Chapter 14
Genetic Grouping

Julius van der Werf

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 and 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

Examples of genetic groupings are:

- breed origin
- animals imported – by country of origin
- animals' birth year

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.

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 animals do not have the same expectation, e.g. because animals are from different breeds, the problem can be solved by incorporating genetic groups in the model.

Hence, instead of the model $y=Xb + Zu + e,$

we used the model $y= Xb + ZQg + Za + e.$

The vector g refers to fixed group effects and the vector a referred to random animal effects within genetic groups. The matrix Q relates animals to groups and **ZQ** relates records to groups.
The estimated breeding value is

$$\hat{u} = Q\hat{g} + \hat{a},$$

and the mixed model is well defined again because the expectation of the vector of random effects is equal to 0. In fact, records are linked to fixed group effects, and random effects are predicted after correction for fixed groups.
The expectation is E**a**= **0** and var(**a**)= $\mathbf{A}\sigma^2_a$, and the vector of breeding values for animals across groups is $\hat{u} = Q\hat{g} + \hat{a}$. Because in this model **g** is just a common fixed effect, the mixed model equations would be:

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

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.

The modified equations are derived by pre-multiplying the coefficient matrix and the right hand side in (2) by

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

and subsequently $\begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -Q & I \end{pmatrix}$ its inverse $\begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & Q & I \end{pmatrix}$ between coefficient matrix and solution vector gives

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

These modified equations have a number of advantages. Firstly, the off-diagonal blocks of groups by fixed effects are zero, as is the right hand side for groups. Secondly, the solutions to the animals within groups are giving across group breeding values (*u* rather than *a*). This has an important numerical advantage in solving mixed model equations. However, the main breakthrough of these modified equations is the insight it gives into a flexible way to define genetic groups. From the equations, you can see that the grouping equations look very similar to equations of animals with no data. The only coefficients are related to the relationship matrix. This was noted by Quaas (1988) and he discovered that this is a key to defining genetic groups for all animals. As we see in the next section, the genetic groups are like ancestors and every animal will have a relationships through such an ancestor through its pedigree. This gives a 'natural way' to define group effects (the Q matrix), something that could otherwise become very tedious, as we see next.

Assigning animals to genetic groups

Groups can 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 1998 and for 0.25 to breed 2 in year 2002, and for 0.50 to breed 2 in year 2004. Because we basically want to derive the contribution of each group relevant to the genetic make up 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, similar to the systematic way of ancestors contributing to an animal through the pedigree. 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.
Base animals, for whom in principle we can not determine their pedigree, will have to be assigned to genetic groups, according to their suspected origin. In the grouping strategy proposed by Quaas, not the base animals themselves ar assigned to groups, but they are assigned unknown 'dummy' parents,

who ar assigned to groups. Such dummy 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 descendents 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 rules fro the coefficients). This creates a very flexible framework to assign animals to genetic groups.

The model is written as:

$$\begin{bmatrix} u_b \\ u \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ P_b & P \end{bmatrix} \begin{bmatrix} u_b \\ u \end{bmatrix} + \begin{bmatrix} u_b \\ w \end{bmatrix} \qquad \mathbf{3}$$

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

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

Quaas shows with numerical examples that the matrix $\mathbf{Q}$ 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 $\mathbf{Q}$ is a n x p matrix and the $(i,j)^{th}$ element of $\mathbf{Q}$ reflects the fraction of the genes of animal i are originating from group j. Hence, genetic groups are like ancestors. As with 'real pedigree', it is not necessary to work out all relationships in a pedigree. Only direct relationships are taken into account, and other relationship are automatically implied, was we have see with rules for building A-inverse. Similarly, we do not need to worry about genetic groups of animals that have parents known, as their expected genetic mean is determined by the parent average. Only animals with one or two unknown parents need an assignment to a genetic group, or better, the missing parents needs be placed in a group where it most likely belongs to

Rules for genetic grouping are derived from the same rules as those for building the (inverse) relationship matrix. If parents are known, we proceed as before, with the normal rules for the relationship matrix. If one or two parents are unknown, we define a genetic group for that unknown parent and treat that genetic group as an ancestor. The only difference with a real ancestor is that genetic groups are fixed effect whereas real ancestors are treated as random

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 i^{th} animal, calculate the inverse (b_i) of the variance of

Mendelian sampling as:

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

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}

Note that when both parents are known, none has to be assigned to groups and $b_i = 2$. The coefficients added are then 2, -1 and ½, i.e. the usual coefficients for the NRm for 2 parents known. If only one parent is known, $b_i = 4/3$ and the coefficients are 4/3, -2/3 and 1/3, i.e. again the same as the case for NRM with one parent known. If no parents are known, both need assignment to groups and $b_i = 1$. The elements added are 1, -½ and 1/4 .
The logic is that if two parents are known, half of the variance of the breeding value has already been explained, leading to a coefficient of 2 (inverse of ½: the variance of the Mendelian sampling term)) to the animals' diagonal. If the animal has no parents known, and his ancestry is explained by groups (e.g. a breed), none of the animals BV has been explained and leading to a coefficient of 1. This distinction between 'fixed groups' and 'random real ancestors' is easier to maintain, and in a way less relevant, if the groups consist of many 'phantom parents', i.e. if they have many 'progeny'. In that case, the difference between random and fixed will be small (as it is with sires with many progeny). But if groups are made up of phantom parents of just one animal, the distinction is not easy to maintain. This leads to the conclusion that there are some theoretical arguments about fitting group effects as random rather than fixed.

