crossbreeding group from the regression model deviates from the breed group model (other than by sampling), than the regression model might lack a certain effect (e.g. maternal effect or heterosis).
Finally, when looking at crossbreeding models at single, or two locus level, it is quickly clear that different crossbred groups can be expected to have different genetic variance (both additive and non-additive). To some extent, the infinitesimal genetic model is not compatible with dominance and inbreeding depression (see next).

Non-additive genetic effects and finite locus models

For genetic models at the locus level non-additive genetic effects are clearly defined as dominance and epistasis. At the population level, for multiple loci, we can define dominance effects and dominance variation. Assuming small contributions from many unlinked loci, genetic covariance between individuals in a non-inbred random mating population is a linear function of the genetic variance components and genetic relationships (Cockerham, 1954).

Including dominance in a mixed model can be done as a 'polygenic' dominance effects, with variance covariance matrix equal to the dominance variance multiplied by the dominance relationships matrix. Dominance relationships can be derived from the additive relationships among the parents

$$d_{xy} = 0.25(a_{s_x s_y} a_{d_x d_y} + a_{s_x d_y} a_{d_x s_y})$$

or using an algorithm for large populations (Hoeschele and Van Raden, 1991) similar to Henderson's rules.

Inbreeding has two effects in a dominance model:

1) with dominance existing, inbreeding will depress phenotypic performance (inbreeding depression)
2) inbreeding complicates the genetic covariance structure of the population. In noninbred populations, the genetic covariance is a function of additive genetic and dominance variance. In inbred populations, additional terms need to accommodate: dominance variance in a completely inbred population, covariance between additive and dominance effects in completely inbred populations and the sum (over loci) of squared effects of complete inbreeding depression (De Boer and Hoeschele, 1993).

A conceptual problem with the infinitesimal model is that it can not accommodate properly inbreeding depression. Inbreeding depression is the result of loss of dominance due to loss of heterozygous loci. We could model dominance effects, but a finite amount of inbreeding depression (decrease of mean per percentage of inbreeding) can not be explained by dominance effects at an infinite number of loci.

De Boer and Hoeschele (1993) have derived rules for exact genetic covariance matrices, including dominance effects, and accounting for inbreeding and inbreeding depression. They noticed that this method is not feasible for larger populations, and compared it with approximate

BLUP, including dominance, but ignoring inbreeding. The inbreeding depression was accommodated as a linear regression of phenotype on inbreeding coefficient. There appeared to be small differences between approximate BLUP and exact BLUP using a genetic model with 64 unlinked loci, biallelic, and with complete dominance.

Conclusion

In analysis of crossbred data, the first worry is to have the first moments right, i.e. the model has to account for breed effects and possible non-additive effects like heterosis and recombination loss. It is important here to realize that breed differences are additive effects and should be added on to within breed effects of additive effects, in order to obtain across breed EBV's.
A second, and of secondary importance, worry is to have the variances right. The fewer loci in the underlying genetic model, the more change that different genotypes (crossbred groups) have different genetic variance. However, as most traits are assumed to be regulated by a large number of loci, and as breed differences (and allele frequencies) are generally not expected to be very high (unless for more extreme crosses), it may be reasonable to assume homogeneity of variance across crossbred groups.

References

Van der Werf, J.H.J. and De Boer, W. 1989. J. Dairy Science. 72:2606.
Quaas, R.L.,1988 Additive genetic model with groups and relationships. J. Dairy Sci. 71:1338-1345.
Cockerham (1954). Genetics 39:859.
De Boer and Hoeschele (1993). Theor. Appl. Genet. 86:245.