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.

In the example, there are different groups for sires and dams, as selected sires may have a different (usually higher) genetic merit than the average of selected dams. However, there is some danger here, as group solutions could become confounded. In this example, if animal 4 was discarded, it would not be possible to estimate a difference between G1 and G2, and the coefficient matrix would be singular.

Genetic Evaluation across Breeds

Many genetic evaluation systems are for one breed at a time. Sometimes there is a good statistical reasons for this, that animals from different breeds are hardly ever found on the same farm, let alone in the same contemporary (management) group. In that case, breed effects can not be estimated from the data, and an across breed evaluation is not justified. A second reason might be more political, as some genetic evaluations are organized by breed societies, that have no interest in crossbreeding or comparing themselves with other breeds (especially if they might look less favourable). From a neutral perspective it would be best to evaluate animals always across breeds and have good linkages between breeds (many farms with more than one breed). In that case, selection ca be optimized across breeds, and use of genetic resources should be optimal (although there are some interesting optimization problems here for animal breeders).
The main issues to consider with across breed evaluation are

- Modeling and estimating the breed differences
- Modeling and estimating crossbreeding effects
- Modeling and estimating differences in variances between breeds

Breed differences (additive genetic effects between breeds) can be dealt with through appropriate genetic grouping. Whether the breed effects are accurately estimable depends on the distribution of different breeds across te different management groups. Breed comparisons can only be made based on data on different breeds within the same fixed effect level (e.g. of contemporary group)
Note that often, both direct and maternal breed effects need to be estimated. For the latter we also need dams of different breeds to be compare in the same herd.

Non additive genetic effects (between breeds)

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. For example, the heterosis coefficient for an animal with a sire having p_s as a proportion from breed A and $1-p_s$ from breed B, and a dam with coefficient p_d and $1-p_d$, would be equal to $p_s.(1-p_d) + (1-p_d).p_d$. This is easy to see as it predicts the proportion of 'heterozygous alleles'.

		Dam alleles	
	Sire alleles	Prop. Breed A	Prop breed B

		p_d	$1-p_d$
Proportion from breed A	p_s	$p_s.p_d$	$p_s.(1-p_d)$
Proportion from breed B	$1-p_s$	$(1-p_s).p_d$	$(1-p_s).(1-p_d)$

There is a relatively simple extension to coefficients for multiple breeds, although gives an additional complication that AxB heterosis may not be the same as BxC heterosis, etc.
The coefficient for epistasis is related to heterozygosity of the parents' genome. This can be derived as e.g. as $p_s(1-p_s) + p_d.(1-p_d)$. This coefficient would represent what is also known as 'recombination loss (Dickerson, 1969). However, there are several epistatic models possible, depending on the actual allelic actions and interactions tat are hypothesized (see Kinghorn, 1983). In any case, additive and non-additive effects in crossbreeding data should be accounted for as these effects influence the mean (as first moments) and genetic evaluation would be baised if they were not accounted for.

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).

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.

Exercises 1.

1) Consider the following data

Animal	sire	dam	breed		performance
1	0	0	Jersey		220
2	0	0	Jersey		260
3	0	0	HF		280
4	0	0	HF		320
5	1	2	Jersey		240
6	3	4	HF		300

Set up mixed model equation with groups according to the regular MME
Set up the modified mixed model equations
Discuss the interpretation of the group solutions as 'phantom parents'
Note the matrix $Q'*A^{-1}$ and discuss the meaning of this.

2) Repeat the first exercise with the following data

Animal	sire	dam	breed		performance
1	0	0	Jersey		220
2	0	0	Jersey		260
3	0	0	HF		280
4	0	0	HF		320
5	1	4	Xbred		265
6	3	2	Xbred		275

Exercise 2

In the following example, set up mixed model equations. Consider only effects of breed. Determine breed contribution of each animal and also EBV's 'across breeds'.

Calf	**Sire**	**Dam**	**%Angus**	**%Nelore**	**Yearling Wgt**
1	unknown	unknown	100	0	320
2	unknown	unknown	0	100	280
3	unknown	unknown	50	50	310
4	1	unknown	50	50	304
5	3	2	follows from above		307
6	1	2	follows from above		296
7	4	5	follows from above		302
8	3	6	follows from above		314